Science & Technology Facilities Council

Conference Proceedings
RAL-CONF-2018-001

PV2018: Proceedings of the 2018 conference on adding value and preserving data

Harwell, UK
15th-17th May, 2018

Esther Conway (editor),
Kate Winfield (editorial assistant)

May 2018

Enquiries concerning this report should be addressed to:

RAL Library
STFC Rutherford Appleton Laboratory
Harwell Oxford
Didcot
OX11 0QX

Tel: +44(0)1235 445384
Fax: +44(0)1235 446403
email: libraryral@stfc.ac.uk

Science and Technology Facilities Council reports are available online at: http://epubs.stfc.ac.uk

ISSN 1362-0231

Proceedings of the 2018 conference on adding value and preserving data

This publication is a Conference report published by the Science and Technology (STFC) Library and Information Service.

Contact information
Name: Esther Conway
Address: STFC, Rutherford Appleton Laboratory
Harwell, Oxon, UK
Email: esther.conway@stfc.ac.uk
Tel.: +44 01235 446367
STFC
https://www.stfc.ac.uk

RAL-CONF-2018-001
ISSN- 1362-0231.

Preface

The PV2018 Conference welcomes you to its 9th edition, to be held 15th – 17th May 2018 at the Rutherford Appleton Laboratory, Harwell Space Cluster (UK), hosted by the UK Space Agency and jointly organised by STFC, NCEO and the Satellite Applications Catapult.

For its ninth edition, the conference series moves to its hosts the UK Space Agency and is located at the Rutherford Appleton Laboratory part of the Harwell Space Cluster in the UK to continue addressing prospects in the domain of data preservation, stewardship and value adding of scientific data and research related information.

As we enter the era of big data; for this conference year, we extended a special invite to large-scale scientific archives so we can facilitate discussion of emergent issues across scientific domains. This conference explores a new and exciting technology age, where are seeing large-scale collaborations occurring on state of art virtual research environments and novel collaborative infrastructures. The PV2018 conference consisted of four session/theme areas had the following objectives

- Facilitate Science Archives and Data Service Providers sharing knowledge, experiences, and lessons learnt and best practices. In addition to fostering cooperation in the areas of Data Exploitation, Preservation and archived Data Stewardship.
- Address key emerging issues for science archives including but not limited to Open Data, Big Data, Managing Heterogeneity, Data Management Planning, Data Usability, Exploitation and Impact.
- Provide a forum for organisations dealing with preservation of own data and value adding to present the status of their activities, plans and expectations. In PV2018 we particularly welcome input from a broad range of science archives and data providers. In addition to space data archives we would like to extend a special invitation to.
- Large science facilities from different domains to facilitate discussion of our common challenges.
- Specialist science archives and data service providers who are integrating data with space-based observation to produce innovative data services.

Session 1: Data stewardship approaches to ensure long-term data and knowledge preservation and data standards. In this session, we consider the best practises for the long-term preservation of the data and other results associated with research across the preservation lifecycle, from the submission of data packages for preservation, to the access of data products. This includes the organisational structures, policies and standards adopted by data centres and archives to assure cost-effective preservation, together with risk management, uncertainty quantification, quality assessment and the evaluation of preservation capabilities. Further, we will consider novel architectures and tools used to realise different preservation strategies, and standards, tools and languages to capture the preservation context, including the preservation of data formats, the use of identifiers, metadata, semantics, data provenance, quality and uncertainty.

Session 2: Adding value to data and facilitation of data use: In this session, we consider activities that add value to archived data, facilitate their use or produce novel data services. Data archivists often focus most of their energy on creating well-formed, well-documented archives with the expectation that they will be available for the next 50 to 100 years. However, archived data are meaningless if they cannot be easily retrieved, understood, and used. As a result we would like to invite submissions from projects or archives who rising to the challenge of enhancing data in order to facilitate exploitation of data assets.

Session 3: Virtual Research Environments for science data exploitation and value adding: This session will consider new challenges, activities and research related to Virtual Research Environments or Collaborative Environments. While Massive data growth is calling for a new paradigm, with a shift towards “bring the user to the data", where scientists can bring their own code and run it where the data actually reside, instead of downloading the data and run their analysis on their computer. There is also an increased need for data, associated documentation and software long-term preservation and accessibility, for scientists to be able to re-run data analysis that was initially applied on the data. Last, scientists are now expecting to share not only their data, but also their software and the results of their research activities, and to work with their collaborators in an easy and effective manner, regardless of their location.

Session 4: Data preservation in practice: past (present) and future: The purpose of this session is to examine existing practices and systems and highlight what has been learnt, including how to best benefit from collaboration between projects and/or disciplines. It will also look forward and attempt to understand how new developments and/or technologies and/or tomorrow's data volumes might influence or even constrain how things will be done in the future. Data preservation is not a static field: we wish to use this session to explore what we have learnt from previous migrations and to consider how to best prepare for the future, including potentially disruptive scenarios. We would also like to facilitate discussion on how different services involved in LTDP interplay and to use this opportunity to consider how we measure success and respond to requirements from funding agencies, such as those for F.A.I.R. data management.

A total of 77 abstracts were submitted for presentation at the conference. Following the peer-review process by members of the conference programme committee, 48 papers were selected for oral presentation. They are complemented by 22 poster presentations (for a total of 310 distinct co-authors with affiliations distributed over different countries from all continents).

The conference introductory welcome, event and dinner talks was given by was given by Tony Hey (STFC), Esther Conway (Centre for Environmental Data Analysis), Chris Mutlow (RAL Space), Beth Greenaway(UKSA), John Remedios (NCEO), Stuart Martin (Satellite Applications Catapult), Sue O’Hare (ESA Business Incubation Centre), Michael Gleaves (Hartree Centre) and Hugh Mortimer(RAL Space)

These proceedings consist of a collection of 43 short papers and 20 abstracts corresponding to the oral and poster presentations delivered at the conference. They are organized in sections according to conference sessions followed by the contributions that were presented during the poster session.

Further to the oral and poster contributions, the conference has been fortunate to receive keynote lectures and discussion panel addressing a variety of data science topics of interest to PV2018

1. The work of the CEOS WGISS group by Mirko Albani (ESA ESRIN)
2. The EVER-EST project by Rosemarie Leone (ESA ESRIN)
3. Data value and curation across countries, across domains discussion lead by Katrin Molch (DLR)
4. The European Open Science Cloud by Juan Bicarregui (STFC)
5. Copernicus C3S planned service by Carlo Buontempo (ECMWF)
6. Open science data management by Rachel Bruce (JISC)

We enjoy participation from projects, organisations and individuals developing novel data services within and as a result of these environments. Participation includes a broad range of scientific disciplines seeking to preserve and derive the maximum value from their data. This edition of the PV conference series is deeply grateful to UKSA and it's organising partners NCEO, STFC and the Satellite Applications Catapult.

RAL Space Science & Technology Facilities Council National Centre for Earth Observation CATAPULT Satellite Applications

Conference Chairs

Conference Chair: Tony Hey - Chief Data Scientist STFC
Conference Co- chair: Harald Rothfuss - EUMETSAT

Organising Committee

Head of Organising Committee: Caroline Callard – STFC/RAL Space
Esther Conway – STFC/CEDA
Brian Matthews – STFC/SCD
Poppy Townsend – STFC/CEDA
Richard Hilton – Satellite Applications Catapult
Anastasia Bolton – Satellite Applications Catapult
Jan Fillingham – NCEO

Local Organising Committee

Poppy Townsend – STFC/CEDA
Ed Williamson - STFC/CEDA
William Tucker – STFC/CEDA
Kate Winfield – STFC/CEDA
Hayley Gray – STFC/CEDA

Scientific Programme Committee

Head of Scientific Programme Committee: Esther Conway - CEDA
David Giaretta - APA
Brian Mathews - STFC
Richard Moreno- CNES
Nancy Ritchey - NOAA
Tom Stein - NASA PDS
Reta Beebe - NASA PDS
Christophe Arviset - ESA ESAC
Mirko Albani - ESA ESRIN
Pascal Lecomte - ESA Climate Change Office
Kevin Ashley - DCC
Jamie Shiers - CERN
Eberhard Mikusch - DLR
Robert Elliott - RAL Space
Phil Kershaw – CEDA
Richard Hilton - Satellite Applications Catapult

Table of contents in order of programme

Oral Presentations

Poster presentation

A Distributed Analytics Framework for Large-scale Heter- ogeneous Geoscience Data

Kwo-Sen Kuo[1], Michael L Rilee[2], Lina Yu[3], Yu Pan[3], Feiyu Zhu[3], Hongfeng Yu[3]

[1] University of Maryland, College Park, US
[2] Rilee Systems Technologies LLC, Derwood, MD, US
[3] University of Nebraska-Lincoln, US

We present a distributed framework in support of large-scale heterogeneous geoscience data analytics. We have developed SpatioTemporal Adaptive-Resolution Encoding (STARE) to represent and co-align heterogeneous data spatiotemporally. We partition and distributed the spatiotemporally aligned data in an interleaved manner to ensure balanced workload among distributed computer nodes. A prototype system has been implemented based on an array database management system, SciDB, which has facilitated us in implementing sophisticated analytics and executing them in a scalable and interactive fashion. Our framework provides a viable solution for geoscientists to break through in their scientific endeavors of using big geo-data.

Keywords: Data placement; Analytics; Indexing; Scalability; Geoscience data

1. Introduction

With advanced computing and observation techniques, geoscientists can routinely access various datasets in support of their scientific studies. These datasets are typically characterized by spatiotemporal hetero- geneity and large volumes, posing a twofold challenge to analyzing them effectively and efficiently. First, different geoscience datasets are generated from diverse sources, such as different instruments and/or various computational models. Multiple datasets employed in a study can be heterogeneous in type, struc- ture, and semantics. Representative examples of data models include Grid, Swath, and Point. This leads to a non-trivial task for scientists to manage these datasets, and further identify features or patterns co- located among these datasets. Second, while many sophisticated techniques have been developed to tackle large volumes of data, among which a common strategy is to employ multiple computer nodes (e.g., a cluster) with shared and/or distributed memory parallel computations (SMP and DMP respectively), it remains a challenging problem to place heterogeneous geoscience data in a distributed environment and simultaneously avoid costly data transfer and repartitioning in analyzing features or patterns.

To address these challenges, we present a distributed framework in support of scalable analytics of large-scale heterogeneous geoscience data. First, we have devised and implemented a key technology, SpatioTemporal Adaptive-Resolution Encoding (STARE), to unify and index geo-spatiotemporal data to address the heterogeneity challenge. In our current study, all three representative geospatial data models (Grid, Swath, and Point) are supported. In fact, the generality of STARE makes it possible to index all geo- datasets in a unified fashion and significantly reduce the cost of data preparation and management. Second, by leveraging STARE-based data access patterns, we have developed a data partition and distribution method that co-locates data chunks from diverse datasets on the nodes of a distributed cluster environment. This method significantly reduces unnecessary data transfer and repartitioning to ensure scalable perfor- mance in data analytics, in particular, for geophysical features co-located among multiple heterogeneous datasets. In addition, this method interleaves data partitions among distributed computer nodes according to user data access patterns and thereby achieves optimal workload balancing. Based on these advanced data indexing and placement techniques, we have

developed a set of analytics functionalities that can effectively and efficiently access and process data in a scalable manner with respect to combinations of datasets and concurrent user numbers.

2. Our Framework

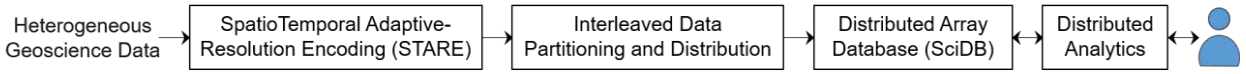


Figure 1: The major components of our framework.

Figure 1 shows the major technical components of our framework in support of effective and efficient ana- lytics for large-scale heterogeneous geoscience data. We describe these components in detail in the fol- lowing sections.

2.1 SpatioTemporal Adaptive-Resolution Encoding (STARE)

We have devised and implemented in C++ the SpatioTemporal Adaptive Resolution Encoding, STARE, indexing scheme to co-align all geo-datasets (Kuo and Rilee 2017). STARE consists of two independent parts, spatial and temporal, each encodes corresponding index and resolution in a 64-bit signed integer. Both are hierarchical.

For the spatial component, we implemented a revised hierarchical triangle mesh (HTM) (Kunszt et al., 2001), with an octahedron as the root (starting) polyhedron, to index geolocation. Its hierarchy forms a quadtree and, with 57 bits, it reaches a precision of ~7 cm in geolocation accuracy. With the sign bit left unset (i.e., the resulting integer index is always positive), we use the least significant 6 bits to denote the hierarchy level best corresponding to (but covering) the data resolution. STARE spatial index is much more amenable to programming manipulations than longitude-latitude, providing a computationally efficient al- ternative.

For the temporal component of STARE, calendrical units are used to form the hierarchy and are there- fore termed hierarchical calendrical encoding (HCE). A 64-bit integer can index hundreds of millennia with millisecond precision. The sign is used to denote before (-) or after (+) the preset start of an epoch. The least significant few bits also denote the best corresponding temporal resolution of the data.

STARE can thus universally index all geo-spatiotemporal data arrays. This universality effectively har- monizes the tremendous diversity in geo-data representation, making STARE a truly transformative inno- vation to offer unprecedented geo-data interoperability. When it is used to index geo-data arrays, it guar- antees spatiotemporal data placement alignment. Furthermore, since STARE indexing is hierarchical with (approximate) data resolution information imbedded, it embodies many additional advantages that are hard to find all together in an alternative (see section 3 of Kuo and Rilee 2017 for details).

2.2 Interleaved Data Partitioning and Distribution

Through a preorder traversal of the quadtree generated by the spatial part of STARE, we can visit the quadtree leaves along a space-filling curve that essentially groups spatially nearby triangular regions to- gether, which provides us with flexibilities on data partitioning and distribution among computer nodes.

If we partition and distribute data points of the triangular mesh along the space-filling curve, each pro- cessor will be assigned a set of contiguous regions on the spherical surface, corresponding to a nearly equal amount of data. However, geoscience data is commonly explored and studied in visual analytics

manners, where users access data within certain viewing regions. Therefore, the computer nodes whose assigned regions are not visible will become idle, resulting in unbalanced workload.

To address this issue, we partition and distribute in the data points along the space-filling curve in a round robin fashion. Through this interleaved method, we can place the neighboring regions among different computer nodes. Therefore, during users' interactive exploration, visible regions will be from many nodes of the cluster, helping to achieve more balanced workload.

2.3 SciDB-based Implementation

Based on our previous research (Doan et al., 2016), we advocate the tightly coupled compute-storage approach to take full advantage of data placement alignment (DPA). By tightly coupled compute-storage, we mean that the same engine not only executes the analysis computations but also manages data storage. Parallel distributed database management systems (DBMSs) exemplify this approach. Because of the tight compute-storage integration, DBMSs know (by loading required chunk distribution maps into DRAM and consulting them) exactly on which node each data chunk resides, in memory or on disk, and are thus able to coordinate the computation optimally, including especially DMP. This is a significant advantage over the loosely coupled compute-storage approach represented by, e.g., Spark (Zaharia et al., 2010) and MapRe- duce (Dean and Ghemawat, 2008).

We implement our framework based on a tightly coupled approach, SciDB (Brown, 2010), an advanced data management and analytics platform featuring complex analytics in a parallel array DBMS (ADBMS), developed, tested, documented, and supported by Paradigm4. SciDB supports atomicity, consistency, iso- lation, durability (ACID) semantics and implements multi-version concurrency control (MVCC). The ACID properties of the system are critical for the protection of valuable datasets, especially in a multiuser envi- ronment.

The data partitioning performed by SciDB (and other DBMSs) is analogous to domain decomposition, done normally on-the-fly in DRAM (e.g. with C++ & Fortran) for DMP. For SciDB, the storage system on each cluster node effectively extends the node's DRAM for the same purpose. When we use STARE as a uniform indexing scheme with our interleaved data partitioning and distribution method, SciDB in effect "pre- domain decomposes" consistently all geo-data arrays, distributes, and stores them on nodes of the cluster to guarantee spatiotemporal DPA. When the analysis is pleasingly parallel (PP), SciDB loads the necessary chucks into the DRAM of their respective nodes and leverages SMP on each node. When the analysis requires DMP and communications among nodes become necessary, SciDB coordinates the work and handles the message passing using its custom protocol. Users of SciDB can thus blissfully issue queries to accomplish their analysis tasks, leveraging both SMP and DMP without parallel programming skills.

2.4 Distributed Analytics

Because it is array-based, SciDB is more suitable for scientific data analytics than traditional relational DBMSs. It provides extensive and flexible operators that can be "wired" together to efficiently accomplish more complex tasks. It is extensible through user-defined types, functions, and operators (UDT/F/Os or collectively UDXs). It features a plugin interface allowing developers to implement distributed plugins in C/C++ leveraging MPI or a streaming interface to pass data chunk by chunk to an analytic process in, e.g., Python or R. With STARE that enables efficient and flexible subsetting through spatiotemporal DPA with SciDB, we have implemented several advanced integrative analytics.

First, we have incorporated re-gridding tools into our framework in support of fundamental comparison operations. Geoscientists routinely compare simulation outputs from different models,

observations ob- tained by different means, and simulation outputs with observations. Geometric properties, including reso- lutions, of the datasets involved are generally dissimilar. A grid cell or instantaneous field of view (IFOV) of a coarser-resolution dataset may cover multiple grid cells or IFOVs of the finer-resolution dataset, yielding a one-to-many comparison that is difficult to interpret. Thus, we need to re-grid the geometry from one to the other or both to a common third, resulting in one-to-one comparisons. We have implemented a re- gridding module as a SciDB UDO that performs distance-weighted re-gridding. We are in the process of adding functionality to this UDO, including flux conservation during re-gridding, which requires DMP. With the integrated re-gridding capability enabling effortless comparison studies, data preparation and analyses can be better automated, supporting systematic data mining and machine learning applications.

Second, we have implemented a highly performing UDO for connected component labeling (CCL) (Oloso et al., 2016), which also requires DMP. The use of CCL is ubiquitous in image processing where the algorithm scans the images and groups pixels into components based on pixel connectivity. CCL is essen- tial for tracking phenomenon events (aka features) and thus for obtaining high-context conditional statistics to enable process-based diagnostics.

Third, we have constructed a high-level visualization interface featuring multiple visual analytics capa- bilities in different views, including user-customized queries, interactive multivariate rendering, and real- time statistical analytics. The different views are performed in a linked fashion, enabling users to simulta- neously explore multiple aspects of data.

3. Result

We have integrated the technology innovations described above culminated in a prototype system capable of supporting rapid-response visual analytics. Through a high-level graphical browser interface (based on Google Maps) supported by a cluster of 16 lightweight nodes running SciDB, users can interact with ani-mation of hourly precipitation data from a diverse set of model output and observations: 2 gridded datasets of very different resolutions and 1 swath dataset from satellite observations. The system supports multiple users to conduct (currently) rudimentary analyses, e.g. aggregate statistics, time series, and percentile, using this prototype system. We believe this prototype represents the embryo of a breakthrough, because in terms of addressing analysis velocity challenge, nothing is more desirable than interactive speed.

4. Conclusion

We have holistically addressed most key components in an end-to-end distributed analytics framework for large-scale heterogeneous geoscience data. Our prototype system has supported non-trivial analytics op- eration and achieved scalable performance. In the future, we plan to leverage Graphics Processing Units (GPUs) to further improve the system performance, and enrich visualizations of analysis results. New data placement and cache techniques will be studied for finer grained data movement between CPUs and GPUs.

Acknowledgements

This research has been sponsored in part by the National Science Foundation through grants ICER-1541043 and ICER 1540542.

References

BROWN, P. G. Overview of SciDB: large scale array storage, processing and analysis. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010. ACM, 963-968.

DEAN, J. & GHEMAWAT, S. 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM,* 51**,** 107-113.

DOAN, K., OLOSO, A. O., KUO, K.-S., CLUNE, T. L., YU, H., NELSON, B. & ZHANG, J. Evaluating the
impact of data placement to spark and SciDB with an Earth Science use case. Big Data (Big Data), 2016 IEEE International Conference on, 2016. IEEE, 341-346.
KUNSZT, P. Z., SZALAY, A. S. & THAKAR, A. R. 2001. The hierarchical triangular mesh. *Mining the sky*.
Springer.
Kuo, K.-S. and M. Rilee, STARE – toward unprecedented geo-data interoperability, in 2017 Conference on Big Data from Space. European Space Agency. 2017: Toulouse, France.
OLOSO, A., KUO, K.-S., CLUNE, T., BROWN, P., POLIAKOV, A. & YU, H. Implementing connected component labeling as a user defined operator for SciDB. Big Data (Big Data), 2016 IEEE International Conference on, 2016. IEEE, 2948-2952.
ZAHARIA, M., CHOWDHURY, M., FRANKLIN, M. J., SHENKER, S. & STOICA, I. 2010. Spark: Cluster
computing with working sets. *HotCloud,* 10**,** 95.

LONG-TERM DATA PRESERVATION DATA LIFECYCLE AND STANDARDISATION PROCESS

Mirko Albani[1], Rosemarie Leone[1], Iolanda Maggio[2], Katrin Molch[3] LTDP WG[4]
European Space Agency[1], Rhea Group[2], DLR[3], International Partners[4]

Science and Earth Observation data represent today a unique and valuable asset for humankind that should be preserved without time constraints and kept accessible and exploitable by current and future generations. In Earth Science, knowledge of the past and tracking of the evolution are at the basis of our capability to effectively respond to the global changes that are putting increasing pressure on the environment, and on human society. This can only be achieved if long time series of data are properly preserved and made accessible to support international initiatives. Within ESA Member States and beyond, Earth Science data holders are increasingly coordinating data preservation efforts to ensure that the valuable data are safeguarded against loss, and kept accessible and useable for current and future generations. This task becomes increasingly challenging in view of the existing 40 year's worth of Earth Science data stored in archives around the world and the massive increase of data volumes expected over the next years from e.g. the European Copernicus Sentinel missions. Long Term Data Preservation (LTDP) aims at maintaining information discoverable and accessible in an independent and understandable way, with supporting information, which helps ensuring authenticity, over the long term. A focal aspect of LTDP is data Curation. Data Curation refers to the management of data throughout its life cycle. Data Curation activities enable data discovery and retrieval, maintain its quality, add value, and allow data re-use over time. It includes all the processes that involve data management, such as pre-ingest initiatives, ingest functions, archival storage and preservation, dissemination, and provision of access for a designated community.
The paper presents specific aspects, of importance during the entire Earth observation data lifecycle, with respect to evolving data volumes and application scenarios. These particular issues are introduced in the section on 'Big Data' and LTDP. The Data Stewardship Reference lifecycle section describes how the data stewardship activities can be efficiently organised, while the following section addresses the overall preservation workflow and shows the technical steps to be taken during Data Curation. Earth Science Data Curation and preservation should be addressed during all mission stages - from the initial mission planning, throughout the entire mission lifetime, and during the post- mission phase. The Data Stewardship Reference Lifecycle gives a high-level overview of the steps useful for implementing Curation and preservation rules on mission data sets from initial conceptualisation or receipt through the iterative Curation cycle.

Keywords: LTDP Data Lifecycle; Preservation Workflow; Data Management and Stewardship Maturity Matrix, PDSC, GSCB, CEOS, GEO

Introduction

The paper presents specific aspects, of importance during the entire Earth observation data lifecycle, with respect to evolving data volumes and application scenarios. These particular issues are introduced in the section on 'Big Data' and LTDP. The Data Stewardship Reference lifecycle section describes how the data stewardship activities can be efficiently organised, while the following section addresses the overall preservation workflow and shows the technical steps to be taken during data curation. The paper concludes with introducing international collaboration for developing coordinated and harmonised lifecycle concepts.

DATA STEWARDSHIP REFERENCE LIFECYCLE

Earth Science data curation and preservation should be addressed during all mission stages - from the initial mission planning, throughout the entire mission lifetime, and during the post-mission phase. The Data Stewardship

Reference Lifecycle (Figure 1) gives a high-level overview of the steps useful for implementing curation and preservation rules on mission data sets from initial conceptualisation or receipt through the iterative curation cycle.

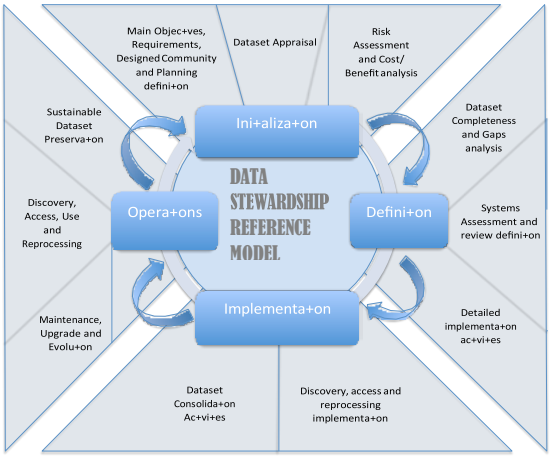


Figure 1 - LTDP Data Stewardship Reference Lifecycle

The core target of the LTDP lifecycle is the preserved data set, composed of consolidated:

- **Data records**: these include raw data, Level 0 data and higher-level products, browses, auxiliary and ancillary data, calibration and validation data sets, and descriptive information.
- **Associated knowledge**: this includes all the processing software used in the product generation, quality control, the product visualization and value adding tools, and documentation needed to make the data records understandable to the designated community. This includes among others mission operation concept, products specifications, instruments characteristics, algorithms description, Cal/Val procedures, mission/instruments performances reports, quality related information, etc. It is necessary to ensure data remain understandable and usable.

The final, consistent, consolidated, and validated "data records" are obtained by applying a consolidation process consisting of the following main steps:

- Data collection
- Cleaning/pre-processing
- Completeness analysis
- Processing/reprocessing

In parallel to the data records consolidation process, the data records knowledge, associated information and processing software are also collected and consolidated.

Data stewardship implements and verifies, for the relevant preserved data sets, a set of preservation and curation activities on the basis of a set of requirements defined during the initial phase of the curation exercise.

Data preservation activities focus on Earth observation data sets long-term preservation, and are tailored according to its mission specific preservation/curation requirements. They consists of all activities required to ensure the "preserved data set" bit integrity over time, its discoverability and accessibility, and to valorise its (re)-use in the long term (e.g. through metadata/catalogue improvement, processor improvement for algorithm and/or auxiliary data changes and related (re)-processing, linking and improvement of context/provenance information, quality assurance). Preservation activities for digital data record acquired from the space segment and processed on ground embrace ensuring continued data records availability, confidentiality, integrity and authenticity as legal evidence to guarantee that data records are not changed or manipulated after generation and reception over the whole continuum of data preservation (archival media technology migration, input/output format alignment, etc.), valorisation and curation activities. The usage of persistent identifier for citation is part of the agency long term data preservation best practices.

Data curation activities aim at establishing and increasing the value of "preserved data sets" over their lifecycle, at favouring their exploitation, possibly through the combination with other data records, and at extending the communities using the data sets. These include activities such as primitive features extraction, exploitation improvement, data mining, and generation/management of long time data series and collections (e.g. from the same sensor family) in support to specific applications and in cooperation with international partners.

Data stewardship activities refer to the management of an EO Data set throughout its mission life cycle phases and include preservation and curation activities. It includes all the processes that involve data management

(ingestion, dissemination and provision of access for the designated community) and data set certification.

PRESERVATION WORKFLOW

The LTDP data stewardship reference lifecycle is also represented through the preservation workflow, which defines a recommended set of actions to be sequentially implemented for the preservation of a "data set", with the goal of ensuring and optimizing its (re)-use in the long term. This preservation workflow, collaboratively developed with European space data holders, ensures that Earth observation mission data sets remain accessible and useable in the long term. Applying this workflow will produce a consolidated, accessible and useable Earth observation data set – consisting of the data records and the associated knowledge – and comprehensive documentation of the preservation procedure. While best initiated during the early mission planning phases, the preservation workflow can also be applied to data sets of current and historic Earth observation missions. The preservation workflow recommended actions/steps are the following:

- EO missions/sensors data set appraisal, definition of designated community & preservation objective (with preservation/curation requirements)
- Tailoring of mission specific consolidation process (on the data records)
- EO missions/sensors data set PDSC tailoring and inventory table filling (including dependencies: Inventory Data Model)
- Tailored PDSC consultation with designated community
- Implementation of tailored consolidation process and collection of documentation and processing software
- Update of EO missions/sensors data set PDSC & inventory table
- Archive & ingestion, master inventory and catalogue population
- Dissemination & Web configuration
- Risk & cost assessment, preservation & cost planning, implementation.

WGISS DATA STEWARDSHIP MATURITY MATRIX WHITE PAPER

The scope of the on-going WGISS Data Stewardship Maturity Matrix definition is to measure the overall preservation lifecycle and to verify the implemented activities needed to preserve and improve the information content, quality, accessibility, and usability of data and metadata. It can be used to create a stewardship maturity scoreboard of dataset(s) and a roadmap for scientific data stewardship improvement; or to provide data quality and usability information to users, stakeholders, and decision makers.

In the extended environment of Maturity Matrices and Models, the Maturity Matrix for "Long-Term Scientific Data Stewardship", of Ge Peng and Jeffrey L. Privette (2015), represents a systematic assessment model for measuring the status of individual datasets. In general, it provides information on all aspects of the data records, including all activities needed to preserve and improve the information content, quality, accessibility, and usability of data and metadata. This was used as a starting point of the WGISS Data Stewardship Maturity Matrix. In parallel, the GEO Data Management Principles Task Force was tasked with defining a common set of GEOSS Data Management Principles (DMP-IG). These principles address the need for discovery, accessibility, usability, preservation, and curation of the resources made available through GEOSS.

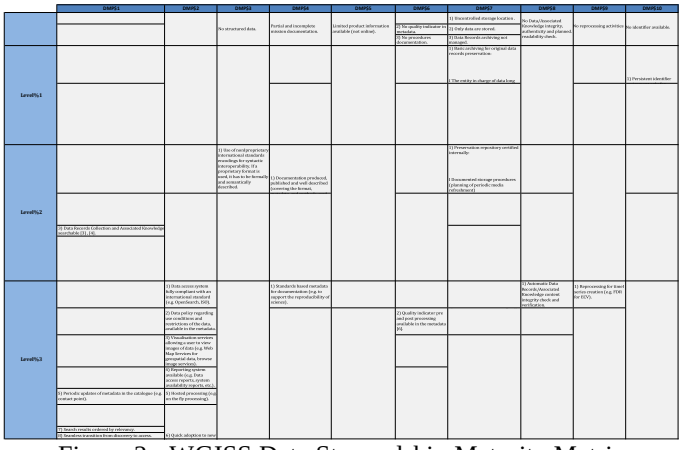


Figure 2 - WGISS Data Stewardship Maturity Matrix

The content of the WGISS Data Stewardship Maturity Matrix represents the result of a combined analysis performed on the DMP-IG and a consultation at European level, with the Long Term Data Preservation Working

Group. The rationales for applying the WGISS Data Stewardship Maturity Matrix are:

- Providing data quality, usability information to users, stakeholders, and decision makers;
- Providing a reference model for stewardship planning and resource allocation;
- Allowing the creation of a roadmap for scientific data stewardship improvement;
- Providing detailed guidelines and recommendations for preservation;
- Evaluating if the preservation follows best practices;
- Giving a technical evaluation of the level of preservation and helping with self-assessment of preservation;
- Providing a status of the preservation, but doesn't offer information on numbers or averages related to preservation;
- Helping to break down problems related to preservation, and to understand the costs associated with each preservation level;
- Funding agencies can define certain goal levels that they would like to reach.
- It is a self-assessment and it is applied at dataset level.

COOPERATION ACTIVITIES

ESA is cooperating in the LTDP domain in Earth observation with European partners through the LTDP Working Group, formed within the Ground Segment Coordination Body (GSCB), and with other international partners, through participation to various working groups and initiatives. The EO LTDP framework international context is shown below:

Figure 3 - EO LTDP Framework international context

The LTDP core documents have also been reviewed and approved at international level within the Committee on Earth Observation Satellites (CEOS) and the Group on Earth Observations (GEO). A review of the Preservation Workflow document is currently on going in the frame of the CCSDS Data Archive Ingestion (DAI) working group.

MEDIA RESCUE ACTIVITY: LESSONS LEARNED

Heritage data preservation activities include the preservation of unique data that can only be recovered from historical media. Therefore, the preservation of these media, together with the hardware that could read the media, should be ensured. During the rescue activity of JERS-1 mission media, some lessons learned were collected. Having no inventory available for the JERS-1 media at the Fucino ground station, several trips to the facility were undertaken in order to manually generate the media inventory. This was later compared against the JERS-1 data already available at ESA, which allowed to identify the missing data. However, this was not a simple task, as a large part of the media labels were either missing crucial information or this information could not be easily read, due to deterioration over time, as the storage environment was not systematically monitored.

The main lesson learned from this media rescue activity is that long-term preservation should be considered, and planned for, from the initial stages of a mission, in order to ensure that long-term data preservation policies are followed throughout the mission lifetime. Preservation of the main information on media labels and in local, digital, inventories should also be ensured, together with other Associated Knowledge. Furthermore, the original media, hardware and software should be preserved until it is certain that all unique data that could be recovered, was retrieved from the historical media. This also implies that the physical archiving storage must be located in a well-controlled environment that would prevent deterioration of the media labels or the media itself.

CONCLUSIONS

Data holdings are growing exponentially in Earth Science data archives worldwide. The European Copernicus program will continue to deliver Petabytes of valuable satellite-based Earth observations for many years to come. Only a systematic approach to data preservation during the entire data lifecycle, coordinated between data holders and application communities, will ensure that these data sets will be accessible and useable to current and future generations, for monitoring long-term variations in environmental parameters as a basis for objectively assessing and predicting effects of global change.

REFERENCES

CEOS, "EO Data Preservation Guidelines Best Practices", http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Data%20Preservation%20Guidelines_v1.0.pdf

CEOS, "EO Preserved Data Set Content", http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf

CEOS, "Long Term Preservation of Earth Observation Space Data: Preservation Workflow", http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/Preservation%20Workflow_v1.0.pdf

CEOS, "Associated Knowledge Preservation Best Practices", http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS%20Best%20Practices/CEOS%20Associated%20Knowledge%20Preservation%20Best%20Practices_v1.0.pdf

CEOS, "Generic Earth Observation Data Set Consolidation Process", http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/GenericEarthObservationDataSetConsolidationProcess_v1.0.pdf

CEOS, "Long-Term Preservation of Earth Observation Space Data: Glossary of Acronyms and Terms", http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/EO-DataStewardshipGlossary_v1.2.pdf

CEOS, "CEOS Persistent Identifier Best Practices", http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS%20Best%20Practices/CEOS%20Persistent%20Identifier%20Best%20Practices_v1.2.pdf

GEOSS, "Data Management Principles", https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf

Peng, Privette, Scientific Data Stewardship Maturity Matrix http://www.slideshare.net/gepeng86/scientific-data-stewardship-maturity-matrix

Giving Access and Value Added Services to CDPP Historic Data Sets

Danièle Boucon[1], Nicolas Dufourg[1], Nicolas Lormant[2], Vincent Cephirins[2], Dominique Heulet[1], Vincent Genot[3], Patrick Canu[4], Baptiste Cecconi[5], Christian Mazelle[3], Myriam Bouchemit[3], Elena Budnik[3]

[1] CNES, 18 av E. Belin, 31401 Toulouse Cedex 9, France,[2] AKKA, 7 Boulevard Henri Ziegler, 31700 Blagnac, France,[3] IRAP, 9 av du Colonel Roche, 31400 Toulouse,[4] LPP, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau, [5] LESIA, Observatoire de Paris, CNRS, PSL, Meudon, France
Corresponding author: Daniele Boucon (daniele.boucon@cnes.fr)

This paper presents the validation and reanalysis of some historic data sets by the CNES CDPP data centre, and how it is planned to facilitate the access and to add value to these data sets. The CDPP is the French national data centre for natural plasmas of the solar system. It assures the long term preservation of data obtained primarily from instruments built using French funding, and renders them readily accessible and exploitable by the international community. It also provides services to enable on-line data analysis. The user communities evolve over the long term, in terms of perimeter and practices, such as new standardized formats, more interoperability needs, common tools … For example, for the Space Physics community, the data format is CDF (Common Data Format), and it was not the case thirty years ago. The CDPP has historic data sets, such as Isee1, Giotto, Voyager. Following scientific interest, these data sets are currently re-analyzed, taking into account the needs of the Space Physics community. It is an opportunity to explain the different difficulties encountered, and to give
everyone some good practices for this kind of activity. These data sets will be first accessible through the SIPAD-NG[1]. The next step is to add value and give access to these data sets. The paper will show how data may be used through tools such as AMDA[2]. The paper concludes on the objective in the future to provide new services in order to facilitate the access and add value to scientific Plasma data.

Keywords: historic data sets, data services, CDPP, user needs, data access, value added

INTRODUCTION

The CDPP (Plasma Physics Data Centre, www.cdpp.eu) is the French national data centre for natural plasmas of the solar system. Created in 1998 jointly by CNES (National Centre for Space Studies) and INSU (National Institute for Earth Sciences and Astronomy), the CDPP assures the long term preservation of data obtained primarily from instruments built using French funding, and renders them readily accessible and exploitable by the international community. The CDPP also provides an evolving set of tools and services, as for example services to enable on-line data analysis (AMDA[2]), 3D data visualization in context (3DView[3]), propagation tool[4], and TREPS[5] which enables to transform vector time series in a choice of heliospheric reference frames. The CDPP is involved in the development of interoperability services, participates in several Virtual Observatory projects, and supports data distribution for scientific missions (Rosetta, Taranis, Solar Orbiter).

The main idea of this paper is to explain how the CDPP facilitates the use of data in a way that is open, flexible, and adaptable to the evolving scientific knowledge and needs of the user communities, based on twenty years of expertise and experience.

WHAT WE HAVE

The CDPP gives access to more than thirty scientific missions, represented by 1355 data sets, 6 211 236 files and 132 Tb, through the SIPAD-NG and AMDA tools. Among
these data sets, two-thirds are preserved at CNES, the remaining third is in the form of scientific parameters preserved at IRAP and extracted from data stored on servers all other the world.

[1] SIPAD-NG -Système d'Information, de Préservation et d'Accès aux Données – Nouvelle Génération" – "Information System for Data Preservation and Access – New Generation" : cdpp-archive.cnes.fr
[2] AMDA -Automated Multi-Dataset Analysis: amda.cdpp.eu
[3] 3DView : 3dview.cdpp.eu, see [4]
[4] Propagation Tool : propagationtool.cdpp.eu, see [5]
[5] TREPS : treps.cdpp.eu, see [6]

The oldest ones are about 40 years old, such as Isee-2 (International Sun-Earth Explorer, studying the interaction between the solar wind and the Earth's magnetosphere, 1977-1987), or GEOS (GEOstationary Scientific Satellite studying the particles, fields and plasmas of the Earth's magnetosphere, 1977-1983). Data from operational missions such as MMS (Magnetospheric Multiscale Mission, since 2015) are also available and are continuously updated with newer data.
The CDPP has also other historic data sets from Isee-1, Giotto, Voyager, pointed out by the scientists for their interest, but not yet open to public access. Such data is of great interest for upcoming missions as existing material, and we are able to find new use for data that was not fully exploitable at its creation (e.g. compare to data issued from new mission; reanalyze in the light of new scientific knowledge; homogenize calibrations).

Giotto (RPA –Rème Plasma Analyzer– data set): the RPA experiment was dedicated to measuring and studying the three-dimensional distributions of plasma particles in the vicinity of the comets Halley and 26P/Grigg- Skjellerup. This data is unique, and of high interest for comparison with other comet approaches (such as Rosetta's encounter with Churyumov Gerasimenko).

Isee-1 (Sounder data set): to study the interaction between the solar wind and the Earth's magnetosphere. The Sounder provides a survey of the frequency spectra of the electric components of natural plasma emissions (0-50 kHz and 0-400 kHz). In its active mode (~ 10 % of time), it triggers the resonance spectra from which the electron density is derived. It is the only mission covering the solar cycle over the period 77-87, and the CNES data is unique.

Voyager (PRA –Planetary Radio Astronomy– data set): the objective of the two probes was to study the outer solar system (Jupiter, Saturn, Uranus and Neptune). The EDR –Experimenter Data Record- and DEDR – Decalibrated EDR- data sets are unique as outer planets mission. They have been of interest for the mission Juice. DEDR data covers Saturn and Jupiter fly-bys. If EDR data with further research could cover Neptune and Uranus this would be an added interest (see [2]).

THE WORK CURRENTLY BEEN DONE

The historic data sets previously pointed out were analyzed and converted in the 1990s from the mainframe used at the time (Control Data, Nos-BE and Nos-VE) towards non-proprietary and more usable formats (IEEE standards). At the same period, in 1995, CNES set up a shared facility, the STAF (Service de Transfert et d'Archivage des Fichiers or file transfer and archiving service), to ensure the long-term integrity of the bit streams accumulated by the projects. The collections of tapes associated with historic data sets were then migrated to this new system with more capacity and automated supports.

Today, twenty years after these analyses and first migration operations, facing new needs and the evolution of communities, the CDPP wants to give access to this historic data. Following is a synthesis of the process and the work currently being done on three missions: Isee-1, Voyager, Giotto. First of all, for each mission, the CDPP has a complete inventory of the data set, regularly updated (such as the volume, format …) and all the available information on the data set (description, documentation, catalogs, software, points of contact …). The inventory is the knowledge network of the data. The process for the re-analysis and validation of the data isthe following:

The first step is a meeting with the scientific point of contact. This is a key point, as nothing can be done without his participation; the lack of a scientist and documentation may limit the re-analysis, or may even lead to abandoning data. Then a set of questions are discussed to identify the real interest for the community, and the feasibility of the operation according to the requirements of the Archive such as: uniqueness of the data, data level, data format, applicable standards, availability of the scientist. The third step is the technical part of the work. The main challenge is to guarantee the quality of the data, the transformation towards current formats and levels of the community, and creation of accurate metadata. The last step (and not the least) is the scientific validation of the data before giving the access.

For the three missions cited here, work is in progress for Isee-1 and Voyager and should succeed, is partially finished for Giotto. Below are some typical difficulties encountered, and lessons learned for the future.

Giotto: a presentation on this subject was made during the PV2011 conference (see [1]). As a reminder, the main difficulties were inconsistencies in the principal documentation in the understanding of the raw telemetry data, and a software to read the data on an old PC programmed with an obsolete OS. Seven years after this article, the work is

nearly at an end and waits for final validation. The status is the same, due to the unavailability of the scientist and other priorities on the Archive side.

Isee-1: data was converted 20 years ago into IEEE format and fixed size, with paper format description. For this case the scientist, Patrick Canu (LPP) is willing to participate and invest time. The challenge, for him, is to, once again, get his hands on information, data and processes frozen more than twenty years ago; he has been reading again and re-analyzing the files. He has re-arranged the data on a day to day basis, and produced associated spectra to facilitate the future use. The next step is a deep validation and transformation towards CDF format, before access and diffusion through the SIPAD.

Voyager: the challenge is to re-analyze the DEDR data to create higher level data that is more convenient for use, and finally to transform it into CDF format. The scientist Baptiste Cecconi (LESIA) has discovered numerous duplications in datation, whose origin is hard to identify. This implies a high level of expertise to retreat this data; this is under development.

Lessons learned:

- the need to take into account the archiving from the beginning of the mission, particularly to collect the information for the understanding and use of data as it is created (data description …);
- a homogeneous terminology at community level;
- we do not know in advance the optimal and complete use of space data;
- the need to maintain and update the scientific reference to the data set, whilst it is being preserved;
- the importance of validation;
- the need to preserve data in physical values and with sufficient resolution, directly usable;
- the importance of a great collaboration between scientist and archivists.

Having learned from these lessons, the CDPP has improved both its procedures and services.

THE TOOLS

Among the tools proposed by the CDPP, this article focuses on SIPAD-NG, AMDA and 3DView.

SIPAD-NG is the software system allowing web consultation of the CDPP scientific data catalog and access to this data. It is an OAIS compliant tool, including also functions for data ingestion, storage, and administration.

Fifteen years after the SIPAD-NG start date, user needs and technology obsolescence lead to a new generic Information System, called REGARDS[6] (see [3]). The user needs are mainly the following: higher performance (due to larger scientific data, often more than 1 gigabyte, and growing databases), faster access and retrieval, more complete and expandable data model, interoperability, flexibility to plug value added services. The data model will take into account the community's dictionary (SPASE) with a consistent terminology, more metadata and interactions with Virtual Observatories. Each piece of data will be accompanied by its metadata. This tool should be operational in 2019.

AMDA is a web-based facility for on-line analyses of space physics. This tool allows the user to perform classical data manipulations such as data visualization, parameter computation, and conditional search. AMDA also offers innovative functionalities such as event searches on the content of the data in either visual or automated ways, generation, use and management of time tables (event lists). The SIPAD-NG proposes to access the AMDA database (scientific parameters) through its interface, but as two different archives. The future REGARDS should be able to improve this interface, at least with consistent dictionary, terminology and descriptions (missions, metadata …).

The CDPP also proposes 3DView, a 3D animated visualization tool of (a hundred) spacecrafts orbit and attitude in the solar system as well as scientific models representation and data display along the orbit, putting back them into their spatial context. Moreover all comets and asteroids in a given volume and for a given time interval can be searched and displayed.

As examples of adding value services, the two figures below plot the Giotto RPA electron density during 10 hours near the comet (between 13/3/1986 15:00h and 14/3/1986 01:00h) using the same data, either with AMDA on the left and 3DView on the right.

[6] REGARDS: REnewal of Generic tools to Access and aRchive Space Data

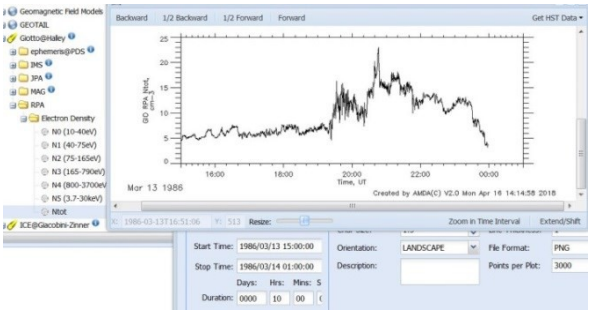


Figure 1: Electron Density E>10 eV is plotted versus time (AMDA)

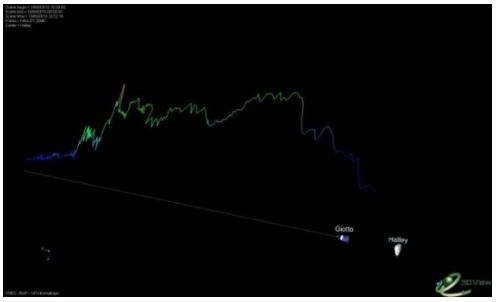


Figure 2: Snapshot of the animation displaying the Giotto trajectory towards the comet the during the time interval; Electron Density E>10 eV is plotted in parallel (3DView)

Taking into account user needs and technology obsolescence, the CDPP has invested in new developments to improve the access and value added services.

AN EVOLVING LOOP OVER THE LONG TERM

Figure 3: Data Cycle

Evolution occurs at different levels; it is not predictable. User communities also evolve.

Data is a living and valuable piece of information, an organic process based on the use and the evolution of user communities. There is an inextricable link between time and data usage, thus creating living data.

IN THE FUTURE

We cannot prejudge the future data usage and the value that will be added as a result of that use.How to add value to the data sets for the user communities in the future in a time driven cycle?

As seen in the previous sections, new needs occur at different levels (data format, metadata, tools …). The evolutions are closely linked to migrations at those levels. For example, we are conscious that migrate all the CDPP database towards the future one for REGARDS will be a large amount of work, that will require the participation of the community.

Technology migrations being apart, a challenge is to maintain an accurate network of information attached to the data, in order to find, access and use it. This is in the hands of both archivists and scientists.

CDPP regularly interacts with scientists, with major yearly meetings (user committee, SWT –Science Working Team-workshops). Between 2016 and 2018, no less that twenty eight scientific publications present results using the tools and databases of the CDPP. For the CDPP, the future is to work hand in hand as a support to the scientists from the beginning of the data life cycle, and anticipate the migrations to maintain the Archive at the best level, acting with the budget constraints.

CONCLUSION

What is perennial? We continuously face evolution at different levels in order to guarantee that data remains understandable, accessible and usable in the long term. Data and tools must be flexible, adaptable and open to the evolving knowledge of the user communities.

CDPP celebrates this year twenty years of accumulated experience (and data):

- it is a real challenge and a long task to rehabilitate data and give access;
- discussion and active listening with the user community is the best method we have found to follow

and anticipate its needs;

- sensitizing and building support is another way to optimize the application of procedures and standards and create an Archive;
- the creation and maintenance of an Archive is an active process following the wheels of progress with implication of all actors on scientific and Archive sides, starting **from the beginning of the mission**.

Perhaps the last words can be summed up by the importance of the relation between the user community and the Archive, as it is the case for the CDPP.

REFERENCES

[1] – Lormant et al., Giotto RPA: How to save unique and invaluable data twenty-five years after their collect?, PV 2011

[2] – Cecconi et al., Natural radio emission of Jupiter as interferences for radar investigations of the icy satellites of Jupiter, Planetary and Space Sci. 61, 32-45

[3] – Bellucci and al., REGARDS, the new CNES generic systemto access and archive space data, PV 2015

[4] – Génot et al., Science data visualization in planetary and heliospheric contexts with 3DView, Planetary and Space Science, doi:10.1016/j.pss.2017.07.007, 2017

[5] – Génot et al., TREPS, a tool for coordinate and time transformations in space physics, Planetary and Space Science, doi:10.1016/j.pss.2017.06.002, 2017

[6] – Rouillard et al., A propagation tool to connect remote-sensing observations with in-situ measurements of heliospheric structures, Planetary and Space Science, doi:10.1016/j.pss.2017.07.001, 2017

Using machine learning to extract data from unstructured research data

Dr Thomas Parsons and Dr Stuart Bowe

Spotlight Data, University of Nottingham Innovation Park, United Kingdom
tom@spotlightdata.co.uk and stuart@spotlightdata.co.uk

Author information

Dr Thomas Parsons is co-founder of Spotlight Data and has worked in research, industry and consultancy across academia, aerospace and pharmaceuticals. He specialises in research data management and applying research to industry.

Dr Stuart Bowe is a Senior Data Scientist at Spotlight Data and specialises in data analysis and machine learning. His background is in Experimental Physics and science.

This practice paper describes the use of the Nanowire software system to extract structured data from datasets and research data landing pages. Nanowire applies machine learning, computer vision, text mining, Natural Language Processing and data analysis techniques to understand files and has been successfully used to analyse and classify research data as part of a research data management project. The paper discusses research data, structuring data using schema.org and Natural Language Processing and machine learning techniques that can be used to understand research data records to aid reuse.

Keywords: Dataset, research, metadata, natural language processing, nanowire software system, research data management

Introduction

This practice paper describes the development of the Nanowire software system to extract structured data from files. Nanowire is being developed in collaboration with the UK Government and is designed to analyse the huge volumes of files that reside in organisations on shared file systems and repositories. Nanowire applies machine learning, computer vision, text mining, Natural Language Processing and data analysis techniques to understand files and has been successfully used to analyse and classify research data as part of a research data management project.

This paper discusses research data, structured data and the opportunities for machine learning to aid the reuse of data.

Reusing research data

Research data has a broad definition and encompasses the majority of work a researcher undertakes:

"Research data, unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results." (University of Edinburgh)[1]

This definition means that in practice, a researcher can generate anything from a simple text file, to an MRI file, to code that runs a simulation. Depending upon whether this data can then be shared, researchers are encouraged to deposit the data within an open data repository (e.g. Dryad[2]) or their institutional repository (e.g. the University of Cambridge data repository[3]) for reuse by others. Research councils and charities usually require researchers to check whether existing data can be used to tackle their research project, before committing further funding[4]. However, the sheer range of potential types of data, means that finding data for reuse can be difficult and is dependent upon the metadata and search capabilities of the data repository the data is stored in. Therefore, opportunities exist to enhance the process of finding data by applying machine learning and Natural Language Processing to understand and classify data, which can then be used to improve search. Machine learning is classified as "*the development of digital systems that improve their performance on a given task over time through experience"* (Brundage et al. 2018). In this

[1] https://www.ed.ac.uk/information-services/research-support/research-data-service
[2] https://datadryad.org/
[3] https://www.data.cam.ac.uk/repository
[4] https://www.ukri.org/funding/information-for-award-holders/data-policy/

case, the task to improve, is the identification of research data for reuse and not the analysis of research data itself. Alongside avoiding replicating existing research by the reuse of data, there is also a strong drive to reproduce the results of prior research in areas such as the pharmaceutical and biomedical field to avoid basing future work on research that has not been validated (Pasquetto et al 2017).

This combination of research data, computational methods and machine learning, can be used to uncover data that would remain hidden, by allowing data to be discovered and then reused. The first step in reusing research data is to understand the data that is available.

Structuring data

Research data repositories allow researchers to find and download data for reuse. A dataset is typically described using a metadata schema such as the Datacite schema[5] or the Jisc RDSS data model[6] to aid search and description. There are numerous schemas available and the type of schema used can be specific to a subject, a research group, an organisation or the type of repository software used. This variety can cause issues for reuse. For example, the email address of the researcher who created the data, could be represented by 'email' or 'contact_email'[7], while the research data files could be described by file type, mime type or software application type. This choice presents the re-user with a problem and can be solved by mapping between the terms, but if the metadata field hasn't been filled in or isn't covered by the schema used, then the user will rely upon keywords for search. If these keywords don't cover the full range of a dataset, then the user may be unable to find the data to reuse. An example of this is MRI datasets within the University of Cambridge data repository, which may be stored in NIFTI format[8] or referenced in an Excel[9] or text file. Both methods are valid to the depositing researcher, but require a re-user to take additional steps to find the data. Therefore, a researcher may know that data should be available to them but simply be unable to find it, which is where machine learning can help.

For example, a dataset stored within a repository may have:

- A landing page
- A metadata record (in a specific schema)
- A JSON or XML record of the metadata
- Links to the data files themselves
- Links to other related data or information

This information allows a human to decide whether the dataset is useful to them, they are likely to make this judgement by reading the landing page information alone and not downloading and opening each file within a record. This means potentially valuable data can be overlooked.

The Nanowire system is designed to take unstructured text, such as web pages, Office documents, PDFs and files and then analyse them using Natural Language Processing and machine learning. It can be used to analyse the information around and a dataset and provide a richer picture of what the dataset is and contains. This analysis can then be stored in a metadata schema, which can then be searched and analysed to provide further detail. The key, is to store the automatically generated metadata in a general schema, which allows flexibility.

Schema.org provides such a general metadata schema for a wide variety of 'things'. Schema.org is a collaborative project that originated from search engine companies such as Google, Yahoo and Microsoft, but has morphed to include schemas outside of this realm and covers a very wide variety of 'things'. It is by no means exhaustive, but underpins the knowledge graphs of the large search engine companies and has

[5] https://schema.datacite.org/
[6] https://github.com/JiscRDSS/rdss-canonical-data-model
[7] http://data-archive.ac.uk/media/398085/rde_metadataprofile_public_02_00.pdf
[8] https://www.repository.cam.ac.uk/handle/1810/243411
[9] https://www.repository.cam.ac.uk/handle/1810/243427

been extended to describe general details about a dataset in a structured way, that is both machine and human readable. This latter statement is important; websites are now incentivised to publish a schema.org record in the source of the webpages to aid search engine results and the expected format is JSON-LD (JavaScript Object Notation for Linked Data). This simple text format is designed for developers and webmasters who have no prior metadata experience to encode data with minimal effort. Thereby making creating records and reusing records significantly simpler than XML based standards, but keeping the Subject-Predicate-Object model of RDF (Resource Description Framework[10]) and allowing extensions to describe specific items where required. For example, a data set may be described using the Dataset schema[11] and individual files described using DataDownload[12], which are lightweight non-subject specific schema designed to capture key details about a dataset. The dataset can be extended to capture general metadata such as its creator (a Person[13] entity), the organisation who created it (Organization[14]) and relevant location data (Place[15]). Further dataset specific metadata can be captured using extensions such as the MedicalImagingTechnique[16] schema to describe MRI data.

Machine learning

To extract value from research data, the metadata record and descriptive text can be augmented using machine learning and methods such as Natural Language Processing (NLP). The Nanowire system is designed to generate structured metadata from unstructured sources and has been applied to process research data landing pages. The aim of was to automatically extract metadata that can then be stored into the relevant schema.org schema, for subsequent search and reuse.

Table 1 illustrates potential sources of additional metadata that can be mined:

Source	Type	Techniques	Purpose
Dataset landing pages, metadata records and text based files within the data	Text	NLP/topic modelling/keywords	To uncover new keywords and descriptions to aid search
Landing pages, metadata records and text based files within the data	Text	Named entity recognition (NER)	To identify creators, organisations, places and locations that are mentioned in the text
Landing pages, metadata records and text based files within the data	Text	Classification and concept assignment using machine learning (e.g. Convolutional Neural Networks).	To classify research data into broad categories and concepts.
Images within the dataset	Images	Image classification and recognition using machine learning	To uncover what images, contain e.g. a classifier can be trained to identify MRI images

[10] https://www.w3.org/TR/rdf11-new/
[11] http://schema.org/Dataset
[12] http://schema.org/DataDownload
[13] http://schema.org/Person
[14] http://schema.org/Organization
[15] http://schema.org/Place
[16] https://health-lifesci.schema.org/MedicalImagingTechnique

Data files within the dataset	Data files	MIME-type identification and metadata extraction	To uncover files in the dataset that may not be described within the metadata record
Audio files within the dataset	Audio	Automatic transcription using machine learning	To allow text search across audio files that are not held within the metadata record
URLs within the landing page and data files	URLs	Web scraping, then NLP/NER etc	To uncover what resources are linked to the dataset and augment the dataset record

Table 1: Sources of metadata

Hence a simple landing page record and the data files themselves can yield a large amount of additional metadata, to then identify data for reuse. Advances in machine learning mean that images can be classified and features identified within them, such as people, faces or objects using libraries such as TensorFlow[17]. Audio can also be translated using services such as Amazon Translate[18] to convert speech to text, that can then be analysed using NLP and NER to further add to the descriptive metadata. Meister's (2017) research suggests that schema.org can also be used to capture a knowledge graph about scholarly articles, thereby raising the opportunity to link both dataset and scholarly articles using the same schema.

Conclusion

Schema.org provides a generic framework with which to store metadata about a dataset, with the ability to extend it for subject specific datasets. This then provides a stable base which can be populated with metadata that is automatically extracted from datasets and landing pages using machine learning and NLP. These steps, can then be used to update the existing metadata landing pages with a greater depth of information or as in the case of Nanowire, to provide an overarching search facility across multiple data repositories and systems. Further analysis work, can then be carried out to find relationships between datasets using knowledge graphs, clustering and graph databases such as Neo4j[19].

References

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. https://doi.org/10.17863/CAM.22520

Meister, V. (2017). Towards a Knowledge Graph for a Research Group with Focus on Qualitative Analysis of Scholarly Papers, *Proceedings of the First Workshop on Enabling Open Semantic Science co-located with 16th International Semantic Web Conference*. http://ceur-ws.org/Vol-1931/

Pasquetto, I, Randles, B. and Borgman, C. (2017). On the Reuse of Scientific Data. Data *Science Journal, 16: 8*, pp. 1–9, DOI: https://doi.org/10.5334/dsj2017-008

[17] https://www.tensorflow.org/
[18] https://aws.amazon.com/translate/
[19] https://neo4j.com/

MASER: A Toolbox for Low Frequency Radio Astronomy

Baptiste Cecconi[1,2], Pierre Le Sidaner[3], Renaud Savalle[3], Xavier Bonnin[1], Philippe Zarka[1,2], Corentin Louis[1], Andrée Coffre[2], Stéphane Aicardi[3], Laurent Lamy[1,2], Laurent Denis[2], Jean-Mathias Grießmeier[4], Jeremy Faden[5], Chris Piker[5], Nicolas André[6], Vincent Génot[6], Stéphane Erard[1], Joseph N Mafi[7], Todd A King[7], Mark Sharlow[7], Jim Sky[8], Markus Demleitner[9]

[1]LESIA, Observatoire de Paris, CNRS, PSL, Sorbonne Université, Meudon, France, [2]Station de Radioastronomie de Nançay, Observatoire de Paris, CNRS, PSL, Université d'Orléans, Nançay, France, [3]DIO, Observatoire de Paris, CNRS, PSL, Paris, France. [4]LPC2E, CNRS, Université d'Orléans, Orléans, France.

[5]Dep. Physics and Astronomy, University of Iowa, Iowa City, Iowa, USA. [6]IRAP, CNRS, Université Paul

Sabatier, Toulouse, France. [7]IGPP, UCLA, Los Angeles, California, USA. [8]Radio Sky Publishing, USA.

[9]Heidelberg Universität, Heidelberg, Germany.

Corresponding author: Baptiste Cecconi (baptiste.cecconi@observatoiredeparis.psl.eu)

The MASER (Measurements, Analysis, and Simulation of Emission in the Radio range) project provides a comprehensive infrastructure dedicated to low frequency radio emissions (typically < 50 to 100 MHz). The four main radio sources observed in this frequency are the Earth, the Sun, Jupiter and Saturn. They are observed either from ground (down to 10 MHz) or from space.
Ground observatories are more sensitive than space observatories and capture high resolution data streams (up to a few TB per day for modern instruments). Conversely, space-borne instruments can observe below the ionospheric cut-off (10 MHz) and can be placed closer to the studied object. Several tools have been developed in the last decade for sharing space physics data. Data visualization tools developed by The CDPP (http://cdpp.eu, Centre de Données de la Physique des Plasmas, in Toulouse, France) and the University of Iowa (Autoplot, http://autoplot.org) are available to display and analyze space physics time series and spectrograms.
Other tools include ExPRES (Exoplanetary and Planetary Radio Emission Simulator) developed at LESIA. The VESPA (Virtual European Solar and Planetary Access) which provides a search interface that allows the discovery of data of interest for scientific users, and is based on IVOA standards (astronomical International Virtual Observatory Alliance). The University of Iowa has developed the Das2 server that allows the distribution of data with adjustable temporal resolution.
MASER is making use of all these tools and standards to distribute datasets from space and ground radio instruments available from the Observatoire de Paris, the Station de Radioastronomie de Nancay and the CDPP deep archive. These datasets include Cassini/RPWS, STEREO/Waves, WIND/Waves, Ulysses/URAP, ISEE3/SBH, Voyager/PRA, Nancay Decameter Array (Routine, NewRoutine, JunoN), RadioJove archive, swedish Viking mission, Inter- ball/POLRAD. MASER also includes a Python software library for reading raw data.

Keywords: Radio astronomy, Tools, Protocols, Interoperability

Introduction

Low frequency radio data are providing remote proxies to study remotely energetic and unstable magnetised plasmas. In the solar system, all magnetized plasma environments are emitting radiation over the full radio frequency range. These radio sources are non-thermal emission phenomena, and are not related to atomic and molecular transitions contrarily to electromagnetic emissions at higher frequencies. The "low frequency" radio emissions are observed in the standard VLF (~3 kHz) to VHF (~30MHz) radio bands. The main radio sources of the solar system are the Sun, Jupiter and Saturn. The Earth, Uranus and Neptune are also hosting natural radio emissions. The planetary radio emissions are linked to the magnetospheric dynamics (i.e., auroral activity, radiation belts, etc.) [1].
The usual data product for low frequency radio emissions observations is "dynamic spectra" (a time varying spectrogram). In this frequency range, it is not yet possible to build imaging radio telescopes, so that the main source of knowledge is this time-frequency representation of the data. Until recently each low

frequency data provider was storing their data products in local formats, or using standard formats with local metadata dictionaries. This prevented interoperability. The NASA space physics community has promoted the CDF [2] format with ISTP [3] guidelines, for day to day usage and archiving. NASA's Planetary Data System (PDS) archive is now accepting CDF/ISTP as an archive format [4], and many space mission teams have adopted the same scheme. Ground based observatories are producing data collections reaching several TB per day [5]. Even with a common file format, it is impossible to download large series of data (even 1 day of data) for local processing. Solutions for space physic data streaming have been developed by University of Iowa and is fitted for low frequency data. The virtual observatory infrastructure is proposing a way to share, discover and retrieve data files with scientific search parameters, e.g., using VESPA (Virtual European Solar and Planetary Access) for solar system sciences [6].

Maser4py library

The MASER (Measurements, Analysis, and Simulation of Emission in the Radio range) library started with development of ExPRES (Exoplanetary and Planetary Radio Emission Simulator) [7], and the Cassini-RPWS Kronos database. Recently, the Solar Orbiter/RPW (Radio and Plasma Waves) instrument science ground segment proposed to join efforts and initiated two community repositories: maser4py (https://github.com/maserlib/maser4py/) and maser4idl (https://github.com/maserlib/maser4idl). They host respectively Python and IDL libraries and software related to low frequency radio science and processing. The code is distributed under a GPLv3 license. Various libraries are currently being transferred (or planned for transfer) into those repositories.

The current main developments on the *maser4py* library concerns data reading modules for legacy and non-standard format radio data collections. With various levels of development, it currently includes data modules for data collections hosted or produced by LESIA (Cassini/RPWS, Voyager/PRA, Solar Orbiter/RPW), by the CDPP [7] (Demeter, Interball, Viking (Swedish auroral mission), ISEE3, Wind), by the PDS/PPI node [8] (Cassini/RPWS, Voyager/PRA), by the Nançay radio telescopes (Nançay Decameter Array, NenuFAR), as well as by the radio amateur RadioJOVE project.

Das2 server interfaces

The Das2 server [9] server interface provides a very efficient way to serve and stream data with adaptive temporal resolution, together with its main client, Autoplot [10,11]. Thanks to the das2 server architecture, setting up data services (so-called "data sources") is simple and straightforward. The process includes writing 2 files: a Data Source Definition File (DSDF), which defines the data source metadata (description, reader script, valid time range, cache level); and a data reader script, which has to write das2stream formatted to standard output for a given input time interval. The das2stream format is documented in its Interface Control Document (ICD) [12].

The MASER team has currently set up two das2 server for distributing LESIA (http://voparis- maser-das.obspm.fr/das2/server) and Nançay (https://das2server.obs-nancay.fr/das2/server) datasets. Three extra servers (for CDPP, LPC2E [Demeter and Taranis], and JAXA [Kaguya] datasets) are being drafted. Data readers are using the maser4py library for reading the data from file.

VESPA interface

VESPA is providing a data discovery framework with well-defined metadata dictionaries, query protocols and a registry of services. Each VESPA service consists in a metadata table following the EPN-TAP [6] dictionary. Each row contains the metadata corresponding to a single product, including a data access URL. The VESPA services respond to TAP protocol [13].

MASER teams are now sharing data files (Raw or CDF formats), and, in the near future, das2 server data source endpoints, as well as Autoplot template (.vap) files. The MASER VESPA services are built on DaCHS [14], and the tables are fed directly reading from the CDF headers or the DSDF files. The VESPA main query portal (http://vespa.obspm.fr) is including capabilities to interact directly with Autoplot (with the SAMP protocol [15]).

Simulation

The ExPRES project [16] is now producing routine simulation data in support to the Juno mission, as daily files of spectrograms of simulated radio emissions induced by the Galilean satellites, for various observatory locations (Juno, Earth, STEREO-A). The simulated data include the time-frequency portions

during which the modeled radio sources of each hemisphere are visible, the 3D locus of the visible radio sources in the Jovian magnetosphere and other intrinsic modelled parameters. A public online interface for ExPRES is under development. The ray tracing code ARTEMIS-P [17] is also aimed at being distributed through MASER.

Applications

The usage of das2 server interface with Autoplot is a huge improvement in data analysis and processing for the MASER teams. Three examples can be cited: the refurbishment of Voyager/PRA data [18] has been consolidated by the das2 server/Autoplot setup, allowing efficient and fast data browsing at all temporal scales; distribution of low frequency data sets together with space observations [5]; and, the NenuFAR instrument is in commission phase in Nançay, and the team is now using das2 server/Autoplot setup for daily routine data checks.

The Juno-Ground-Radio team is gathering ground-based radio observatories from all over the world (France [lead], USA, Ukraine, Japan, Poland…) and provides data supporting the Juno science team. The data files are now distributed through VESPA, using CDF files formats when possible. Das2 server interfaces are under study for Ukrainian and Polish teams.

Future steps

New data formats and collection compatibility will be included in the maser4py library in the future. The MASER team will add new readers and tools, and will also reach out to the community for participation. This requires a consolidation of interfaces (classes and methods) and tests.

The ExPRES simulations are now used by the Juno team for Juno/Waves instrument. Recently, the need for a similar radio ground support scheme has been identified by the Solar Orbiter and Parker Solar Probe teams. The MASER tools will be proposed to the solar radio ground observatories.

Finally, there is a growing need for large scale coordination of open source library and software (especially for python-based developments). Several groups are pushing for this, and the MASER team will participate to this effort (e.g., http://openplanetary.co for planetary sciences).

Acknowledgments

The Europlanet H2020 Research Infrastructure project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654208.

References

[1] Zarka, Philippe. 1998. "Auroral Radio Emissions at the Outer Planets: Observations and Theories." *J. Geophys. Res.* 103: 20159–94.
[2] Common Data Format (CDF) http://cdf.gsfc.nasa.gov
[3] International Solar Terrestrial Program (ISTP) https://spdf.gsfc.nasa.gov/istp_guide/istp_guide.html
[4] PDS4/CDF Specification (2015) https://pds-ppi.igpp.ucla.edu/doc/CDF-A-Specification-v1.0.pdf
[5] Lamy, L, P Zarka, B Cecconi, K-L Klein, S Masson, L Denis, A Coffre, and C Dumez-Viou (2017). "1977– 2017: 40 Years of Decametric Observations of Jupiter and the Sun with the Nançay Decameter Array." *Planetary, Solar and Heliospheric Radio Emissions (PRE VIII).*
[6] Erard, S., B Cecconi, P Le Sidaner, A Pio Rossi, M T Capria, B Schmitt, V Génot, et al. 2017. "VESPA: a Community-Driven Virtual Observatory in Planetary Science." *Planet. Space Sci.* doi:10.1016/j.pss.2017.05.013.
[7] Centre de Données de la Physique des Plasma (CDPP) – Plasma Physics Data Centre. http://cdpp.eu
[8] NASA/PDS (Planetary Data System) PPI (Planetary Plasma Interactions). https://pds-ppi.igpp.ucla.edu
[9] Das2 Server, University of Iowa. http://das2.org
[10] Autoplot, University of Iowa. http://autoplot.org
[11] Faden, J, R S Weigel, J Merka, and R H W Friedel. 2010. "Autoplot: a Browser for Scientific Data on the Web." *Earth Sci. Inform.* 3: 41–49. doi:10.1007/s12145-010-0049-0.

[12] Piker, C, et al. Das2 server ICD, University of Iowa. http://das2.org/Das2.2.2-ICD_2017-05-09.pdf
[13] Table Access Protocol (TAP). http://www.ivoa.net/documents/TAP/
[14] Data Centre Helper Suite (DaCHS). http://dachs-doc.readthedocs.io
[15] Simple Application Messaging Protocol (SAMP) http://www.ivoa.net/documents/SAMP/
[16] Louis, C K, L Lamy, P Zarka, B Cecconi, M Imai, W S Kurth, G Hospodarsky, et al. 2017. “Io-Jupiter Decametric Arcs Observed by Juno/Waves Compared to ExPRES Simulations.” *Geophys. Res. Lett.*, 1–17. doi:10.1002/2017GL073036.
[17] Gautier, A-L, B Cecconi, and P Zarka. 2013. “ARTEMIS-P: a General Ray Tracing Code in Anisotropic Plasma for Radioastronomical Applications..” *Proceedings of the 2013 International Symposium on Electromagnetic Theory"*.
[18] Cecconi, B, A Pruvot, L Lamy, P Zarka, C Louis, S L G Hess, D.~R. Evans, and D Boucon. 2017. “Refurbishing Voyager 1 & 2 Planetary Radio Astronomy Data.” *Planetary Radio Emissions VIII*.

Enhancing access to environmental research data for a wider user community

Matthew James Fry
Centre for Ecology and Hydrology, Wallingford, UK
mfry@ceh.ac.uk

Data outputs from hydrological research have a potentially wide audience in both research and non-research communities. The UK Centre for Ecology and Hydrology (CEH) produces hydro-meteorological datasets that are useful for a wide range of users and applications. Factors such as discoverability, dataset complexity, data volume, and availability of information about dataset contents can result in datasets that are hard to use, meaning the potential impacts of the datasets are often not fully realised. Some users do not have the technical skills to make use of large and complex research datasets and require additional mechanisms to enable access to the data they need. This paper describes the range of users of this data and their requirements, and demonstrates how CEH is meeting diverse user needs through the development of data infrastructures and bespoke applications to enhance access to some of the key datasets within its data centres.

Background to the CEH data centres and data holdings

The Centre for Ecology and Hydrology (CEH) hosts the Environmental Information Data Centre (EIDC), the UK's Terrestrial and Freshwater Sciences data centre. Typical of a research data centre, the contents comprise datasets across a range of scales, from single experiments and field studies to larger campaigns and derived data products with national and international scale coverages. The core user functionality is also typical of such data centres: a metadata catalogue with an interface for searching discovery level metadata, provision of Digital Object Identifiers for long-term citation of datasets, use of supporting documentation to describe datasets in more detail and the processes by which they were created, and bulk file download facilities for those interested in accessing datasets. File formats are generally simple text (csv), common GIS formats, and netCDF used for larger gridded datasets. This functionality meets the essential requirements of both the funding research council (provision of long-term access to datasets from funded research for scientists) and the research user (access to datasets for further research).

The need for wider access to research outputs

Many of the national scale datasets produced by CEH and stored within the EIDC are of wider interest beyond the research community. For example, the National River Flow Archive is mandated by UK government (Defra) and the devolved administrations of Northern Ireland, Scotland and Wales to maintain and provide access to river flow data and provide information on water resources nationally. Access to data is provided via its website, where additional information on data quality, utility, and catchment statistics such as land use and geology, is provided alongside interactive maps and graphs, to ensure the data is as usable as possible. Information on who these users are is gathered at the point of data access: users are required to identify the type of institution they are affiliated to prior to downloading data from the NRFA website. A breakdown of the download statistics for 2017 is provided in Figure 1. Whilst "University" users download most data (72%, 32000 downloads), a significant proportion are from companies (12%, ~5000), the public sector (3%, ~1500) or describe themselves as "individuals" (2%, 1000). This breakdown is believed to be indicative of the range of users of other national scale hydro-meteorological datasets stored within the EIDC, though evidence for this is more ad hoc, based on data requests and direct communication with data users, e.g. through engagement at industry events and conferences.

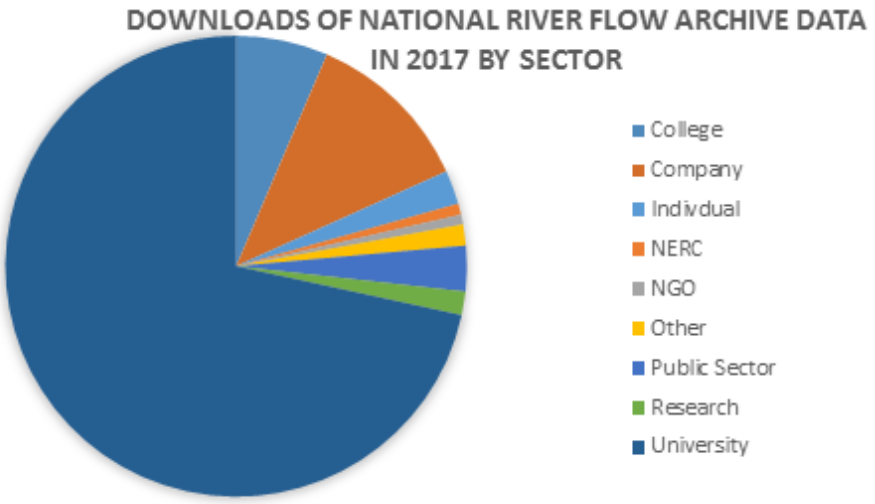


Figure 1. Breakdown of users of the National River Flow Archive

Data access requirements of non-research users

Users of this hydro-meteorological data have a wide range of data access requirements, depending on their specific need for the data and their level of technical expertise. Many users are interested in data for a specific location and time, while others require data over the entire country for longer periods. Some users want to process raw data, whereas others require only metadata, derived statistics or a visualisation of a dataset. Whilst many research users and consultants are increasingly able to write computer programs and software to access and process data, others require the use of software such as Excel for analysis, and cannot automate processing of, or access to, data.

A current Knowledge Exchange project for the UK Drought and Water Scarcity Research Programme is aiming to make a number of hydro-meteorological datasets available and accessible to diverse stakeholders. Many of these datasets are large (up to TB in size) and data access methods are being developed to allow the data to be used appropriately. Extensive user-engagement has been undertaken with users from different sectors. A recent knowledge exchange showcase was attended by over 120 people from a range of organisations, including water companies, UK regulating agencies, hydrological consultancies, as well as researchers. Attendees with an interest in datasets were shown a number of data access options, and surveyed to identify preferences for data access, voting electronically to identify each option as “Not useful”, “Quite useful” or “Very useful”.

Survey options provided were:

- Just download all of the data
- Programmatic access over the web to extract the data I need (i.e. API access)
- Virtual Research Environments for advanced analysis
- Simple web interface to select location of interest and download data in simple formats
- More complex applications to dig into and understand the data

Results are shown in Figure 2. Whilst a few users would like to download all of the data, many more were interested in programmatic API access and simple access via web interfaces.

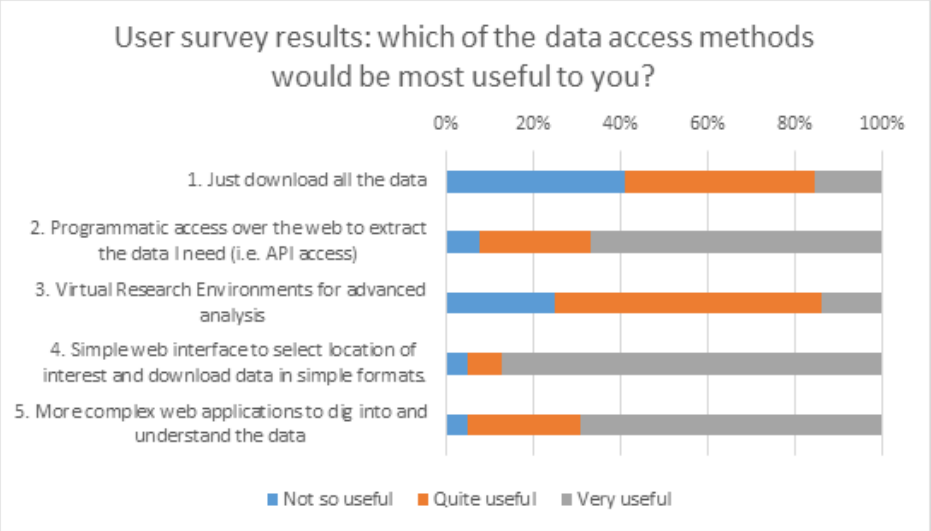


Figure 2: summary results of a survey of user preferences for data access methods.

The CEH Environmental Information Platform

To meet these user needs, CEH will be extending its existing platform for data access. CEH created its Environmental Information Platform to provide enhanced access to CEH's key data holdings via web-based tools and programming interfaces. It contains a number of applications relating to different datasets, and aimed specifically at users of each dataset, some of which are described below.

UK Rainfall portal. The UK Rainfall Portal (CEH, 2016) allows users to explore, visualise and extract data from the CEH Gridded Estimates of Areal Rainfall (CEH-GEAR) dataset (Tanguy et al, 2016). This is a large (~650GB) dataset of daily rainfall at a 1km grid scale for the UK from 1890 onwards, stored in netCDF format. Many users are interested in the rainfall at a particular location, or in viewing spatial variation of rainfall for a given day or month, and therefore do not want to download the entire dataset. Many non-research users are still not familiar with the netCDF format or do not have programming skills required to make use of data in this format. The UK Rainfall application therefore makes the data more accessible, allowing users to view maps and graphs of the data and to download subsets of data in csv formats. The application is supported by the THREDDS Data Server (TDS) software stack (Unidata, 2015), which provides web-accessible end-points to netCDF data stores, such as Web Map Services. The TDS system reads directly from the data files stored within the data centre, meaning users are accessing the same long-term archive as those downloading data. The TDS system also provides programmatic end-points using OpenDAP, and a number of users making use the CEH-GEAR data, including from water companies and consultants, do so using this API.

UK Droughts portal. The UK Droughts Portal (CEH, 2015) makes use of similar technologies to the UK Rainfall portal. The drought data made available are updated in near-real time, at the beginning of each month to reflect rainfall deficits for the previous months over river basins, and at a 5km grid-scale, over the UK. The application enables users to view both the spatial distribution of drought and graphs showing changes in drought statistics over time. The UK Drought portal allows researchers, regulators and hydrologists to undertake drought and water situation monitoring without the need to download the underpinning data. For this application the TDS system aggregates an archived dataset from the data centre with more recent provisional data.

UK Lakes. CEH hosts the UK Lakes database of over 40000 lakes across the UK, which includes information on physical parameters (perimeters, areas, catchment areas), typology (by altitude, geology, etc.), chemistry, and catchment land cover. The UK Lakes Portal (CEH, 2016) allows users to search for a lake and view the data held, via maps and charts. The portal also integrates data from the National Biodiversity Network, retrieving biological recordings that overlap spatially with the lake area for each lake and allowing it to be readily filtered by species, etc., and even allows the addition of new observations.

Shiny applications for advanced data analysis

Shiny is a software library that enables building of web applications in the R statistical programming language. The widespread use of and familiarity with R within the scientific community means it provides a relatively simple means for scientists to realise interactive web-based analysis of their own data. Within the UK Drought and Water Scarcity Research Programme a number of Shiny applications have been developed to provide access to detailed modelling results. The Historic Droughts "UK Reconstructed Flow Data Explorer" (CEH, 2018) enables users to visualise results from the modelling of over 300 UK catchments, comparing the quality of the modelling between catchments, as well as spatially. This information helps users of the modelled data to understand the data prior to downloading the full dataset from the EIDC. The "UK Hydrological Drought Explorer" (CEH, 2018) allows users to use modelled flow data to understand past droughts based on given drought thresholds and durations, producing tables of droughts ranked by importance, and graphs of past drought severity. Such applications bring large and complicated datasets to life, and provide bespoke analysis functionality for users who may not be able to undertake such analyses themselves.

Summary and Future work

A wide range of users, with varying requirements, need to access hydro-meteorological data from the CEH Environmental Information Data Centre. The CEH Environmental Information Platform provides infrastructure and applications to enable this data access. Through these tools CEH has increased the access to these datasets and improved the capability of wider sectors within the UK environmental space, with users within environmental regulators, water companies and environmental consultancies. Work currently underway aims to extend the use of these applications to further enhance data access, improving the speed of data access from the TDS system and providing interfaces to allow users to upload shapefiles of data and extract subsets of data from large gridded datasets, and increasing the number of datasets available via documented APIs.

References

CEH. 2015. The UK Droughts Portal. Available at https://eip.ceh.ac.uk/droughts

CEH. 2016. The UK Rainfall Portal. Available at https://eip.ceh.ac.uk/rainfall

CEH. 2016. The UK Lakes Portal. Available at: https://eip.ceh.ac.uk/apps/lakes/

CEH. 2018. UK Reconstructed Flow Data Explorer. Available at https://shiny-apps.ceh.ac.uk/reconstruction_explorer/

CEH. 2018. The UK Hydrological Drought Explorer. Available at https://shiny-apps.ceh.ac.uk/hydro_drought_explorer/

Tanguy, M.; Dixon, H.; Prosdocimi, I.; Morris, D. G.; Keller, V. D. J. (2016). Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2015) [CEH-GEAR]. NERC Environmental Information Data Centre. https://doi.org/10.5285/33604ea0-c238-4488-813d-0ad9ab7c51ca

Unidata, (2015): THREDDS Data Server (TDS) version 4.6.2 [software]. Boulder, CO: UCAR/Unidata. (http://doi.org/10.5065/D6N014KG)

Policy, Infrastructure, Skills and Incentives Driving African Data Sharing: The African Open Science Platform Project

Ina SMITH[1], Tshiamo MOTSHEGWA[2], Susan VELDSMAN[3]
[1,3]African Open Science Platform, Academy of Science of South Africa (ASSAf), Persequor Park, Meiring Naudé Road, Lynnwood 0020, Pretoria, South Africa
Email: ina@assaf.org.za, susan@assaf.org.za
[2]Computer Science, University of Botswana, 4775 Notwane Road, Gaborone, Botswana
Email: tshiamo.motshegwa@mopipi.ub.bw

The Science International Accord on Open Data in a Big Data World presents an inclusive vision of the need for and the benefits of Open Data for science internationally and in particular for Lower and Middle Income Countries. In addition to benefiting from Data from the international community, African countries have much to contribute in terms of Data for all to benefit in making progress towards implementing the United Nations Development Programme (UNDP) 2030 Sustainable Development Goals (SDGs). Providing a comprehensive view of what is happening on National and continental level, will not only assist fellow researchers and potential funders in identifying gaps – it will also assist African countries identify opportunities for inter-regional links to strengthen collaboration, and towards international alignment with science activities on the international level. The African Open Science Platform (AOSP) initiative is an important outcome of the Accord on Open Data in a Big Data World, and through it, great progress has been made towards a better understanding of the following: - buy-in and support from countries and institutions through data policies; capacity building and developing skills; how sharing of data by researchers is rewarded (incentives); and current infrastructures that exist to support the sharing of data.
This paper introduces the AOSP project, discusses its alignment with the SDGs, and presents some outcomes of the project to date.
Keywords – Policy, Open Data, Open Science, SDGs, Africa

1. Introduction

1.1 Background

Open Data (Kitchin, 2014) has been demonstrated nationally and internationally to be an enabler for exploiting opportunities with benefits to economies and societies (Science-International, 2015). It is also widely accepted that Open Data is a basic prerequisite and key requirement for maintaining acceptable quality and standards of scientific rigor through facilitating reproducibility of results. To accrue the benefits of open data, it requires infrastructure, open access, skills and capacity. Countries are responding to this dispensation and are developing infrastructure, national systems for funding, and research management plans to address challenges. There is however – as it is with the digital divide - a real risk that least developed countries and their poorly resourced national research systems, may fail to respond to these opportunities, which are vital to the attainment of Sustainable Development Goals and the gains through the Digital revolution (AOSP, 2016). Furthermore, an increasing number of governments around the world are defining and implementing "open data" strategies in order to increase transparency, participation and/or government efficiency (Huijboom & Van den Broek , 2011).

1.2 Sustainable Development Goals

(Robert, et al., 2005) provides a detailed discussion on what constitutes Sustainable Development – highlighting the necessary goals, indicators, values, and desirable practices. Furthermore, (Sachs & Reid, 2006) discusses desirable investments towards sustainable development, and proposed *an "investment strategy for sustainable development in low-income countries"* – a strategy that incorporates investing in the interdisciplinary science of sustainable development, and that called for a "*Millennium Ecosystem Fund to give poor countries the wherewithal to incorporate environmental sustainability into national development strategies"*

(Griggs, et al., 2013) argued that planetary stability must be integrated with United Nations targets to fight poverty and secure human well-being. They proposed a Unified Framework with a set of six sustainable development goals (SDGs) following from combining the Millennium Development Goals (MDGs) (UN, 2012) with conditions necessary to assure the stability of Earth's systems.

In 2015, through a General Assembly resolution on 25 September 2015, (UN, 2015), Resolution 70/1 - *Transforming our world: the 2030 Agenda for Sustainable Development*, United Nations member

states adopted a set of goals to end poverty, protect the planet and ensure prosperity for all as part of a new sustainable development agenda. The Resolution identified a set of 17 goals and stated that the goals and targets were to stimulate action in the period 2015-2030 in the areas of critical importance for humanity and the planet – citing people, planet, prosperity, peace and partnerships themes. The 17 goals span a spectrum ranging from - ending poverty in all its forms everywhere; taking urgent action to combat climate change and its impacts; building resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation through to strengthening the means of implementation and revitalize the global partnership for sustainable development, to name a few.

The role of Data in furthering Sustainable Development Goals has received considerable attention in the literature. (Gijzen, 2013) for example discussed Big Data (Khan, et al., 2014; Chen, et al., 2014) for a sustainable future. (Griggs, et al., 2013) also further opined that ;

> "*We should be collecting big data that can be used to model and test an array of different scenarios for sustainably transforming the production and consumption of energy, improving food and water security, and eradicating poverty. Managing these issues will also help to rebalance important biogeochemical cycles (especially the carbon, nitrogen and phosphorus cycles), mitigate climate change, reverse ocean acidification and reduce the loss of biodiversity. Big data will help to illuminate the origins, nature and scale of these challenges, and how they relate to one another. National databases and research centres can be linked to create huge databases. Initiatives similar to those of the Intergovernmental Panel on Climate Change and the Global Ocean Observing System could fill the gaps in scientific, technical and socio-economic data.*"

The role of information and communications technologies (ICTs) on contributing to achievement of the SDGs has also been addressed by various authors. (Sachs, et al., n.d.) discussed in detail the role of ICTs in achieving SDGs, (Bothwell, et al., 2015) identified technology trends, opportunities and innovative case studies that global leaders can harness as they begin to strategize on how to implement the SDGs and finally (Batchelor, et al., 2003) shared lessons learned from seventeen information development projects contributing to the goals.

This paper will provide a general overview of sustainable development goals, and then discuss the methodology adopted in the AOSP project – including an overview of the landscape study on data intensive activities in the African continent and its key findings. In addition, the paper will discuss the draft frameworks proposed for each of the focus areas: policy, infrastructure, capacity building and incentives in the implementation of the project.

1.3 About the African Open Science Platform

The African Open Science Platform (AOSP) – is an outcome of the international accord on Open Data in a Big Data World. The project is managed by the Academy of Science of South Africa (ASSAf) with input from the International Council for Science Regional Office for Africa (ICSU ROA) hosted by ASSAf. The project is funded through the South African National Research Foundation (under the Department of Science and Technology), in collaboration with the International Council for Science (ICSU) Committee on Data for Science and Technology (CODATA).

The project aims to develop an open science and innovation dialogue platform in order to increase awareness, accessibility and visibility of African science and data, at the same time reflecting on progress made on the African continent in terms of the following areas - Open science/data policy and strategy; Open science/data information technology (IT) infrastructure; Capacity building/training to support open science/data; and Incentives for sharing science output and specifically the underlying data sets, in an open and transparent way.

At the end of the three year period, the project aims to have tangible outcomes in terms of the above, which is expected to include:

1. A framework for each of the mentioned focus areas, to promote the development and coordination of data policies, data training, data incentives and data infrastructure;
2. A database with information on open science (incl. open data) initiatives and key role players/experts in open science, across all disciplines;
3. A policy toolkit guiding governments on implementing open science policies;
4. An advocacy toolkit to assist representatives from various African countries to create more awareness of the benefits of open science;
5. A competency index for all working with data; and

6. A training toolkit, building on existing resources.

AOSP was launched in December 2016. Year 1 of the project focused on creating awareness and conducting an Open Data landscape survey, while Year 2 – which commenced on 1 November 2017 – is focusing on building capacity and delivering training workshops. Initiatives indicated above are captured in a database and will inform future recommendations. This platform is expected to enhance accessibility to and increase the impact of African science, and specifically the data sets underlying the final research output. Table one[20] shows the project schedule, activities and expected outcomes.

2. Methodology

The first phase of the AOSP project involved a continental data initiatives landscape analysis. This was done through engaging in many meetings, workshops, presentations by experts, desktop research and literature reviews. An African Open Science Platform landscape study was conducted to compliment these efforts. The survey instrument interrogated respondents to solicit information about data activities and some dimensions. Details are given in Table two[21].

3. Results

An annotated map of the continent visualizing the initial data on coverage and distribution of data initiatives in the continent is given[22]. Table three[23] provides an overview of the findings of the landscape analysis - It shows and highlights some selected initiatives in the AOSP frameworks of policy, infrastructure, capacity building and incentives.

Furthermore regarding infrastructure, there has been a lot of progress in network connectivity through development of National Research and Education Networks (NREN) and regional network of NRENs. An NREN is a specialised internet service provider dedicated to supporting the needs of the research and education communities within a country. It is usually distinguished by support for a high-speed backbone network that provides connectivity to the necessary cyber-infrastructure and its platforms - science gateways, data repositories and research clouds etc., the World Bank has published a report on *"The role and status of African National Research and Education Networ*ks" (Foley, 2016) that gives a detailed account of progress in the development of African NRENs, against the NREN Capability Maturity Model (CMM). There is a wide ranging capability - 9 African NRENs qualify as Level 6 (mature) NRENs. The following NRENs already allow for running data-intensive applications and sharing of high end computing assets, bio-modelling and computation: KENET (Kenya), TENET (South Africa), RENU (Uganda), and ZAMREM (Zambia) (AAS, 2016). Great progress has also been made in Algeria, Egypt, Kenya, Morocco, Senegal, Tunisia, South Africa, Uganda, and Zambia, connecting universities and research institutions.

However, the challenge frequently cited is the cost of data - being extremely high in some African countries – especially towards the South and in landlocked countries (Kunda & Khunga, 2015). There has been support from the European Union for these NRENS through the AfricaConnect programme on a cost-share basis and also through EUMEDCONNECT for North African NRENS to help develop NRENs.

3.1 Some Selected Projects

There are significant number of data (and compute) intensive projects and initiatives of National, regional and international consequence hosted or driven in Africa – including in the areas of earth and spaces sciences – for example, The Square Kilometre Array (Dewdeney, et al., 2009). In this area, Africa will host the largest high resolution radio telescope to do fundamental Astrophysiscs science (SKA-Organisation, 2015) as the continent has access to the rich Southern hemisphere skies, and quiet zones with respect to minimal radio frequency interference. For this reason - South Africa and SKA African partner countries are investing in space science and astronomy as building blocks for a knowledge economy through developing necessary supporting platforms. Data challenges for next generation telescopes including the SKA have been widely discussed (Norris, 2010).

[20] https://drive.google.com/file/d/1QdSefOdhVY_BsF7J2Y3PpnR2z1koV_b4/view?usp=sharing
[21] https://drive.google.com/file/d/1QdSefOdhVY_BsF7J2Y3PpnR2z1koV_b4/view?usp=sharing
[22] https://www.targetmap.com/viewer.aspx?reportId=56245
[23] https://drive.google.com/file/d/1QdSefOdhVY_BsF7J2Y3PpnR2z1koV_b4/view?usp=sharing

In Genomics and Human Heridity, there is the H3Africa (Rotimi, 2014) and its network - H3ABioNET (Mulder, et al., 2016), where the continent is making gains post the human genome project to take advantage of advances in genomics to help bridge the gaps and aleviate the glaring lack of research in these areas despite the heavy burden of Non- Commucable Diseases (NCDs) facing the continent that can be addressed by genomics research. (Mulder, et al., 2017) provides a detailed discussion of challenges in genomic research data generation, analysis and sharing in the African setting. The full H3Africa policy framework, negotiating fairness in genomics is discussed in (de Vries, et al., 2015).

There are also signifcant projects in Earth observation, enviroment, climate and weather modelling and applications therein in early warning systems and applications including in agrometeorology and hydrometeorology. For example, (SADC-CSC, 2017) details a project funded by the African Development Bank that recognises the need to generate, disseminate and use reliable and high quality climate information to help decrease the negative impacts of extreme weather and climate related phenomena and risks in the Southern Africa Development Community (SADC) region.

Collectively, all these projects will necessitate and facilitate the creation of requisite computational, network and data infrastructure through regional cyber-infrastructures; facilitate collaborative networks; stimulate discussions around data policy frameworks; address issues around data generation, data analysis, data sharing challenges and negotiating fairness in these domains. For the SKA readiness, the Southern African Development Community countries are developing an HPC ecosystem and the SADC Cyberinfrastructure (Motshegwa, et al., 2018) through a SADC Cyber-infrastructure Framework regional policy instrument (SADC, 2016). For H3Africa, a H3ABioNet, a bioinformatics network and infrastructure which aim to build capacity for large-scale genomics projects in Africa are being developed (Mulder, et al., 2016).

Higher up the value chain, through these projects, it is envisaged they will provide necessary impetus in creating new cohorts of scientists, engineers, researchers and technicians ready to work on world-class projects. This will drive the human capital development programme including in transferable skills such as in data science and as consequence, contribute to stemming the historical brain drain away from the continent and spillovers into countries' economies.

As a result of all this activity, there is an increasing need for governments to put in place the necessary legislation, policies and governance structures to facilitate data sharing and its exploitation for research, innovation and for development. In addition to policy, there is awareness for the need of substantial investments in reliable ICT infrastructure that can sufficiently support data hosting, sharing and analysis. Furthermore, for researchers to generate, curate, manage and preserve and analyse data, they also need specialised skill sets and requisite training. Finally, where there are no internal motivation, governments and funders need to find ways to incentivise institutions and researchers to embrace a data sharing culture paradigm shift regarding Open and FAIR data (Wilkinson, et al., 2016), all this in the midst of current impediments including costs of generation, storing data and effort expended and addressing the challenges of limited capacity, deep mistrust and who benefits (Serwadda, et al., 2018).

4. SUMMARY

This paper has presented the ongoing phased African Open Science platform project. The project aims to drive the dialogue on data, data policies and development of data infrastructure in the African continent to allow countries and regions to contribute and benefit from the global science enterprise through open data and open science and attain sustainable development goals. A landscape survey has been conducted around a framework of focus areas - policy, infrastructure, capacity building and incentives. The survey reveals a significant number of initiatives and developments in infrastructure and human capital development, primarily driven by project needs and preparedness. There is however limited advances in policy and incentive schemes. The main challenge in the project was to direct the focus on research data, although the need for open government data cannot be diluted. Policy is a process, and although the project managed to initiate the dialogue in many African countries, there is still a long way to go before policy will be implemented. From a strategic viewpoint, Phase 2 of the project will be focussing on developing an actual African Open Science Commons, similar to the European Open Science Cloud (EOSC), integrating open science initiatives, policy, infrastructure, skills development and incentives.

About Authors

Dr Tshiamo Motshegwa is based at the Department of Computer Science, Faculty of Science at the University of Botswana. He leads High Performance Computing and Data Science Research cluster. He has been Chair of the SADC Technical Experts Working Group developing the SADC Regional Cyber-infrastructure Framework. He serves on the Botswana Government's Ministry of Tertiary Education Science and Technology task team for the Botswana Space Science strategy overarching developments and opportunities in space sciences and technologies. He is also a member of the Botswana SKA & African Very Long Base Interferometer Network (AVN) Projects technical Committee. He also serves on the Botswana Open Data Open Science (ODOS) committee. He is in the International Steering Committee of the International Data Week Conference (IDW) and SCIDATACon-IDW 2018.

References

AAS, 2016. *Riding the National Research and Education Networking Train in Africa,* s.l.: Association of African Universities.

AOSP, 2016. *African Open Science Platform Concept*, Pretoria: AOSP.

Batchelor, S. et al., 2003. *ICT for Development : Contributing to the Millennium Development Goals - Lessons Learned from Seventeen infoDev Projects,* Washington DC: World Bank.

Bothwell, C. et al., 2015. *SGD ICT Playbook,* Fairfax, USA: s.n.

de Vries, J. et al., 2015. The H3Africa policy framework: negotiating fairness in genomics. *Trends in Genetics,* 1 March, 31(3), pp. 117-119.

Dewdeney, P., Hall, P., Schilizzi, R. & Lazio, T., 2009. The Square Kilometre Array. *Proceedings of the IEEE,* August, 97(8), pp. 1482-1496.

Gijzen, H., 2013. Big data for a sustainable future. *Nature,* October, Volume 502, p. 38.

Griggs, D. et al., 2013. Policy: Sustainable development goals for people and planet. March, Volume 95, p. 305–307.

Huijboom , N. & Van den Broek , T., 2011. Open data: an international comparison of strategies. *European journal of ePractice,* March/April.Volume 12.

Kitchin, R., 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* First ed. London: SAGE.

Kunda, D. & Khunga, B., 2015. *Implementing National Research and Education Networks (NRENs) in LandLocked African Countries: Critical Success Factors.* Porto, Portugal, s.n.

Motshegwa, T., Wright, C., Sithole, H. & Morgan, A., 2018. *Developing a Cyber-infrastructure for Enhancing Regional Collaboration on Education, Research, Science, Technology and Innovation.* Gaborone, IIMC International Information Management Corporation,.

Mulder, N. et al., 2017. Genomic Research Data Generation, Analysis and Sharing – Challenges in the African Setting. *Data Science Journal,* 16(49), pp. 1-15.

Mulder, N. et al., 2016. H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Research,* 1 December, Volume 26, pp. 271-277.

Norris, R., 2010. *Data Challenges for Next-generation Radio Telescopes.* s.l., s.n., pp. 21-24.

Robert, K. W., Parris, T. M. & Leiserowitz, A. A., 2005. What is Sustainable Development? Goals, Indicators, Values, and Practice. *Environment: Science and Policy for Sustainable Development,* 47(3), pp. 8-21.

Rotimi, C., 2014. Enabling the genomic revolution in Africa. *Science,* 20 June, 344(6190), pp. 1346-1348.

Sachs, J. D. & Reid, W. V., 2006. Investments Toward Sustainable Development. *Nature,* 19 May, 312(5776), p. 1002.

Sachs, J. et al., 2015. *How Information and Communications Technology Can Achieve The Sustainable Development Goals,* New York

SADC, 2016. *SADC Cyberinfrastructure Framework,* Gaborone: SADC Printer.

SADC-CSC, 2017. *Southern African Regional Climate Information Services for Disaster Resilience Development,* Gaborone: SADC.

Science-International, 2015. *Open Data in a. Big Data. World. An international accord,* Paris: International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS),InterAcademy Partnership (IAP).
Serwadda, D. et al., 2018. Open data sharing and the Global South—Who benefits?. *Science,* 359(6376), pp. 62-643.
SKA-Organisation, 2015. *Advancing Astrophysics with the Square Kilometre Array: 1.* 1st ed. City Giardini Naxos, Sicily, Italy: Dolman Scott Ltd.
UN, 2012. *The Millennium Development Goals Report 2012,* New York: UN.
UN, 2015. *Resolution adopted by the General Assembly on 25 September 2015,* New York: UN.
Wilkinson, M. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Nature - Scientific Data,* 15 March.

OAIS proposed new concepts

David Giaretta[1], J. Steven Hughes[2], John Garrett[3], Mark Conrad[4], Mike Kearney[5], Felix Engel[6], Matthias Hemmje[6], Robert R. Downs[7], Terry Longstreth[8], Bruce Ambacher[9]

[1]PTAB Ltd, Dorset, UK, david@giaretta.org
[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA, John.S.Hughes@jpl.nasa.gov
[3]Consultant, Colombia, MD, USA, garrett@his.com
[4]NARA, Washington DC, USA, mark.conrad@nara.gov
[5]Sponsored by Google, Huntsville, AL 35803, USA, kearneysolutions@gmail.com
[6]FernUniversität in Hagen, Hagen, Germany, mattias.hemmje@fernuni-hagen.de, felix.engel@fernuni-hagen.de
[7]Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY, USA, rdowns@ciesin.columbia.edu
[8]Data and Information Standards Consultant, Laurel, MD, USA, terry.longstreth@comcast.net
[9]Consultant, Columbia, MD, USA, bambacher@verizon.net

The first version of the Reference Model for an Open Archival Information System (OAIS), also known as ISO 14721, was published in 2002 and the second, updated version was published 10 years later, after a very long review process.

The standard has been enormously influential and is used as a basis for sundry guidance and for much of the work on preserving all kinds of digitally encoded information in an enormous variety of archives and repositories. Now, the review and revision process is underway again and clarifications, additions, and other improvements to the standard are being identified, examined, and recorded for subsequent review.

The review process is being conducted in a publicly open and transparent manner, involving a large number of individual and organisational contributions, as can be seen in the description and registry that are available on the http://review.oais.info review website. Many ideas for new and improved concepts have been contributed as well as suggestions for improving processes and their descriptions.

This paper discusses several of the additional concepts that have been proposed for the revision of OAIS. In the initial review process, which will produce an updated version of the standard for further review, some proposed changes have been accepted, others have been rejected, while others have been modified before acceptance. In addition, many suggested changes have identified the need for additional revisions to related concepts and definitions.

The suggestions, comments, and discussions are all available on the review website. Nevertheless, we believe this paper offers additional value because it reflects the insight and understanding that have been attained from the suggestions that have been contributed and offers insight into the revisions that are being proposed for the new version of OAIS. We also describe some of the broad underlying reasons for the decisions that have been taken at this stage. These include the influences that arise from the need to maintain the implementation agnostic perspective of the OAIS requirements, while offering flexibility for interpreting opportunities for practical implementation and enabling auditability through ISO 16363 audits.

Keywords

Digital preservation, Information Model, Digital Repository

Introduction

OAIS is the responsibility of a working group of the Consultative Committee for Space Data Systems (CCSDS), currently named CCSDS Data Archive Interoperability (CCSDS-DAI); and, like other standards published by the CCSDS and International Organization for Standardization (ISO), goes through the CCSDS and ISO review processes. The second version of OAIS was published in 2012. To maintain the 5-year review cycle, the next round of updates was begun in 2016 with the aim of collecting ideas for the new draft in the most open way possible. The new draft will then be put into the official review processes. One significant improvement that improves the transparency of this review is to conduct discussions on a publicly accessible web site rather than by email. ,This improvement ensures that anyone can read the discussions and, through a simple registration process, can also contribute to the review process. The website, http://review.oais.info uses a customized version of Bugzilla, which is normally used to track bugs in software, but, in this case is being used to track proposed changes to the text.

As stated on the website homepage, the aim of the review was to reduce ambiguities and to fill in any missing or weak concepts and to add useful terminology while ensuring backward compatibility with regard to major terminology and concepts. Further, for consistency, the general level of detail should not be changed nor, should the standard be changed from a reference model to an implementation design.

The process is to collect together what are termed "suggested changes" (SC). The proposer of the SC described the change, generally providing specific wording in the form "From …. To …..", accompanied by justification and discussion of the change. Others can then make comments and criticisms of the SC, and an exchange of views then ensues. At the weekly, open, CCSDS-DAI WebEx meetings, the SCs are discussed and, after more discussions, the group makes decisions as to whether to accept the change, reject the change, combine one change within another or accept the change with modifications to the wording. Pieces of text may be affected by separate SCs; in this case the challenge towards the end of the process is to produce the final text taking all changes into account, as far as possible. For example, one set of changes may involve a significant change of wording and this would have to override, for example, a typographical correction to the existing text, while taking such suggested corrections into account, if they remain applicable.

The updated draft of OAIS which will be the output of this exercise will then be subject to the extensive review processes of CCSDS and ISO procedures described on the homepage and which govern the final update and review processes. As noted, this is "*a long and involved process but is the one which has been well proven by CCSDS and ISO and has contributed to the success of OAIS*".

A total of 212 SCs have been submitted for OAIS from 26 national and international organisations. At the time of writing the new draft is not yet finalized so what follows is a description and discussion of what suggestions have been made rather than what changes have been accepted. The suggestions described here may be roughly grouped as (1) straightforward corrections to typographical or factual errors, (2) corrections to diagrams and/or the text which accompanies them, (3) changes which would improve the testability of OAIS compliance, (4) changes in discussion of preservation techniques and (5) other suggestions. About one third of the total number of SCs are corrections to diagrams and minor improvements to the text.

In the following sections we try to give a flavour of the variety of suggested changes through examples of these, but we focus on those SCs which would affect OAIS conformance and those which provide insights into digital preservation. The changes that have a significant effect on conformance will certainly have implications for ISO 16363 when it is revised, probably six months after the OAIS draft is ready. SCs for ISO 16363 are collected on the same web site.

Suggested changes affecting conformance

As stated in section 1.4 of OAIS, a conformant OAIS must "*support the model of information described in 2.2. ….. and…fulfil the responsibilities listed in 3.1*." where the numbers are the OAIS section numbers. The model of information is further elucidated in section 4.2 of OAIS while the mandatory responsibilities are further described in section 3.2 of OAIS. The separation of the text into these sections was designed to minimize forward references, so that concepts are introduced and explained to some extent before they are used elsewhere.

About half the SCs were declared by their reporters as not affecting conformance. Of the other half, which could affect conformance, only 36, directly concern those sections. Changes affecting conformance will certainly need to be reflected in ISO 16363 and examples of these potentially very significant changes include the following.

Preservation Objectives

OAIS defines Long Term Preservation as "*The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term*". The term Independently Understandable is defined as "*A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.*"

SC #115 (http://review.oais.info/show_bug.cgi?id=115) attempts to clarify the concept of "interpreted, understood and used", so that this is more easily testable and hence auditable, by adding the concept of Preservation Objective. The definition is "*A specific achievable aim which can be carried out using the Information Object*". Extending that definition is "*An example of what a member of the Designated Community should be able to achieve with the Information Object now and into the future. Preservation Objectives may be for example one or more of the following: being able to render an image, play a sound recording, display a document or calculate a scientific value or otherwise use digitally encoded information.*

A Preservation Objective should have the following attributes:

- *Specific, well defined and clear to anyone with a basic knowledge of the domain*
- *Actionable, the objective should be achievable currently and into the future.*
- *Measurable, it should be possible to know whether or not the objective has been attained at a given point in time.*"[2],

There would have to be a change the definition of Independently Understandable to "*A characteristic of information that is sufficiently complete to allow it to be understood and/or used by the Designated Community as exemplified by the associated Preservation Objectives without having to resort to special resources not widely available, including named individuals.*"

There are a number of consequent changes which are enumerated in the discussion of the SC.

Designated Community

The Designated Community is a fundamental concept in OAIS. SC#42 proposed to change the definition to make it easier for repositories to describe their designated Community(ies), which is of course important for auditing.

From: *Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.*

To: *Determine, either by itself or in conjunction with other parties, which entities should become the Designated Community and, therefore, should be able to understand the information provided. Definition of the Designated Community includes a determination of their Knowledge Base.*

Content Data Object

SC#222 addresses the concern that currently OAIS defines things like Fixity in terms of the Content Information, which is the combination of the Content Data Object and the Representation Information. However, the latter is potentially a network of linked information which is likely to grow and extend if required to ensure the Designated Community can understand and use the information. This makes it potentially difficult in practice, for example, to calculate Fixity.

The proposal is to change Preservation Description Information (PDI) from: "*The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information*" to "*The information which is necessary for adequate preservation of the Content Data Object and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information.*"

This change is believed to (1) reflect what archives currently do (2) make audits easier when the change goes through to ISO 16363 and (3) improve the logical consistency of OAIS.

Insights into Digital Preservation

Fundamental preservation options

Section 5 of OAIS is entitled Preservation Perspectives and was originally aimed at providing guidance about digital preservation. However, it is recognized that the section emphasizes Migration to the exclusion of other options, except for a brief discussion of emulation techniques.

The proposal was to make major changes in section 5 which may be summarized as expanding on the following ideas:

Fundamentally, approaches to information preservation in the face of changing technologies, resource availability and Designated Community requirements require a number of basic strategies which can be motivated as follows.
To preserve Content Information the information encoded in the Content Data Object being preserved may be
(1) kept by the archive unchanged or
(2) kept by the archive but may be changed or
(3) not kept by the archive, but instead be handed on to another archive which will preserve the information
Option (2) covers Migration, the current focus, while (1) includes the use of Emulation, which is mentioned in the current version of OAIS but is extended to more general addition of Representation Information, of which emulation software is one example. Option (3) recognizes that an archive may not be able to guarantee its funding, or even its continued existence, over the long term and so it must be prepared at some level to hand over its holdings and all the information contained in its AIPs. This makes auditing more practical in that whereas future resources cannot be audited, plans for a hand-over can be.

Suggested Changes which do not affect conformance

Many of the SCs are to correct or clarify diagrams, for example by re-positioning labels on arrows in figures or correcting the direction of arrows, and many are to ensure that the diagram and its accompanying text are consistent. All of these are important.

A great number of SCs refer to the Functional Model. which, although not mentioned in terms of conformance, nevertheless provides a rich and useful set of concepts, terminology and checklists for archives. The following are examples of these.

Preservation Watch

SC #120 points out the usefulness of the term Preservation Watch within the Preservation Planning functional entity, generalizing the functions “Monitor Designated Community” and “Monitor Technology” and explicitly recognizing that there are changes in things other than technology and the designated community which may be a threat to preservation.

Distributed Archives with Distributed Functional Entities

The review also addressed several SCs (SC#127,133,134,135, 155) that argued that a distributed OAIS with Distributed Functional Entities can appear in the form of, e.g., an association involving an OAIS with distributed Functional Entities that has entered into agreements with other OAISs to link or integrate their distributed functionalities or/and services with each other in a complementary way.

Such a distribution of Functional Entities of an OAIS as well their collaborative composition into distributed OAIS, can be of a physical, organizational, or administrative nature. The motivation for such a distribution of Functional Entities of an OAIS may be to distribute resources to achieve the complete set of Functional Entities to establish an OAIS in a complementary and collaborative way.

This type of association of Functional Entities of an OAIS is fundamentally different from the previously existing architectural distribution examples, in that it does not only federate, share or cooperate with respect to Functional Entities but really physically, logically or organizationally distributes them in accordance with competencies and capacities of contributing archives, locations, organizations or administrations.

Conclusions

This paper has just scratched the surface of the suggested changes. Looking at the SCs as a whole, it is clear that they are all very valuable, whether they are recommended for acceptance or not, because of the discussions that they have stimulated. Of those mentioned above, some have been rejected, others accepted and yet others, at the time of writing, are still under discussion.

Looking at the discussions and the decision process it is clear that a number of fundamental ideas are being adhered to, namely:

(1) Backward compatibility with the current version of OAIS must be maintained, for example changing the name of the functional entity *Archival Storage* to *Preservation Storage* would break this backward compatibility.
(2) The general level of detail should not be changed and
(3) The standard should remain a reference model and continue to be implementation agnostic. For example, going into detail about new technologies such as "cloud" may be seen to be too implementation specific, although this may need to be addressed in ISO 16363.
(4) Where there is a topic on which nothing useful can be written, then nothing should be written.
(5) Concepts should, as far as possible, be re-used rather than creating new concepts.
(6) Clarity and specificity of what is meant is preferred to enumeration of what is not meant. However, there is clearly a balance to be reached to ensure clarity. For example, should it be stated explicitly that a functional entity may be distributed, or is it simply enough to omit specifying that it is a single unit?
(7) An important use of OAIS is now to form the basis of ISO 16363 for Audit and Certification of Trustworthy Digital Repositories and therefore importance should be given to making auditing more practical.

Acknowledgements

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (NASA).

Support for Robert Downs was received from NASA under contract NNG13HQ04C for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC).

References

[1] Reference Model for an Open Archival Information System (OAIS), 2012, CCSDS 650.0-M-2 and ISO 14721:2012. Available from https://public.ccsds.org/Pubs/650x0m2.pdf
[2] Curating Scientific Research Data for the Long Term: A Preservation Analysis Method in Context, available from http://www.ijdc.net/index.php/ijdc/article/view/182
[3] Audit and Certification of Trustworthy Digital Repositories, 2011, CCSDS 652.0-M-1 and ISO 16363:2012. Available from https://public.ccsds.org/Pubs/652x0m1.pdf
[4] The website for the 5-year review of OAIS and ISO 16363, http://review.oais.info/

ESA Data Preservation System

Mirko Albani[1], Michel Douzal[1], Domenico Castrovillari[2], Paolo Boezi[3], Daniele Iozzino[3], Iolanda Maggio[3]

[1] European Space Agency, Intecs Soution, [3] Rhea Group

The European Space Agency has the mandate to assure the long-term preservation, sharing and exploitation of space data and its associated knowledge. ESA's aim is to turn space exploration and space-related activities into an overall societal project involving a wide variety of stakeholders. To this end, it brings together and coordinates as many countries as possible under the banner of space missions. It is a basic principle that ESA deals with its stakeholders openly and with real transparency, an approach that has contributed to its long-term success.
The EO Data Preservation System has the main objective of providing the required infrastructure and services to assure ESA and Third Party Missions (TPM) EO Data Records and Associated Knowledge preservation and accessibility, and to support the cooperation activities with national and international organizations in the data preservation domain. The generic "EO Missions/Sensors Preserved Dataset" content includes Data Records and Associated Knowledge.

Keywords: Long Term Archive; Provenance; Data Preservation; Processes

Introduction

The main components of the EO Data Preservation System involved in the preservation process are the Management System for Data and Associated Knowledge (KMS), Data Information System (DIS), Master Archive (MAR), Cold Back-up Archive (CBA).

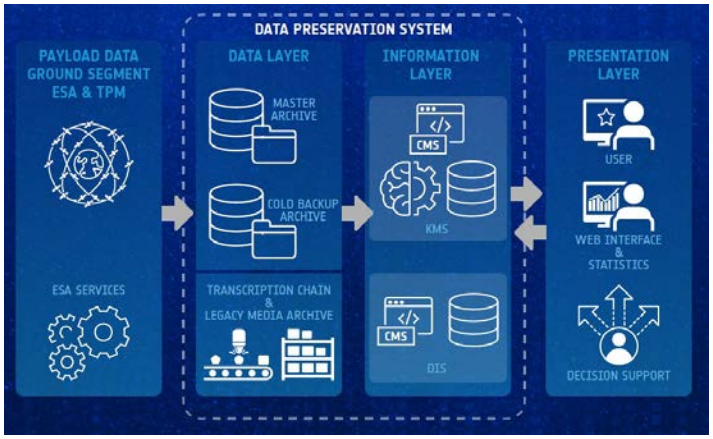


Figure 1 – EO Data Preservation System

The Master Archive and the Cold Back-up archives cover the archiving functionalities whereas the Data Information Service (DIS) provides the information, the history and the provenance of the data archived. The concept of Data Information Service arises from the service requirements ESA included within the Data Service Initiative (DSI) program modelling systems and processes to enable the management of ESA's EO Data as a standard asset. These requirements ensure that metadata for each product is collected in a database and the data element itself is systematically stored with full track of all value added to the data during the service activities. The underlying database storing all information is well suited to report on the activities within the DSI service but also to present zoom-able level of detail for any EO Data asset held by ESA, making the tool useful to staff involved in operational support to management and decision-making. In addition, there are other sources of operational data and repositories around the data payload ground segment. The huge value of the information contained in all these systems is enhanced by providing a harmonised, service-independent view and control of the data assets held across system in order to provide end-to-end operational analysis of data assets to pinpoint changes, errors or discrepancies. The crucial factor

determining the success of the DIS is the ability to receive the metadata, recognise products across source services, adjust obtained information and produce the unequivocal set of attributes for any data asset. The scope of this paper is to describe the features of the system and any relevant preservation processes.

Master Archive

The ESA Master Archive is implemented through a dedicated service (DAS) awarded to industry through an open ITT in 2016. ESA has outsourced EO data archiving activities to a single provider with the goal to benefit from economy of scales and standardization of data archival & delivery processes and interfaces.

The industrial consortium is made of ACRI-ST (FR), adwäisEO (LU) and KSAT (NO), with the following distribution of roles:

- ACRI-ST: data archiving (and delivery) service provider – Prime contractor
- adwäisEO: data archiving (and delivery) operation provider with its leading edge IT and connected secured infrastructure in Luxembourg
- KSAT: data knowledge provider and overall ingestion validation

In order to guarantee the data safety, the archive is distributed between two locations sited > 200km apart; a master archive in Luxembourg and an archive back up in Sophia Antipolis (France). The archival operational flow between the two facilities is depicted in Figure 2, which also lists the technical solutions for the infrastructure in each data centre. Several data quality checks in terms of reliability of the process during data transfer are performed all along the flow (data not corrupted during transcription or during copy to the backup center, data still readable on tape).

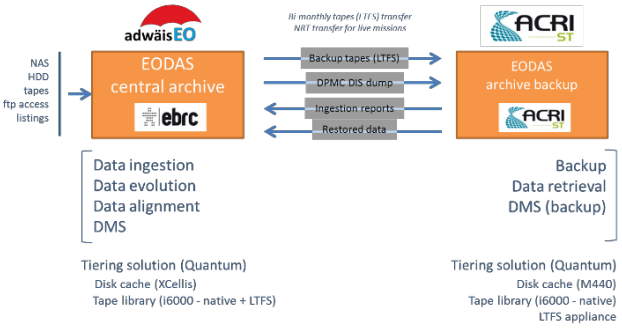


Figure 2 – Master Archive operational flow

The Master Archive infrastructure is mainly based on two similar Quantum iScalar 6000 libraries connected via 10+ GE and 16 Gbps SAN links to DELL M1000e + M6x0 blade enclosure and servers. A StorNext System is used to manage the data in Hierarchical Storage Management (HSM) mode. In this environment, the disk structure containing the data is exported to NFS clients as a classical Linux volume and specific policies determine the way to store the data (keep on disk and/or tape, automatic generation of several copies…). Data are copied to LTO7 tapes in native Quantum format (ANTF) for data transfers between main and backup archive centres as shown in Figure 3.

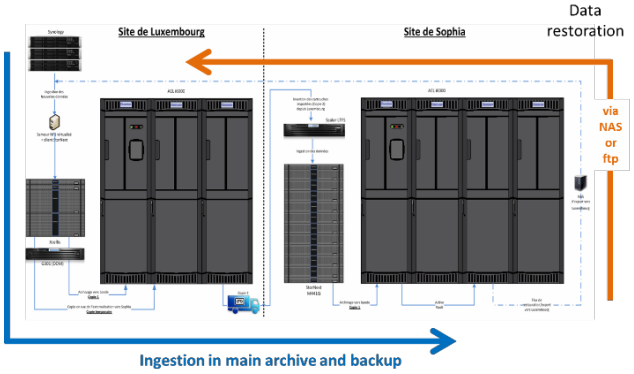


Figure 3 – Master Archive infrastructure

The Master Archive ingests data from historical missions according to a data ingestion plan constantly maintained by ESA Data Librarian as well as data coming from live missions (currently Cryosat-2 and SMOS) upon formal requests by the relevant ESA Operations Managers.

The status and progress of the data ingestion is visible in real time at the service web portal at http://www.eodas.info.

Figure 4 - EODAS Service Web Portal

The EODAS Service Web Portal is updated on hourly basis with continuous injection of fresh information coming from the core processes that drive the data archiving. Lots of views and filters are available to select the information of interest allowing also making exports in pdf format or excel tables for any further analysis and statistics of the archived products.

Cold Back-Up Archive

The Cold Back-up Archive (CBA) has been implemented using hardware already available in ESRIN in 2014. Throughout the years, it has been upgraded to enhance performances and allow seamless archive and extraction capabilities. It currently contains:

- Unique data
  - Historical data from ESA Processing and Archiving Centres
  - EO Unconsolidated data
- Live Mission data
  - Swarm (started in mid-2014)
  - Odin
  - SCISAT
  - Proba-1
  - Cryosat-2 (started in March 2015)
  - SMOS (started in May 2016)
- Second Copy of the Master Archive Data
- ESA Science Disaster Recovery Data

The CBA has been the core storage of the Processing and Archiving ESA External Centres Phase-out project, in which it stored and extracted around 4.5 PB of data in solely 3 years.

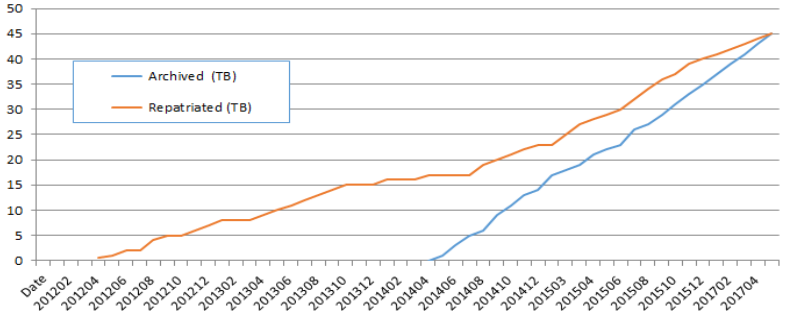


Figure 5 - ESA External PAC Phase-out data

The CBA stores two copies of the data in its two archive centres, located in separate buildings on ESRIN premises. Synchronization between the two sites is performed through a dedicated 16 Gb/s fibre optic connection.
In the scope of the Inter-directorate Joint Activities, a dedicated circulator software has been implemented in order to allow reception of ESA Science data from both the internet and ESA WAN connected centres.

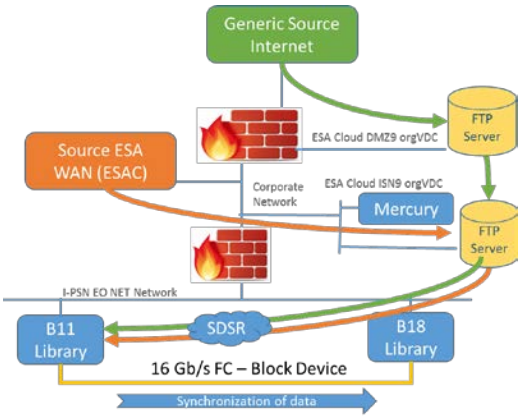


Figure 6 - CBA Workflow of Science Data

Live missions data are circulated by the Preservation Element Front-End (PE-FE). Once the data have been ingested and validated, confirmation reports are sent by the CBA to the relevant Mission Payload Data Ground Segment.

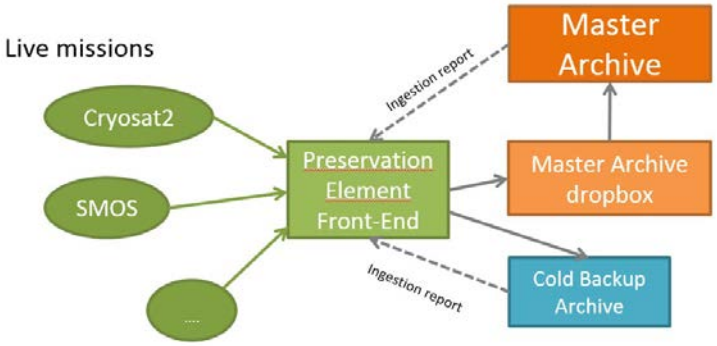


Figure 7 - Live Mission Data Workflow

The infrastructure of the CBA consists of two Robotic Libraries, capable of storing 24 PB of online data each. The Main Library is a 3000 slot STK SL8500, equipped with 8 T10000D and 10 T10000B drives, whereas the Disaster Recovery Library is a 3000 slot STK SL3000 with 4 LTO-7 drives. Migration to LTO-8 is foreseen, allowing 36 PB of online data.
Both Libraries are seen as a huge virtual file-system through the ORACLE HSM. Extraction and Archiving capabilities are maximized using first tier of Solid State Drives pools, a 10 Gb/s Network and a fully 16 Gb/s fibre optic infrastructure.

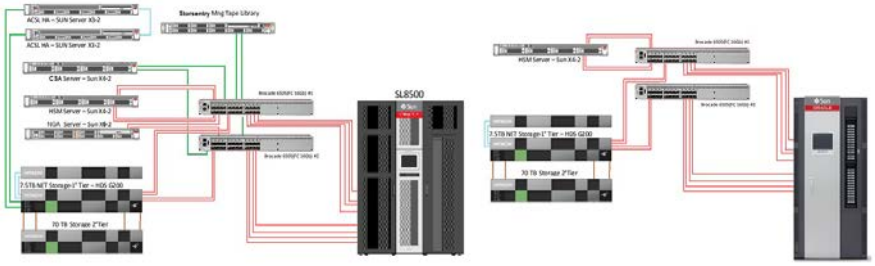


Figure 8 - CBA HW Infrastructure

The CBA currently hosts approximately 6 PB of data for a total of 105 million of files.

Figure 9 - CBA data holdings

Both EO-SIP packages and native format data are archived using a dedicated archiving software capable of validating the data, extracting metadata and generating confirmation reports. Data can be either bulk ingested/extracted or automatically ingested via inbox baskets from local or remote locations.

A web portal is used and updated regularly with the status of the ingestions, extractions and figures of the data stored

Figure 10 - CBA Web portal Dashboard

2. Data Information System

The Data Information System (DIS) is a management support system based on EO data products. DIS is in charge of managing metadata information for all products generated, used and distributed by any EOP-GES project for all the ESA and Third Party Missions. DIS has been developed within the same Contract, which implements the service for data consolidation and reprocessing (DSI) as an extension of the internal inventory system. DSI is managed by Serco Italy.

The original core of the system has been upgraded to manage the information about products available at many different systems and the relationship among them and the history of data. DIS is supporting data awareness and control, as the system is keeping trace of any metadata for all products stored at different processing and archiving sites. DIS is consisting of a database repository, an ETL module and a Business presentation layer. DIS facilitates improved control of the EO data owned by ESA, concentrating information spread over many different external services in a single place, and supporting verification that all data is aligned at all different sites and distribution of the right data for the projects requesting it. DIS is fed from different services at ESA; each service is providing all the available information about its data, how it is organized and the changes applied to it. DIS is collecting all the information, integrating what has been received to build up the full set of metadata for each product. A major requirement for DIS is to trace history of data, the changes applied in the past, what are the datasets including this product, what the different versions of the datasets consolidated over the time, what is the version accessible for users requesting it.

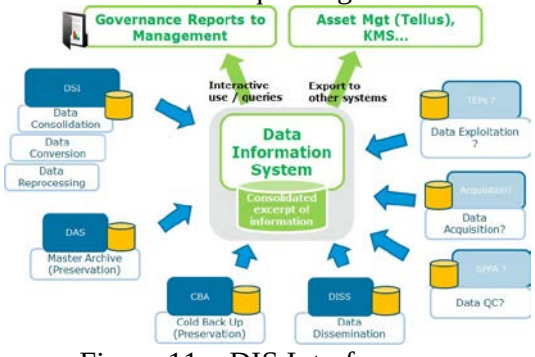


Figure 11 – DIS Interfaces

DIS has initially started collecting data from the DSI projects, and now it is integrating with DAS. In a near future additional service will be added to complement the scenario as repository of data, specifically Cold Backup and Dissemination Services. Further extensions of DIS coverage will be analysed later on. The information available in DIS is accessible through the Business Intelligence layer on top of it, allowing to design and publish on web the reports defined by users to monitor and control their data. A set of predefined reports supports direct access to the most common views over the data, and dedicated reports are being developed for specific tasks. The presentation layer of DIS is supporting drill-in on the charts presented, allowing to select the specific data of interest to retrieve high level information up to the full detail.

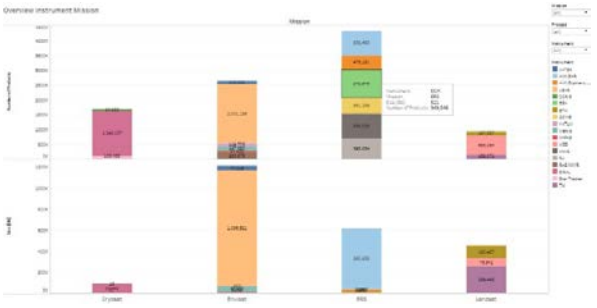


Figure 12 – DIS presentation charts

For each product DIS is storing and referencing not only metadata, but cross-references to other products and processes related to the product, as

- The datasets to which the product is associated / disassociated, and the versions involved
- The changes applied to the product, and history of different versions available
- The auxiliary files used to generate the product, and the versions of each auxiliary file used
- The lower level products used to generate it
- The products generated from the product during any further processing activity

Most of the above information is already available on the presentation layer, or it will be added in the next future. Analysing large sets of data at product level cannot easily managed, and therefore DIS is implementing full support for the dataset level, where products can be freely aggregated to highlight common characteristics. For each dataset, DIS stores:

- Information available at each archiving site
- List of products associated to each dataset
- History of all versions created for the dataset, and the changes applied a dataset version for migrating to the next version
- How the dataset was generated, using which datasets and auxiliary datasets, and by means of which transformations

The provenance view on the DIS website shows graphically the relationship between a dataset and the others, and the processes applied to it, using an high-level view, easy but very helpful for the understanding of the data involved.

Conclusion

Since its birth, the EO Data Preservation System has allowed ESA and its Heritage Data Program (LTDP+) to preserve both the data and the knowledge, and being compatible with the OAIS and CCSDS preservation standards and guidelines.

The Data Preservation System will be continuously supported and enhanced in terms of both Hardware and Software, in order to ensure the preservation and valorisation of the ESA Earth Observation Assets.

British Geological Survey (BGS) Practices in Data Curation

Pinnick, Jaana [1]; Baker, Garry; and Riddick, Andrew
Informatics Directorate, British Geological Survey, Keyworth, United Kingdom
[1] jpak@bgs.ac.uk

Introduction

The long validity of geoscience data, often unique and unrepeatable, means permanent retention is often required (Pinnick, 2017). As previous data must remain available for future research and interpretation, a generic 10-year retention is therefore insufficient. The best practice guidance by the UK Research and Innovation (2015) states '…data that by their nature cannot be re-measured or re-created such as earth observations […] may often warrant indefinite storage and preservation'. Future research questions cannot often be predicted, but collecting a large variety of data types over decades whilst striving for standardised contextual metadata is challenging. BGS has legislative obligations to manage some data types, but its wide variety of stakeholders have sometimes conflicting priorities. The role of the NGDC (National Geoscience Data Centre) as a national repository requires it to monitor the condition of digital data and to collaborate with stakeholders to ensure the continuing usability, trustworthiness and interoperability of data in its care.

Appraisal and Value of Research Data

The need to preserve research data has to be considered in the light of several factors. Who are we preserving data for? Who appraises it? Who pays for its long-term storage? The availability of reliable and up-to-date data is a prerequisite for developing high-quality data products and policies as well as the basis on which funding decisions of future scientific research projects are often made. However, the money spent on data generation and preservation does not necessarily correspond to its possible future reuse value. In the era of 'big data', processing and appraising datasets of several terabytes can be challenging even without talking about the preservation costs over time, which in itself is difficult to evaluate. BGS needs to plan for 'big data' generated by sensor networks and environmental monitoring and to assess the need to keep them in the longer-term. It is more than likely that in the event of such large growths in data volumes the role of data appraisal and the need for more reliable costing predictions will increase considerably.

Preservation Policy Development

As an initial step before developing a preservation policy for the data centre, an exploratory MSc dissertation (Pinnick, 2016) studied the preservation requirements and corporate drivers of the BGS and the NGDC within their wider stakeholder community and the organisational framework. The study also looked at the efficiency of the existing data management procedures in terms of digital continuity within the current challenging funding climate. A small-scale end user survey, which discovered that 40 per cent of the respondents have used NGDC data for 10 years or longer, also found that over 70 per cent of them use raw data. A great majority of theparticipants were concerned about file format obsolescence and the lack of required preservation skills and resources.

Following the initial research, publicly available digital preservation policies and strategies from memory organisations (e.g. British Library, UK National Archives, UK Parliamentary Archives) were studied to inform the selection of preservation policy elements and structure. A useful resource at this stage was the Digital Preservation Coalition Digital Preservation Handbook (2015), which contains a comprehensive section on developing institutional policies and strategies as well as writing a final 'business case'.

Although the Informatics senior management fully appreciate the importance of developing BGS preservation capabilities, a business case was written prior to initiating the work, placing it into a wider organisational context and to summarise the objectives and benefits for BGS. The case outlined a modular digital preservation solution based on the OAIS model, developing the existing in-house resources e.g. data ingestion tools, data access pathways and metadata. It also suggested introducing new tools and technologies, where applicable, to increase automation and to create further efficiencies reducing the cost of the planned work. Finally, it estimated staff costs over the next four years and proposed some measures of success.

For the purposes of clarity, we defined the preservation policy as the overall guiding fixed principles of why we need to preserve data, giving us the mandate to do the work. The preservation strategy, when completed, will be a flexible action plan implementing the policy, describing the methodologies employed to achieve the preservation goals, and designing the workflows to be put into place. The BGS policy (BGS, 2017) has been made publicly available and will be reviewed every 3-5 years, whereas the strategy will be a living document aligned to current corporate strategy and resources and therefore kept internal.

The policy has been approved by the Senior Informatics Directorate Management and applies to both born- digital

and digitised research data. It is used in conjunction with existing policies and legal and regulatory framework and defines the roles and responsibilities of all stakeholders and lists standards and best practice the organisation will follow in its long-term preservation of research data.

Programme Development and Risk Mitigation

The next task is to develop a comprehensive preservation strategy. The aim is to create a modular preservation programme starting from the 'parsimonious preservation' principle developed at the National Archives, integrating preservation workflows throughout the existing data management processes to keep the costs down. Gollins (2012) argued that technological obsolescence is not, as often stated, the primary threat but that 'a much more imminent threat is poor capture and inability to achieve safe and secure storage of the original material'. A key challenge within the research data lifecycle for the NGDC is the capture and transfer of data from the researcher(s) to the repository. It is therefore essential to clarify roles and responsibilities of each party to facilitate the data capture and to ensure the early creation and capture of high-quality metadata and to document clear terms and conditions for reuse. The NGDC Data Deposit Portal has been designed with this in mind. Work is ongoing to add a PREMIS preservation metadata element to the existing BGS Discovery Metadata schema, which complies with the ISO standard 19115:2003.

Over the next couple of years we will create a digital data asset register for research datasets by loosely implementing the Data Asset Framework (DAF) methodology (2009), which is freely available on the DCC website, although ironically the associated DAF online tool appears currently to be inaccessible. As part of this project, a survey will be conducted across internal data creators and users to collect information about data assets. The asset register will provide information for prioritisation of at-risk data and datasets and support improved decision-making for preservation planning and actions. It will be integrated with the corporate retention schedule, currently under review.

Open Data and Research Data Management Training

NERC data policy guidance notes (2016) state: 'The environmental data produced by NERC-funded activities will be openly available for others to use and must be submitted to NERC for long-term management and dissemination'. The NGDC facilitates data sharing for NERC grant holders as well as other stakeholders. The human element will be addressed by increasing awareness about preservation best practices amongst data creators, by delivering data management training to all stakeholders to ensure preservation is relevant to them, and by enhancing contextual metadata capture.

Collaboration and Sharing Preservation Expertise

BGS is a member of the Digital Preservation Coalition (DPC) Workforce Planning Sub-Committee and participated in their BitList crowd-sourcing exercise in 2017 to identify the most at-risk digital materials. We also work with the Digital Curation Centre (DCC) and the UK National Archives (TNA). Nobody can preserve data in isolation and we want to share our knowledge and experience with other Data Centres and digital repositories. This helps develop capabilities in both generic preservation and with more discipline-specific requirements. To ensure continuity of large volume datasets of long-term value BGS already works with CEDA using its JASMIN super- computing facilities, and using social media is a way to build up professional networks and to keep up with recent developments.

Business Changes and Benefits

Our business case was written to justify a strategic allocation of resources for data preservation. BGS want to introduce a better accountability for managing assets across the data lifespan, from streamlining IT infrastructure components to enhanced long-term interoperability of data and systems. To minimise the need for expensive data rescue efforts, best practice must be implemented from the pre-ingestion stage onwards. As part of the culture change we will work to improve the organisation-wide understanding of the corporate value of the data, its reuse and preservation. Capturing complete and consistent rights metadata strengthens data reuse minimising the risk of its misuse, and providing this information with the data is a core function in the online Data Deposit Portal process.A comprehensive preservation programme will enable a planned approach to technological and strategy changes and enhance cost and resource effectiveness. It will reduce risks to corporate reputation due to unavailability of data and ensure data remains accessible to stakeholders. Updating organisational skills and technologies will strengthen the data quality and uphold the value of the investment made over time to curate the unique data assets at the repository, be it in time, money or professional expertise. The preservation work has contributed to the NGDC obtaining the CoreTrustSeal accreditation, and offering secure open access to reliable and reusable geoscience data will support innovation and new opportunities for all stakeholders using those assets.

References

British Geological Survey 2017 *BGS Digital Preservation Policy*. Available at http://www.bgs.ac.uk/downloads/start.cfm?id=3173 [Last accessed 12 April 2018]

Digital Curation Centre 2009 Data Asset Framework Implementation Guide Available at http://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf [Last accessed on 12 April 2018]

Digital Preservation Coalition 2015 Digital Preservation Handbook. 2nd ed. Available at https://dpconline.org/handbook [Last accessed 12 April 2018].

Gollins, T 2009 Parsimonious preservation: preventing pointless processes! Available at http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf [Last accessed 12 April 2018]

Gollins, T 2012 Putting parsimonious preservation into action. Available at http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation-in-practice.pdf [Last accessed 12 April 2018]

Natural Environment Research Council 2016 NERC Data Policy guidance notes. Available at https://nerc.ukri.org/research/sites/data/policy/data-policy/ [Last accessed 12 April 2018]

Pinnick, J 2016 Exploring digital preservation requirements: A case study from the National Geoscience Data Centre (NGDC). Unpublished dissertation (MSc), Northumbria University.

Pinnick, J 2017 Exploring digital preservation requirements: A case study from the National Geoscience Data Centre (NGDC). *Records Management Journal* 27(2), pp.175-191. DOI: https://doi.org/10.1108/RMJ-04-2017- 0009

UK Research and Innovation 2015 *Guidance on best practice in the management of research data*. Available at https://www.ukri.org/files/legacy/documents/rcukcommonprinciplesondatapolicy-pdf/ [Last accessed: 12 April 2018].

Pioneering Steps towards Use of Data-cubes in the Global Earth Observation System of Systems

Joakim Adrup joakim.adrup@esa.int
ESA Centre for Earth Observation, EOP-GEE, Frascati, Italy
Royal Institute of Technology, School of Electrical Engineering and Computer Science, Stockholm, Sweden

The purpose of the activity is to explore the possibility of incorporating the use of data cubes in the Global Earth Observation System of Systems (GEOSS). The platform could benefit from the ability to also provide data cube capabilities. Data cubes allows more flexible access that can be utilized for visualizing data. The focus in this paper has been on allowing users extended tools for inspecting the data visually and how this is best achieved. The goal always being to accom- modate the users requirements while also keeping the tools general enough to be useful for a wide variety of purposes and users. However viewing changes with respect to time has been identified as one of the most valuable aspect that data cubes can accommodate for users in the GEOSS platform. The reason being due to the wide range of different applications that can come from time depen- dent analysis. This paper has investigated and presents a selection of methods for inspecting and comparing time-points on a map interface.

Background

Finding out opportunities and methods for utilizing data cube capabilities can provide added value to the users of the GEOSS platform. On a larger scale this investigation can serve to give hints and guidelines for other platforms that want to incorporate data cubes.

1 Earth Observation datacubes

Earth observation (EO) data cubes is an initiative to try and harmonize EO data access and format (Strobl et al. 2017). So the task for one person to access multiple different sources of EO information is not insurmountable. Data cubes usually supply analysis ready data, which means that the data is projected and preprocessed in a way that allows the user to start analyzing without additional efforts on their side. An example of preprocessing done is spatial alignment, meaning that pixels in the same position have been aligned to all have the same geographical location. The data ”cube” might give the notion that a cube is limited to the three spatial dimensions. But in reality it is much more flexible. It can handle any number of dimensions in practice being more of a hyper-cube (Baumann 2017).

1.1 GEOSS platform and portal

Global Earth Observation System of Systems (GEOSS) platform aims to combine ground segment data from multiple different sources. Data come in various forms, not only as satellite data. It can also be so called in situ data, which refers to data collected on ground or near ground (Anderson et al. 2017). It could be things like water buoys, weather stations, etc. All this data are made available through the portal via meta data (Max Craglia et al. 2017) that describes the location together with many other aspects. The data is managed by an architecture and middleware called the DAB (Data Access Broker) (Nativi, Massimo Craglia, and Pearlman 2012). This manager software is aimed to tackle the difficulties of the many different data formats as well as maintaining where the data is located.

1.2 Research Qyestion

How can the GEOSS portal be extended to allow the user to interact with and inspect the content of the Earth Observation data cube focusing on aspects that fulfill the users requirements?

2 Implementation

2.1 User Scenario

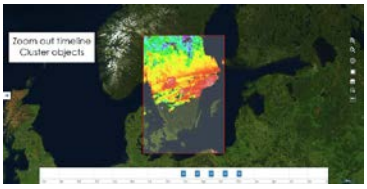
In the initial phase of the project a brief background research was conducted. This information inquiry was done to understand what aspects of data cubes that com- munities would find interesting and useful. A couple of different organizations were studied to find out their goal and what data they were interested in gaining from the GEOSS platform. Organizations were picked either by their previous interest or col- laboration with GEOSS or because of their work aligning with data that the GEOSS platform could provide. After the study it became clear that investigating change and

specifically change with respect to time was the most interesting point to many of these organizations. Convincing use cases could easily be constructed that aligned with research many of these organizations conducted. An example of how a scenario could look like is shown in figure 1.

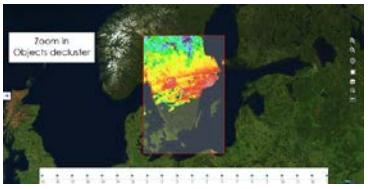
Figure 1: User scenario Global Forest Observations Initiative (GFOI)

2.2 Inspecting tool

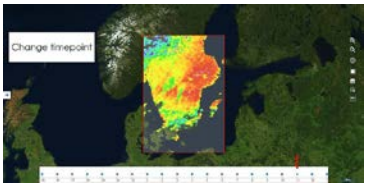
Inspecting in this context refers to the act of browsing through a set of data with the intent of understanding its content and features. The user often wants to understand what data is available in terms of content and quality before committing to down- loading the data. Inspecting data can allow the user to find interesting sections in the data that the user did not anticipate before opening the data set. Inspecting should be quick and responsive in order to be useful for the user. The method of conveying time information to the user was done by using a timeline (see figure 2). The purpose of the timeline is for the user to understand when and to what extend there exists data acquisitions. The design of the timeline should be minimal, limiting the amount of visual clutter while still conveying essential data.



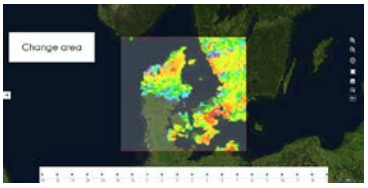
(a) Clustering of time-line objects



(b) Declustering of time-line objects



(c) Change of timepoint in the time-line

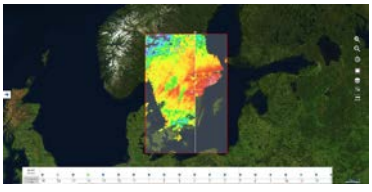


(d) Change of area of interest

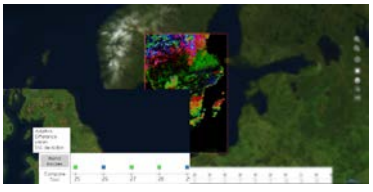
Figure 2: Timeline tool, example data is MODIS surface temperature

2.3 Comparing tool

Comparing different time-points can be a beneficial tool for a user to gain a quick understanding of changes in time. Take the example of an urban area, the user might want to quickly check if there is visible difference in the growth of the urban area. Viewing if houses have been built, forests cut down, roads laid, etc. A responsive and intuitive tool to compare time points could in this case tell the user where to put their effort. Easily showing which areas are of importance. How do you go about com- paring time-points then? Two promising ways of giving this ability to the user have been identified. The first one is to overlay two different time-points then create an interactive slider that the user can slide back and forth alternating between revealing the overlaying and underlaying image layer. The second method is to allow the user some simple computational functions. The HTML5 canvas comes with a selection of predefined methods of combining image data called composite operations (S. Fulton



(b) Compare Slider



(b) Image Composite Modes with difference

Figure 3: Compare tools

and J. Fulton 2013). These functions can quickly combine image with additive, differ- ence and plenty of other options. Giving the user a toolbox of functions to use when they want to compare layers. In fig. 3b a difference operation is used.

3 Discussion

The use of a timeline is a quick, common and well explored tool for the purpose of conveying time related data. The connotations that users associate with timelines are already there. They already know its purpose on first glance. Giving the advantage in ease of use and a better experience for the user. Understanding these aspects can be utilized to great effect in designing tools that behave according to what the user expects and wants. An important aspect for the user to feel freedom while they inspect dates is to allow actions such as easy movement along the timeline and zooming for more precise date information. Giving the user a means of comparing data can be of added value to the user. It can give them an understanding of how the data could come to be used and to what benefit the data is for the user. The amount of different use cases for this wide range of different datasets makes it difficult to cover all if not even impossible. The best approach is supplying a set of general tools that can cover as much as possible without being too specific. Defining some general tasks like inspecting and comparing helps in understanding what tools can be implemented to support these tasks.

4 Ongoing Work

In the near future users will be brought in to test and evaluate the prototype imple- mentation. With the purpose of gathering user insight on flaws in the design, potential opportunities and to better understand how their requirements can be met. The time- line tool is finished with reservations to minor changes. But the tools for comparison is still in progress and has not yet reached full implementation.

5 References

Anderson, Katherine et al. (2017). "Earth observation in service of the 2030 Agenda for Sustainable Development". In: *Geo Spat. Inf. Sci.* 20.2, pp. 77–96.

Baumann, P (2017). *The Datacube Manifesto*. http://www.earthserver.eu/sites/default/files/upload by users/The-Datacube- Manifesto.pdf.

Craglia, Max et al. (2017). "Exploring the depths of the global earth observation system of systems". In: *Big Earth Data* 1.1-2, pp. 21–46.

Fulton, Steve and Jeff Fulton (2013). *HTML5 canvas: native interactivity and animation for the web.* " O'Reilly Media, Inc."

Nativi, Stefano, Massimo Craglia, and Jay Pearlman (2012). "The Brokering Approach for Multidisciplinary Interoperability: A Position Paper*". In: *International Journal of Spatial Data Infrastructures Research* 7, pp. 1–15.

Strobl, Peter et al. (2017). "The Six Faces of the Data Cube". In: *The Six Faces of the Data Cube*, p. 4.

DATACUBE as a NATIONAL GEO-SPATIAL INFORMATION SYSTEM

As with other countries in Eastern Africa, the climate of Uganda is characterised by the frequent occurrence of droughts and floods that often cause major disruption to the population through their impacts on water resources, agriculture, food security and disease. Recent trends over the region indicate changes in rainfall profile in recent years result in longer and more intense periods of drought.[1]

The impact of these events may be mitigated or reduced by Forecasting and Early Warning Systems (EWS) which allow appropriate actionable responses to be taken in advance. Existing early warning systems in Uganda are piecemeal in nature, often incomplete and seen as unsustainable and ineffective as a whole[2].

The Drought & Flood Mitigation Service (DFMS) takes a multi-faceted, inclusive and open approach to ensure that decision makers in Uganda are provided with practical information that will improve knowledge about and help reduce the impact of weather and climate hazards.

The primary objective of DFMS project is to reduce the impact of flood and drought events on the livelihoods of Ugandan communities. It will do this by providing localised monitoring and forecasts of environmental conditions relevant to many day-to-day activities.

The legacy of this project is a long term data store of space and non-space based environmental information which is increasingly viewed as national geospatial asset.

The DFMS project is being led by the RHEA group who have joined forces with a number of UK based organisations, including the UK Met Office as well as other NGOs based in Uganda. The Ministry of Water and Environment (MWE) leads the Government of Uganda's involvement, in formal cooperation with the Uganda National Meteorological Authority (UNMA), Office of the Prime Minister (OPM), Ministry of Agriculture, Animal Industry and Fisheries (MAAIF) and National Agricultural Research Organisation (NARO)

The project is part of the UK Space Agency's International Partnership Programme, a 5-year, £152 million programme designed to partner UK space related expertise with governments for sustainable economic or societal benefit in partner countries.

The DFMS monitoring and forecast services provided in collaboration with Ugandan government ministries and institutes will apply across the country, rolled out initially to pilot agricultural communities in Karamoja and the Cattle Corridor as well as the Kakira Sugar Corporation near Jinja. DFMS services are also offered to the wider community including NGO, agri-businesses and commercial information providers. In addition to direct agricultural activity, other potential uses include Disaster Risk Financing, Insurance, Health, Nutrition, and Education programmes.

The DFMS is due to become fully operational in 2019, with prototypes available incrementally from Spring 2018.

DFMS PLATFORM

DFMS provides an innovative, open platform for the collection, processing, storage and dissemination of high resolution information from a diverse range of sources. The information includes:

- A range of tailored, timely satellite-derived products including: soil moisture, land surface temperature, vegetation indices, water level extents, water height and land cover classifications.
- Local measurements from Uganda's network of ground based environmental sensors including meteorology, river flows, and soil moisture.

[1] See http://onlinelibrary.wiley.com/doi/10.1002/2016RG000544/full

[2] Building the concept and plan for the Uganda National Early Warning System (NEWS) – Darren Lumbroso on behalf of UK Department for International Development (DFID) July 2016

- Agricultural community-level measurements including crop and livestock condition.
- A suite of weather forecasts from daily to seasonal timescales.
- A hydrological forecast driven using consistent set of weather forecast information to forecast river flows, runoff, soil moisture and evaporation allowing prospective drought and flood conditions to be calculated.
- The flexible platform provides a wide range of outputs such as environment monitoring data, risk maps and early warning alerts. Expert and non-expert users are catered for. These may be configured for specific needs and location to support decision making at national or local level. An applications programming interface (API) allows independent organisations to access the key data for making own assessments.

Each of the elements provided by DFMS significantly improves what is currently available to Uganda's decision makers and communities. The key additional innovation offered by DFMS is the ability to combine, process and disseminate them together within a safe and secure environment.

Combining data allows cross-calibration which enhances the accuracy and reliability of the information products and capability to produce timely and frequently updated high resolution forecasts.

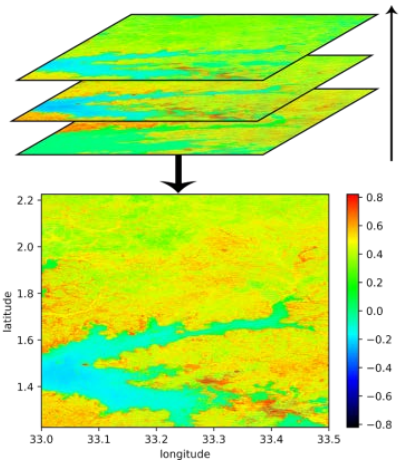
DFMS is deployed in the cloud to provide a reliable and performant service. The computing requirements for supporting these datasets are significant and a cloud based system avoids the cost and complication of a dedicated data centre. The system is accessed via web browser, mobile applications and an Applications Programming Interface

OPEN DATA CUBE

Various big data collaborative platforms exist today, or are in development, with the aim to enable the exploitation of EO data and the creation of value-added products and services. Data cubes represent one type of such platforms, and have the following key benefits:

- The ability to conveniently access, manipulate and visualise long time series of geospatial data (space-based EO products but also other data sets) in order to monitor or detect changes and inform decision makers of such changes;
- The ability to access 'Analysis-Ready' Data (ARD): essentially, ARD data are data that have been processed and organized so that users are not required to invest time and resources in specialized skills to pre-process it (e.g. corrections for instrument calibration, geolocation, radiometry). In addition, ARD products are organised in a standardised structure with associated metadata.
- The possibility to access and combine multiple data sources (e.g. from satellites, airborne vehicles, ground-based instruments);

- The ease of access, down to the pixel level.
- Software open source and packaged as useable as part of a wider application such as DFMS



The open data cube implementation is gaining momentum and enjoys a growing user base. Initially promoted by Geoscience Australia and CSIRO, it is now promoted and developed under the aegis of the Committee on Earth Observation Systems (CEOS). Typically, open data cubes have been so far developed by governments for government use, however RHEA is one of the few commercial companies to develop an open data cube. CEOS estimates that over 25 open data cubes will be operational or in various stages of development by 2020.

The DFMS platform software is also open source and freely accessible in order to foster a community to develop, sustain, and grow the breadth and depth of applications. Other resources such as risk atlases, soil maps, water resources database may be included. In addition models for derived calculations such as crop yields, health etc may be integrated.

DFMS INFORMATION products

Satellite Monitoring

DFMS makes use of the best freely available satellite data from the United States and the European Union includling the new fleet of Copernicus Sentinel satellites (these sources may always be enhanced with new satellites or commercial providers). This is processed into Analysis Ready Data (ARD) for Data Cube storage and further processing and analysis.

The satellite imagery is used by DFMS to support the generation of information about land cover, vegetation, water height, land surface temperature and emissivity and soil moisture.

The products are oriented to local needs and resolutions. For example the vegetation productivity (NDVI) is offered on a per-parish basis or at different spatial scales from 10 km down to 10m if required.

The satellite products are used to directly monitor ground conditions, cross verify with other products, ground data and forecasts. The quality of all products is regularly reviewed and updated as part of the service.

Weather Forecast

DFMS includes a dedicated weather forecast feed from the UK Met Office to support UNMA in delivering forecast information from short range forecasts out to seasonal forecasts.

The forecasts are produced with the needs of users in mind. For example, many stakeholders require reliable information regarding the onset, cessation and intensity profile of Uganda's rainy seasons.

On longer timescales, data from climate models are included for planning purposes.

Hydrological Forecast

DFMS utilises a pair of models that utilise the precipitation forecasts together with information on land cover, soil properties and a land model grid. The model produces a range of forecast outputs including soil moisture, evaporation, surface runoff, and base flow. These outputs are used to predict drought and flood conditions.

Agricultural Services

DFMS makes use of Pictorial Evaluation Technique (PET, already recognised by the World Food Programme) as a means of assessing crop and livestock condition and available as mobile app.

By making visual comparisons of what they see in the field with the photo-indicators in the manuals, assessors can determine the approximate yield of an individual field or area of forage; or score the body condition of an animal. The data collected is also used to inform and calibrate the satellite data products such as vegetation condition and land coverage.

SUMMARY

The Open Data Cube is a key enabler for the provision of a wide range of environmental services

- Uganda-dedicated high precision satellite-derived products
- Seamless weather & climate forecast across multiple timeframes
- Linked hydrological modelling for drought and flood predictions
- Localised two-way assessment in target farming areas
- Integrated on a secure, reliable accessible platform (technology / security)
- Enriching flow of information between all stakeholders
- Anchored in government institutional responsibilities
- First steps toward a fully integrated National Early Warning System (NEWS)
- Supports achievement and monitoring of Sustainable Development Goals
- Extensible and adaptable for many uses
- Building an accessible store of environmental information as a national asset

The AgroMet Data Cube: Developing Smart Data Access for Pest Risk Research and forecasting

Taylor Day[1*], Gerardo López Saldaña[1], Jon Styles[1], Bethan Perkins[1] and Jane Lewis[1].

[1] Assimila Ltd, Reading Enterprise Centre, University of Reading, Earley Gate, Reading. RG6 6BU, UK.

* Taylor.Day@Assimila.eu

Pest outbreaks are devastating to crop yields yet are difficult to predict due to the complex relationship between pest development and environmental conditions. PRISE (Pest Risk Information SErvice) is a 5-year project, funded by the UK Space Agency International Partnership programme (IPP), which integrates openly available Earth Observation (EO) and meteorological data into complex pest development models to provide Sub-Saharan African (SSA) farmers with advanced pest risk warnings to help minimise crop losses. This integration produces indicators of pest biofix dates, lifestage and risk to agricultural crops at national scales, considering the environmental and pest population variability, to provide farmers with location-specific pest risk information. To facilitate this process, Assimila is developing an AgroMet Data Cube (AMDQ) that acquires and organises the input environmental data into a structure that facilitates simple, yet rapid, data access. The AMDQ will automatically access remote data portals, store the data in a central structure and creates seamless and homogenous analysis-ready environmental datasets, irrespective of its format and resolution. It will also support smart spatio-temporal data access by latitude, longitude and time in natural language, as well as traditional indexing using rows, columns and layers. With the derived pest risk information also archived in the AMDQ, real- time pest risk information can be compared with historical trends to enhance our understanding of the current risk to farmers. The project demonstrates the importance of sophisticated data archiving and provides an example of using EO data to produce novel datasets to solve one of the largest problems facing SSA farmers.

Introduction

Crop pests[1] have been prevalent in agricultural ecosystems since their inception some 10,000 years ago (Stukenbrock and McDonald, 2008). They are widespread globally and are estimated to be responsible for global productivity losses of up to 40% (Oerke, 2006; Savary *et al.*, 2012). Pest outbreaks are devastating, respect no political boundaries and are becoming increasingly unpredictable due to climate change (Dhanush *et al.*, 2015; Lamichhane *et al.*, 2015; Boggs, 2016) and are exacerbated by globalisation, international trade and population growth (CGIAR, 2017). The losses caused by crop pests impact farmer livelihoods and international food supply chains globally. Whilst these losses threaten food safety in developed countries (Savary *et al.*, 2012), the impacts of these losses are more prolific in developing countries, such as those in Sub-Saharan Africa (SSA) (Costa, 2014; Paini *et al.*, 2016). This is mainly due to the heavy reliance of rural economies and community development on agricultural productivity, and as such, crop losses have been found to hamper the pursuit of food security, improved nutrition and the ending of poverty (Govereh and Jayne, 2003; Cline, 2007; Diao *et al.*, 2010; Wheeler and Braun, 2013). With smallholder farmers producing up to 80% of the food supply in SSA (IFAD, 2010), it is imperative that innovative methods are developed to help these farmers mitigate against pest-induced crop losses.

Pest development is strongly linked to environmental conditions, with temperature, rainfall and humidity amongst some of the most important factors driving development (Westbrook *et al.*, 2016; Orlandini *et al.*, 2017; Glatz *et al.*, 2017). Therefore, analysing these environmental parameters allows pest development to be modelled and forecasted. Traditionally, methods to do this have used environmental data obtained from a network of *in situ* meteorological stations (Chattopadhyay *et al.*, 2002; Wharton *et al.*, 2008; Landschoot *et al.*, 2013), but their sparse geographic coverage limits the applicability of this approach, particularly when applied to large rural areas. To overcome these issues,

[1] The term 'pests' is used throughout this paper to include insects, mites and plant pathogens.

more recent studies have highlighted the advantages of using EO data for this application, due to its ability to provide well-calibrated and spatially continuous data, presenting a reliable source of environmental information (Da Silva *et al.*, 2015). However, until recently, difficulties associated with accessing and analysing large EO datasets have often been a barrier to their exploitation in this field.

PRISE (Pest Risk Information SErvice) is a 5-year project funded by the UK Space Agency International Partnership Programme (IPP), which exploits EO and meteorological data to provide SSA farmers with advanced pest risk warnings to help minimise crop losses. This will be achieved by:

1. Using a multitude of openly available EO and meteorological data sources to derive the environmental parameters required to model crop pest development, such as air and crop canopy temperature and rainfall.
2. Using these derived parameters to model the development of pests such as *Tuta absoluta* (Tomato Leafminer) and *Spodoptera frugiperda* (Fall Armyworm) to understand pest risk in SSA.
3. Developing an AgroMet Data Cube (AMDQ) to facilitate geospatial data can be access and analysis.
4. With the above datasets archived in the AMDQ, create long-term climatologies that allow current conditions to be compared with historic trends.

The remainder of this paper describes each of these four aims in more detail, outlining the planned approach and innovations that will be developed throughout the project duration.

Using Earth Observation to Understand Pest Risk

The importance of using environmental conditions to model pest development is widely accepted, with temperature being the most frequently used parameter because of its pivotal role in driving development (Wilson and Barnett, 1983). Between an upper and lower development threshold (the optimal zone for pest development), warmer temperatures result in a decrease in development time, as more 'optimal heat' can be accumulated in the same time period (Briere *et al.*, 1999). This heat accumulation is often measured in degree days; a well-documented measure used to model biological development. Through the rearing of pests at a range of constant temperatures, previous studies have calculated the number of degree days required pests to reach each of its life stages (larvae, adult etc.) and to fully develop (Herms, 2004; Tonnang *et al.*, 2015). Using this information, many previous studies have monitored the number of degree days accumulated over a timeframe to understand the current development status of a pest or, when considering future temperatures, the timing of future biological events such as when adult pests will emerge (Chattopadhyay *et al.*, 2002; Wharton *et al.*, 2008; Landschoot *et al.*, 2013). However, many of these studies obtain the input temperature data from the nearest available meteorological station, limiting the applicability of the results for areas distant from a station. High installation and maintenance costs also present challenges when relying on meteorological stations for pest monitoring, particularly over large areas of interest.

EO data can overcome the problems associated with using meteorological stations. It has the advantage of providing a source of spatially continuous and reliable information, reflecting the geographic changes in environmental conditions as opposed to the conditions at specific location. PRISE will primarily use the Meteosat Second Generation (MSG) constellation to obtain the required temperature parameters. MSG is a series of geostationary satellites that acquire data for the whole of Africa at 15-minute intervals with a spatial resolution of ~3-5km (EUMETSAT, 2017). Its onboard SEVIRI (Spinning Enhanced Visible and InfraRed Imager) sensor acquires thermal infrared data which can be used to derive the radiative skin temperature of the land, known as the Land Surface Temperature (LST). Combined with its frequent observations, MSG provides high resolution diurnal temperature data which is critical for modelling pest development during the day and at night (Chen *et al.*, 2015). However, LST is not always indicative of the temperature experienced by the pest, as different pests and life stages can be found in different parts of the environment, such as the in the soil or crop canopy. As a result, PRISE will model environmental

component temperatures which can be used to reflect the pest's actual environment, which should be advantageous compared to using LST for all life stages as in other studies (da Silva, 2015). Furthermore, some pests require more complex development models requiring additional parameters detailing rainfall, humidity and soil moisture. Therefore, PRISE will use existing datasets or derive these parameters to fulfil the input requirements of these complex development models. To ensure the accuracy of the derived parameters and datasets, extensive *in situ* validation also forms a central part of the project.

However, it must be noted that there are some challenges associated with using EO data. One of the main challenges is the impact of cloud cover which prevents the acquisition of data. Through the use of land surface models and integration of modelled meteorological forecasts and reanalyses, PRISE will fill cloud gaps to create seamless and continuous datasets that can be used by the pest development models. A further challenge is the integration of these derived datasets with algorithms to produce information detailing pest biofix dates and lifestage that can be used to detail pest risk at national scales, considering the variability in landscape, climate and pest populations. The translation of EO- scale data to field-level information is critical to provide SSA farmers with location-specific pest risk information to help minimise crop losses. It is paramount to get this information as accurate as possible, and as such, part of the PRISE project involves the rigorous *in situ* calibration of these pest risk indicators through extensive field calibration studies and the development of a crowd-sourcing network, to provide critical feedback that can be used to improve model outputs.

In addition, Assimila will derive and archive the above parameters for a period spanning ~15 years. This allows the generation of 15-year climatologies and allows the current conditions to be compared with the historic trends, providing an insight into whether pest risk in the current season will be worse or in line with previous seasons. Moreover, throughout the duration of the project, data detailing historic pest outbreak records may be obtained, and the long-term archive enables this data to be compared with the environmental conditions at the time to assess whether a relationship exists and may help to extend our understanding beyond simple temperature-driven degree day models.

The AgroMet Data Cube

The proliferation of EO data in recent years has provided both opportunities and challenges to the scientific community (Overpeck *et al.,* 2011), namely the way in which these 'Big Data' sources should be approached. With historical EO archives and further planned missions, there is an urgent need for solutions to organise many different data archives that can be multiple terabytes in size, in an easy-to- use structure.

The concept of a data cube was born from the need to develop a sophisticated 'Big Data' infrastructure to remove the challenges associated with using large multidimensional datasets (Lewis *et al.,* 2017). These challenges can be grouped into three broad categories: access, pre-processing and analysis. The former describes the difficulty associated with initially downloading or getting access to particular datasets, whilst pre-processing challenges capture the complexity of analysis-ready data creation using multiple data sources, each with varying file formats, resolutions and projections. The final challenge relates to a simple way to address this vast wealth of data such that the scientific analysis can take place. Even when all the required data can be stored into a single multidimensional structure, accessing array elements in terms of location and time can be challenging. Facilitating data access is pivotal to ensure that this wealth of data can be better exploited by both advanced and entry level users, particularly if the latter find the initial uptake of EO data to be overwhelming for their application.

Assimila is developing an AgroMet Data Cube (AMDQ) to provide the environmental and pest risk information to PRISE and to other interested parties working in this area, in a structure that also addresses the aforementioned challenges. It removes the data download difficulties through the provision of a set of tools that access multiple data portals and intelligently stores this data in a central structure, forming the Data Cube. The acquired data can then be combined with the state-of-the-art methods described in section 2 to create seamless and homogenous analysis-ready datasets at the continental scale, which can also be

stored in the AMDQ. Examples of these products are, a gap-free LST product derived from MSG and a climatology of surface reflectance derived using the entire MODIS (Moderate Resolution Imaging Spectroradiometer) data record that will allow the inter- and intra- annual vegetation status to be monitored. The process of acquiring and archiving this data will occur in near real-time, providing up-to-date information of the current environmental conditions and pest risk status. Once archived, these datasets then support smart spatio-temporal querying, allowing users to ask for data corresponding to a latitude, longitude and time in natural language, as well as traditional indexing using rows, columns and layers, simplifying the analysis of multidimensional data. All datasets can be stored in their native format, resolution and projection. The AMDQ will allow users to extract data from different sources by specifying a non-native spatial resolution and geographic projection. For example, the AMDQ allows a user to resample a 25m land cover map to 500m using a statistical mode and reproject this data from one projection to another, or to downscale a 4km LST product to 500m using a downscaling algorithm with associated uncertainties calculated.

The AMDQ is currently under development using the Python 3.6 programming language. It consists of two main components. The first contains the server-side operation that automatically downloads and stores data in the Data Cube and subsequently generates analysis-ready data. The second component is an Application Programme Interface (API) that allows users to search for and load data (and metadata) from the AMDQ in a cloud environment, with a suite of tools to analyse each dataset, including the generation of maps and plots. The AMDQ API uses the capabilities of the Geospatial Data Abstraction Library (GDAL) (GDAL, 2018) to handle vast amounts of geospatial datasets in raster and vector format, and the xarray module[2] to perform powerful array operations. An on-the-go AMDQ solution will be provided as well, packaging all the functionalities of the Data Cube in a single file using a Python pickle, allowing users to store and process data locally.

Conclusion

This paper describes the approach and challenges associated with deriving novel environmental and pest-specific parameters from EO and meteorological data for the PRISE project, including the removal of cloud gaps and the derivation of a long-term historic archive. It outlines a smart data structure, known as the AMDQ, which helps to overcome the issues associated with using these data sources and promises to provide geospatial data with a simple interface to remove the barriers linked to the uptake of geospatial data for a wide range of applications. Overall, the project demonstrates the importance of sophisticated data archiving and provides an example of using EO data and novel approaches to solve one of the largest problems facing SSA farmers.

[2] https://xarray.pydata.org/en/stable/

References

Boggs, C.L., 2016. The fingerprints of global climate change on insect populations. *Current opinion in insect science*, *17*, pp. 69-73. DOI:10.1016/j.cois.2016.07.004

Briere, J.F., Pracros, P., Le Roux, A.Y. and Pierre, J.S., 1999. A novel rate model of temperature-dependent development for arthropods. *Environmental Entomology*, *28*(1), pp. 22-29. DOI:10.1093/ee/28.1.22

CGIAR, 2017. *Pests and Diseases*. [Online] Available at: <http://ciat.cgiar.org/what-we-do/pests-and-diseases/> [Accessed 13 April 2018].

Chattopadhyay, C., Agrawal, R., Kumar, A., Singh, Y.P., Roy, S.K., Khan, S.A., Bhar, L.M., Chakravarthy, N.V.K., Srivastava, A., Patel, B.S. and Srivastava, B., 2005. Forecasting of Lipaphis erysimi on oilseed Brassicas in India— a case study. *Crop Protection*, *24*(12), pp. 1042-1053. DOI:10.1016/j.cropro.2005.02.010

Chen, S., Fleischer, S.J., Saunders, M.C. and Thomas, M.B., 2015. The influence of diurnal temperature variation on degree-day accumulation and insect life history. *PloS one*, *10*(3), p.e0120772. DOI:10.1371/journal.pone.0120772

Cline, W.R., 2007. *Global warming and agriculture: Impact estimates by country*. Peterson Institute.

Costa, S. J., 2014. Reducing Food Losses in Sub-Saharan Africa, *An action research evaluation trial from Uganda and Burkina Faso*, pp. 1-21.

Da Silva, J.M., Damásio, C.V., Sousa, A.M., Bugalho, L., Pessanha, L. and Quaresma, P., 2015. Agriculture pest and disease risk maps considering MSG satellite data and land surface temperature. *International Journal of Applied Earth Observation and Geoinformation*, *38*, pp. 40-50. DOI:10.1016/j.jag.2014.12.016

Dhanush, D., Bett, B.K., Boone, R., Grace, D., Kinyangi, J., Lindahl, J., Mohan, C.V., Ramirez-Villegas, J., Robinson, T.P., Rosenstock, T.S. and Smith, J., 2015. Impact of climate change on African agriculture: focus on pests and diseases. *Findings from CCAFS submissions to the UNFCCC SBSTA.* pp. 1-4.

Diao, X., Hazell, P. and Thurlow, J., 2010. The role of agriculture in African development. *World development*, *38*(10), pp. 1375-1383. DOI:10.1016/j.worlddev.2009.06.011

EUMETSAT, 2017. *Meteosat Second Generation (MSG) provides images of the full Earth disc, and data for weather forecasts.* [Online] Available at: https://www.eumetsat.int/website/home/Satellites/CurrentSatellites/Meteosat/index.html [Accessed 25 April 2018].

GDAL. 2018. *GDAL - Geospatial Data Abstraction Library: Version 2.2.4*, Open Source Geospatial Foundation. [Online] Available at: <http://gdal.osgeo.org> [Accessed 13 April 2018].

Glatz, J., Du Plessis, H. and Van den Berg, J., 2017. The effect of temperature on the development and reproduction of Busseola fusca (Lepidoptera: Noctuidae). *Bulletin of entomological research*, *107*(1), pp.39-48. DOI:10.1017/S0007485316000572

Govereh, J. and Jayne, T.S., 2003. Cash cropping and food crop productivity: synergies or trade-offs?. *Agricultural economics*, *28*(1), pp. 39-50. DOI:10.1016/S0169-5150(02)00066-X

Herms, D.A., 2004. Using degree-days and plant phenology to predict pest activity. *IPM (integrated pest management) of midwest landscapes*, pp. 49-59.

IFAD, 2010. Viewpoint: smallholders can feed the world. [pdf] Available at:

<https://maintenance.ifad.org/documents/10180/ca86ab2d-74f0-42a5-b4b6-5e476d321619> [Accessed 16

April 2018].

Lamichhane, J.R., Barzman, M., Booij, K., Boonekamp, P., Desneux, N., Huber, L., Kudsk, P., Langrell, S.R., Ratnadass, A., Ricci, P. and Sarah, J.L., 2015. Robust cropping systems to tackle pests under climate change. A review. *Agronomy for Sustainable Development*, *35*(2), pp. 443-459. DOI:10.1007/s13593-014-0275-9

Landschoot, S., Waegeman, W., Audenaert, K., Van Damme, P., Vandepitte, J., De Baets, B. and Haesaert, G., 2013. A field-specific web tool for the prediction of Fusarium head blight and deoxynivalenol content in Belgium. *Computers and electronics in agriculture*, *93*, pp. 140-148. DOI:10.1016/j.compag.2013.02.011

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., Raevksi, G., Hooke, J., Woodcock, R., Sixsmith, J. and Wu, W., 2017. The Australian geoscience data cube—Foundations and lessons learned. *Remote Sensing of Environment*, *202*, pp. 276-292. DOI:10.1016/j.rse.2017.03.015

Oerke, E.C., 2006. Crop losses to pests. *The Journal of Agricultural Science*, 144 (1), pp. 31-43. DOI:10.1017/S0021859605005708

rlandini, S., Magarey, R.D., Park, E.W., Sporleder, M. and Kroschel, J., 2017. Methods of Agroclimatology: Modeling Approaches for Pests and Diseases. *Agroclimatology: Linking Agriculture to Climate*, pp. 1-10. DOI:10.2134/agronmonogr60.2016.0027

Overpeck, J.T., Meehl, G.A., Bony, S. and Easterling, D.R., 2011. Climate data challenges in the 21st century. *Science*, *331*(6018), pp. 700-702. DOI:10.1126/science.1197869

Paini, D.R., Sheppard, A.W., Cook, D.C., De Barro, P.J., Worner, S.P. and Thomas, M.B., 2016. Global threat to agriculture from invasive species. *Proceedings of the National Academy of Sciences*, *113*(27), pp. 7575-7579. DOI:10.1017/S0021859605005708

Savary, S., Ficke, A., Aubertot, J.N. and Hollier, C., 2012. Crop losses due to diseases and their implications for global food production losses and food security. *Food Security,* 4 (2), pp. 1-6. DOI:10.1007/s12571-012-0200-5 Stukenbrock, E.H. and McDonald, B.A., 2008. The origins of plant pathogens in agro-ecosystems. *Annual Review*

of Phytopathology, *46*, pp. 75-100. DOI:10.1146/annurev.phyto.010708.154114

Tonnang, H.E., Mohamed, S.F., Khamis, F. and Ekesi, S., 2015. Identification and risk assessment for worldwide invasion and spread of Tuta absoluta with a focus on Sub-Saharan Africa: implications for phytosanitary measures and management. *PloS one*, *10*(8), p. e0135283. DOI:10.1371/journal.pone.0135283

Westbrook, J.K., Nagoshi, R.N., Meagher, R.L., Fleischer, S.J. and Jairam, S., 2016. Modelling seasonal migration of fall armyworm moths. *International journal of biometeorology*, *60*(2), pp. 255-267. DOI:10.1007/s00484-015- 1022-x

Wharton, P.S., Kirk, W.W., Baker, K.M. and Duynslager, L., 2008. A web-based interactive system for risk management of potato late blight in Michigan. *Computers and Electronics in agriculture*, *61*(2), pp.136-148. DOI:10.1016/j.compag.2007.10.002

Wheeler, T. and Von Braun, J., 2013. Climate change impacts on global food security. *Science*, *341*(6145), pp. 508-513. DOI:10.1126/science.1239402

Wilson, L. and Barnett, W., 1983. Degree-days: an aid in crop and pest management. *California Agriculture*, *37*(1), pp. 4-7.

Data Policy of Institute of Space and Astronautical Science (ISAS) at JAXA

Ken EBISAWA (ISAS, JAXA)

Institute of Space and Astronautical Science (ISAS) is the sole science center at Japan Aerospace Exploration Agency (JAXA). So far, ISAS has not had an established data policy, while necessity of such a data policy has been recognized. The new data policy has been discussed extensively in 2017 and finally approved in March 2018. That will be publicly announced at ISAS home page shortly.

1. Background

ISAS is the sole center of space science in Japan, and archives science data primarily taken by JAXA's space science missions. So far, ISAS has not have an established data policy. Necessity of the data policy has been recognized recently, and data policy issues were discussed extensively in 2017 under a committee under ISAS established to discuss data-related matters. The policy was finally approved in March 2018. The policy is written in Japanese, and the English version is being drafted. In this paper, I will introduce essence of the ISAS data policy in the form of draft of the English version.

2. Data policy of ISAS (Draft)

2-1. Purpose

This data policy describes the principle of handling data when Institute of Space and Astronautical Science (ISAS) carries out its own scientific research and development concerning space sciences.

2-2. Application

This policy applies to the data created or processed by ISAS. Furthermore, for the data taken or processed by other establishments (e.g., Japanese universities, foreign space institutes), where ISAS provided the required instruments, launch opportunity, budget, human resources, etc., we hope the philosophy of this data policy is respected by those parties.

2-3. Scope of the data where this data policy is concerned.

We define the scope of the "data" covered by this policy: *information having broad scientific values that can be universal and long-term usages, not relying on particular physical medium.*

For example, followings are not included in the current scope of the "data"; personal notes or photos, informal minutes of meetings, temporal information not intended for permanent use, physical samples from celestial bodies (such as asteroids) or samples used by micro-gravity experiments.

Below, we give representative examples of the data in the current scope, together with their brief explanations:

- Source data/Raw data : Satellite telemetries, as such, which have to be parsed according to specific formats, and will be the source of all the following data processing. After data processing, "Observation data" and "Engineering data" (see below) will be produced
- Observation data : Such numerical data taken by artificial satellites, space probes, balloons, sounding rockets etc., that describe physical conditions of celestial bodies or space phenomena that *cannot* be controlled by observers. Many space phenomena show significant time variations, so that observation data may not be reproduced.
- Engineering data : Numerical data of the physical conditions such as orbits, attitudes, temperature etc. of the *data-measuring side* (satellites, space probes or instruments etc.).
- Experimental data : Numerical data obtained by *intentionally controlling* the objects by the observers. In most cases, identical or similar data may be reproduced by repeating the same experiments

- Simulation data：To simulate observation data, engineering data, experimental data etc., numerical data artificially produced by computation. To repeat computations, the identical data can be reproduced.
- Digitized data resulting from analysis of the samples of celestial bodies or microgravity experiments. Numerical shape-model data representing shapes of observed celestial bodies. These data may be updated when accuracy of the analysis or reproduction is improved.
- Digitized documents, photos, pictures, videos, etc., which are created aiming for long-term and universal use.

In addition, we consider the followings entities as the data covered by the present policy; data to describe the data (meta-data), tools and software required to use the data, algorism to create the data, explaining documents of the data.

2-4. Proprietary data and open data

All the data have to be either "open/public" or "proprietary/close". We will judge if each type of the data will be open or private, considering individual circumstances.

Those data which will be evidence of published scientific results have to be open. Generally, data are expected to be made public, except when (1) it may cause harm for public security or privacy if made open, (2) it may cause scientific demerits if made open (for instance, calibration is uncertain), (3) the data are used exclusively by the team who created data for a certain period of time, and (4) right to use the data is granted to particular people or parties for a certain period of time.

In principle, we will indicate presence of the proprietary data with the reasons that they cannot be open, except when not clarifying presence of the proprietary data has some merit. For proprietary data, who can access the data and the proprietary period are defined. When the proprietary period is over, the data may become public, continue to be proprietary, or be discarded.

For the observation data taken by satellites, space probes, balloons, and sounding rockets etc., if proprietary period is required for the instrument team to calibrate the instruments, or for the proposers/observers to analyze the data exclusively, the nominal proprietary period is one year.

2-5. Policy of open data

Based on the belief that global and long-term use of the open-data will continue to the progress of science, we will carry out the following measures:

1. Carry out proper data processing or provide data explanation, so that using the open data does not require any proprietary knowledge.
2. Open data are kept at least 30 years in a usable condition.
3. Provide free services to easily find open-data and enhance usability of the data.
4. Open data will be made citable using, e.g., Digital Object Identifier (DOI).

2-6. Rule when using the open data

When using open data, we request to use the following rules. These rules follow Japanese government standard of data-use (second version), and compatible with Creative Commons BY 4.0:

1. In principle, open data can be used for free of charge, either commercial or non-commercial, including copying, sending and modifying. However, for some open data kept at ISAS, those establishments

beside ISAS concerned with production of the data may impose other restrictions.

2. When using data, please indicate origin of the data as "ISAS/JAXA". However, there is a case that those establishments contributed to the data production request additional acknowledgments; data users are expected to follow those requests.
3. When the data are modified, please indicate the fact that the data are modified, and indicate, as much as possible, what kind of modification is made.
4. We are not responsible for whatever actions resulted by data-users.

2-7. Preservation of data

In principle, we try to keep all the data as long as possible. However, considering costs and resources, we may have to discard data in some cases. Following is the guide line to preserve/discard data:

- Public data are kept for a long-term, no less than 30 years.
- Data which cannot be in principle reproducible will be kept for a long-term.
- Source data (raw data) from which other data are produced by data processing are kept for a long-term in a condition that the reprocessing is possible.
- Data which can be in principle reproducible but requiring a large amount of resources, are not discarded as much as possible.
- Data which can be easily reproducible may be discarded.

Developing improved workflows and tools for preserving and exploiting environmental research data – a case study from the NERC National Geoscience Data Centre (NGDC)

A.T. Riddick[1], A. Fernie[2], J.O. Pinnick[1], and G. R. Baker[1]

[1]British Geological Survey, Environmental Science Centre, Nicker Hill, Keyworth, NG12 5GG

[2] British Geological Survey, The Lyell Centre, Research Avenue South, Edinburgh, EH14 4AP

The increase in the volume and variety of environmental research data available over recent years has resulted in a number of challenges for repositories seeking to ingest, store, and make available this data. Some of the key challenges include the need to rapidly ingest data from a diverse range of data providers and in a range of file formats, an issue common to the environmental sciences as well as to other science disciplines. In parallel with these trends many public sector repositories are also realising a requirement to achieve more, often with restricted resource allocations. These data driven and economic factors have prompted the NERC National Geoscience Data Centre (NGDC) to invest in the development of improved workflows and new software tools, in order to continue to meet our objectives ensuring the future usability of our geoscience and environmental datasets long into the future.

One of the key requirements for the NGDC is to ingest and manage geoscience sub-surface data from a range of geoscience sub-disciplines. This data is obtained from a range of sources which includes Natural Environment Research Council (NERC) funded grant projects undertaken by universities and or research organisations. Other data is sourced through statutory data donations, such as under the Mines and Quarries Act 1954 and the Science and Technology Act 1965. The NGDC receive notification to drill and subsequent borehole data where the depth of the borehole is more than 100 ft. Similarly, NGDC also receives borehole data where a well is sunk at a depth of 50ft or more for the purposes of searching or abstracting water. Other data (for example, additional geotechnical data and details of other sub-surface investigations) is received through partnerships with other government departments and local authorities through voluntary deposits.

NGDC faces a requirement to hold a wide range of data from historically datasets captured prior to the advent of more modern data management processes through to real time data about geological processes sourced from sensor or monitoring networks for example. This diverse data can include a wide variety of file formats which may include proprietary data formats. In addition since the NGDC stores and manages data which ranges from academic research data through to commercial and other sensitive data there are challenges in understanding and recording the conditions of deposit, so that these are correctly understood by current and future users of the data.

In order to focus efforts on making data usable and accessible in the longer term, a data preservation strategy is being developed, and this includes an “acceptable” list of file formats for donation has been developed. This list focusses on software independent file formats (such as .csv and .txt for data), and general office based file formats that although proprietary are XML based. This constraint is currently only a recommendation, in view of the wide variety of file formats produced across the earth sciences, however we are strengthening this further year upon year.

As well as the format, the diversity in the source of the data received by NGDC has presented a particular set of requirements in prioritising and appraising data. The workflow for data ingestion involves accessioning the data (by recording high level metadata including a description of the data received), and then undertaking an evaluation of the value of the data to the data centre’s designated community using a “Data Value Checklist”. At this stage any terms and conditions which govern the on-going use of the data (for example an embargo on other researchers using that data, or commercial restrictions due to

confidentiality) are also considered and recorded within the metadata.

Another challenge faced by NGDC concerns the initial receipt of data, particularly where data is not deposited with the NGDC directly, for example British Geological Survey (BGS) owned research centre data funded by NERC, may have originally been provided to a BGS scientist because of a professional connection or a particular project, and this scenario may prevent the crucial collection of terms and conditions before accepting data. A key requirement for NGDC is providing one consistent and simple method to deposit data.

With the increasing volume of data donations received a requirement was identified for a streamlined and automated system for users to donate their data and feed this directly into the data centre's appraisal and archiving workflows. The NGDC Data Deposit portal (http://transfer.bgs.ac.uk/ingestion) was designed and constructed to meet this requirement, providing one method to upload data with metadata. The portal captures a contextual description, the ownership and geographic location of the data. Users can also input access restrictions such as whether any of the data is confidential, and the intended release date (if appropriate and in accordance with our policies). This provides a means of capturing the essential information about a data deposit without requiring onerous data input by the user. The final step is to upload the data files to be deposited through the portal. Providing a simple method to deposit data and crucially capture the terms of the deposit helps encourage all deposits to follow this route.

The data deposit portal automatically feeds the data submitted into the data ingestion workflow allowing the data to be processed into the detailed accessions system. Historically, some components of our original data centre workflow had developed independently of each other, and the development of the Data Deposit portal has provided an opportunity to streamline the flow of data through the system and replace several previously manual steps with database transactions to increase efficiency and reduce data errors.Staff resources are inevitably limited, but have always been used flexibly and to maximum benefit to for example accession new data received into the organisation, appraise the value of this data and then ingest the data into relevant data stores, databases and collections.When new datasets appeared, there was often a need to build a new database to store and provide access or the data never went beyond the first accession level. The single onshore borehole index (SOBI) is a shining example of data being processed through to a database built to hold a particular type of dataset, in this case a National Dataset containing borehole data (i.e. a dataset adding value by combining and integrating several appropriate source datasets). Processing the data to that level of granularity takes a lot of resource.
Realistically taking into account economic constraints it is becoming more difficult to be able to process all of the data received through this workflow cycle (and continue to keep building specific databases for datatypes). The original Accessions system was seen as too basic and our corporate databases too complex and some middle ground was required to help provide more discovery and access to the data without having to, build new databases. There has also been an increasing need to introduce some level of prioritisation in order to identify the data which is most needed by the repositories stakeholders at an early stage in the process.

In order to approach this data description, appraisal and prioritisation step in a systematic manner and to provide an audit trail of the operations undertaken by data centre staff, recent developments have been progressed to streamline and improve workflow processes. An important new feature is an extension to database systems to record additional details about the accession and incorporate some of the functionality previously provided by existing corporate databases, such as linking to digital files and to site the location of the data on a map. Other contextual metadata recorded at this stage includes assigning a data category (e.g. grant funded, statutory or voluntary donation).

Whilst simple in concept this "Detailed Accessions" procedure provides a rich level of contextual metadata

about the deposit, linked to information about the terms and conditions. Some deposits will never go beyond this stage, but the resulting output is a database which can be utilised by data centre staff in order to make decisions about prioritisations for archiving of the data within the repository data stores, and also by management to support strategic longer term decisions about on-going operation of the data centre. Importantly this information can also be searched by users to obtain more detailed information about the data than would be available from the initial accession information alone.

The data deposited and ingested within the NGDC can be searched using a web based search interface located upon the NGDC pages within the BGS website, including statutory donations, NERC funded grants (from academia or research centres) (http://www.bgs.ac.uk/services/NGDC/dataDeposited.html). This interface searches the data held in the detailed accessions database and allows a number of search optionsincluding a free text search, supplemented by a more advanced search allowing search by attributes such as title or description, subject depositor and the media upon which the data is available. There is also a web based geographic search facility. The searches produce a results listing which provides details of any access restrictions relating to the data. Where there are restrictions on use details are provided of who to contact to provide further information about the data. The majority of openly available data deposits can also be down-loaded from this interface

Finally, having reviewed the workflow for all data deposited and having the ability to review the types of data being deposited, a specific new automated workflow for the AGS (Association of Geotechnical and Geoenvironmental Specialists) data exchange format has now also been developed. The AGS format is a self-describing data transfer format which can be programmatically ingested, and the intention is to further automate ingestion workflows for other data types as resources permit.

Following the implementation of the improved workflows, and web interface developments NGDC submitted an application for accreditation under the CoreTrustSeal scheme, and in January 2018 became the first of the NERC data centres to receive this accreditation.

The CoreTrustSeal of approval respository certification is administered by the ICSU World Data System (WDS) and the Data Seal of Approval (DSA) and provides accreditation of a trustworthy digital repository. This certification is designed to confirm that the data held by a repository is managed in line with current and emerging best practice so that the data remain accessible and usable into the future. The CoreTrustSeal accreditation process covers a number of facets of repository operation and data management, including for example data integrity and authenticity, storage procedures, workflows, and data re-use. The process to attain certification involves submitting a self-assessment for each of these categories.

The NGDC was assessed as having fully implemented a number of the requirements of the certifying body including those regarding workflows, and data re-use. A compliance estimate of either three (being implemented) or 4 (fully implemented) for most of the other categories assessed (particularly data appraisal, storage procedures, and data preservation) was attained. The positive feedback received has allowed NGDC to further hone a number of its procedures.

A Standard Reference Model for Data Archives

John S. Hughes[1], Daniel J. Crichton[2], David Giaretta[3], Mike Kearney[4], Robert R. Downs[5], Matthias Hemmje[6], Rosemarie Leone[7], Terry Longstreth[8], John Garrett[9], Sean Hardman[2], Ronald Joyner[2], Michael McAuley[2], Costin Radulescu[2], Bruce Ambacher[10]

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA, John.S.Hughes@jpl.nasa.gov

[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA

[3]PTAB Ltd, Dorset, UK

[4]Sponsored by Google, Huntsville, AL 35803, USA

[5]Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY, USA

[6]FernUniversität in Hagen, Hagen, Germany

[7]European Space Agency (ESA), Frascati, Rome, Italy

[8]Data and Information Standards Consultant, Laurel, MD, USA

[9]PTAB Consultant, Columbia, MD, USA

[10]Research Associate, Digital Curation Innovation Center, College of Information Studies, University of Maryland, College Park, MD

An implementable Data Archive Architecture is being developed for trusted digital repositories based on the Reference Model for an Open Archival Information System (OAIS) – ISO 14721. A set of interoperable protocols and interface specifications are planned that will offer capabilities for accessing, merging, and re-using data, both within and across the operational boundaries of trustworthy digital repositories. The model will also provide support for the fundamental scientific need to verify the reproducibility of results.

This standards development task is being performed by the Data Archive Interoperability (DAI) working group within the Consultative Committee for Space Data Systems (CCSDS). The architecture integrates concepts from the OAIS Reference Model, the ISO/IEC 11179 Metadata Registry (MDR) standard, the CCSDS Reference Architecture for Space Information Management (RASIM), the proposed draft recommended practice document, Information Preparation to Enable Long Term Use (IPELTU), and three decades of digital repository development for science research.

Keywords

Interoperability, Information Architecture, Information Model, Digital Repository

1.0 Introduction

Long-term digital preservation methods are critical practices to preserve the benefits of space missions for the mission customers and our society as a whole. A new concept is presented, a Data Archive Architecture called the OAIS Interoperability Framework (OAIS-IF) [2]; a software architecture suite of standards that will support the OAIS and add capabilities for interoperability between users and all archives that comply with the OAIS-IF standards. This architecture is derived from and supports the Open Archival Information System (OAIS) Reference Model (RM) [1] approach developed by the Consultative Committee for Space Data Systems (CCSDS) and adopted by archives world-wide.

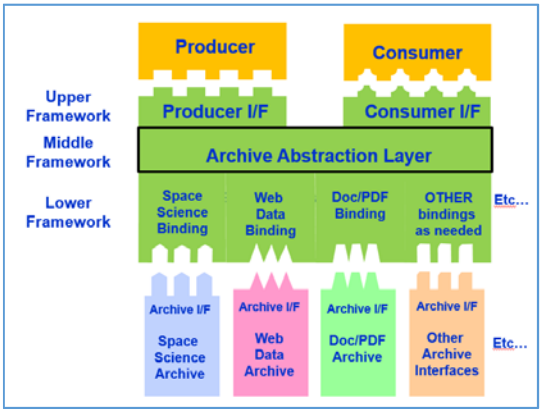
Establishing the OAIS-IF provides stakeholders with a conceptual understanding of the components and the relationships among those components needed to implement systems that can foster acquisition, stewardship, and continuing access to data products and related information resources and services that have been selected for use by designated communities. Identifying and describing the components and relationships described within the OAIS-IF enables stakeholders, including data producers, archivists,

repository managers, and sponsors to plan, design, procure, implement, and operate systems that support the long-term use of data that are currently being collected or stored. Furthermore, understanding the components and the relationships depicted within the OAIS-IF enables repository system developers to create and acquire the services and features needed to provide the functionality that data stewards can utilize to provide data producers and users with the capabilities needed for sharing and using data, respectively.

2.0 The OAIS Interoperability Framework

The first OAIS-IF product that will be produced by the DAI WG is the Data Archive Architecture Description Document (DA-ADD). The ADD will describe the overall framework from top to bottom. It is a prerequisite to insure a compatible end-to-end design of the other components in the OAIS-IF architecture. A conceptual framework [2] is illustrated in Figure 1. The upper framework provides the user interfaces for the producers and consumers of the OAIS archive data.

The middle framework is an Archive Abstraction Layer (AAL) with the function of obscuring and encompassing the implementation details to facilitate interoperability and archive platform independence. Example interactions that pass through the AAL would include methods to allow producers and consumers to interact with an archive using OAIS standard informational entities including the Submission Information Package (SIP); Archival Information Packages (AIPs), Dissemination Information Package (DIP) and their components.



Finally, below the framework are the specific plug- in components that connect the framework to a domain archive. These components map the standard OAIS information entities and bind OAIS-
IP service interfaces to their counterparts in thedomain specific archives. These plug-in components

*Figure 1 - The overall high-level structure of OAIS-IF*also expose domain specific information models and services.

3.0 An Example Interoperable Infrastructure – The Planetary Data System

In 1982 and 1986, the National Research Council (NRC) Committee on Data Management and Computing (CODMAC) issued reports that set guidelines for the development of science data archives. Based on the CODMAC guidelines, the Planetary Data System (PDS) was established in 1989 and since then has aggressively promoted long-term preservation in the Planetary Science community. In Oct 2013 the PDS released PDS4 [3, 4], a complete redesign of its system using lessons learned and principles set forth in the Open Archival Information System (OAIS) and ISO/IEC 11179 [8] reference models. Soon afterwards, the International Planetary Data Alliance (IPDA), a consortium of international space agencies, endorsed the PDS standards and set the course for bringing together the entire science discipline under one interoperable infrastructure.

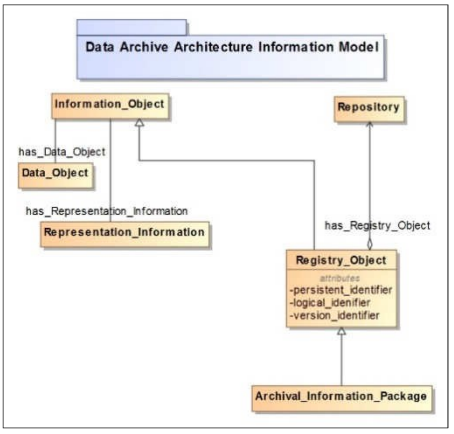


Figure 2- Information Components

4.0 Status of the Architecture Description Document

A draft of the Data Archive Architecture has been created using concepts from sources that include the Reference Architecture for Space Information

Management (RASIM) [5], PDS4 specifications [3, 4], the Data Archive Ingest (DAI) WG Report to the CCSDS Management Council [6], SCIence Data Infrastructure for Preservation – Earth Science (SCIDIP-ES) [7], and reports and recommendations from the Research Data Alliance (RDA) Data Fabric WG. The draft model is being captured and managed in the Cornerstone Framework (NPO-49832), the framework used to in the development of the PDS4 Information Model. A few of the information model and service components are illustrated in Figures 2 and 3 respectively.

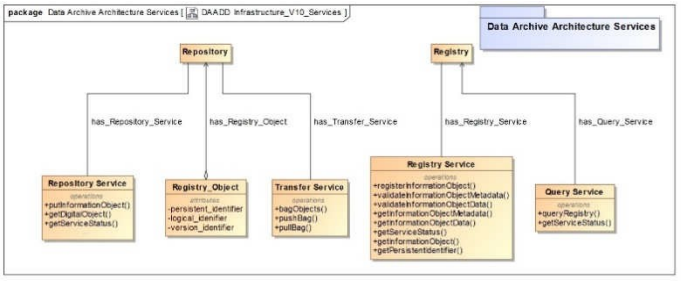


Figure 3 - Service Components

5.0 Conclusion

In 2002, the National Virtual Observatory Science Definition Team said, "It is probably safe to say that no other professional community has reached the level of data interchange standards (both syntax and semantics) that we have reached in astronomy." [9] The international planetary science community has met this challenge with the development and adoption of the PDS4 Information System. The OAIS-IF is introduced as a conceptual model for improving understanding of the components and relationships needed for planning, designing, developing, acquiring, and operating digital repositories that enable the submission, management, preservation, and continuing use of data products and related resources that have been identified as having enduring value for designated communities. The universal acceptance of the OAIS RM provides the opportunity to extend these successes by developing an OAIS Interoperability Framework that will enable interoperability across data archives in general.

6.0 Acknowledgements

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (NASA). Support for Robert Downs was received from NASA under contract NNG13HQ04C for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC).

References

[1] ISO 14721:2003: Reference Model for an Open Archival Information System (OAIS), ISO, 2003.

[2] Kearney, M., et. al., Digital Preservation Archives – A New Future Architecture for Long-term Interoperability, To be published, 15th International Conference on Space Operations, Marseille France, May 2018.

[3] Planetary Data System - PDS4 Information Model Specification, Version 1.8.0.0, March 2017.

[4] Planetary Data System - PDS4 System Architecture Specification September 1, 2013, Version 1.3.

[5] Reference Architecture for Space Information Management (RASIM) CCSDS 311.0-M-1

[6] Data Archive Ingest (DAI) WG Report to the CCSDS Management Council (CMC), Figure 2: Notional Data Archive Architecture, March 2017

[7] Science Data Infrastructure for Preservation-Earth Science Data. (2012). Retrieved April 28, 2018, from www.scidip-es.eu.

[8] ISO/IEC 11179: Information Technology -- Metadata registries (MDR), ISO/IEC, 2008.

[9] NVO Science Definition Team, “Towards the National Virtual Observatory”, A Report Prepared by the National Virtual Observatory Science Definition Team, 2002.

Evolution of CNES Tools and Processes for Long Term Preservation of Space Science Data

Claire Caillet, Benoît Chausserie-Laprée, Dominique Heulet

CNES, 18 av E. Belin, 31401 Toulouse Cedex 9, France

This paper presents future evolutions of CNES activities to ensure long-term preservation of data from space missions involving CNES. During the next few years, support to projects activities, data inventory activities and data access services will evolve significantly. The main goal of the inventory is to identify, for space missions involving CNES, all data collections and associated information, for CNES self-use and for data discovery through a dedicated website (REFLECS). Despite the CNES efforts to keep up to date and to improve the inventory, this activity is reaching its limits: many manual actions, redundancy of information, there is no link between data discovery and data access. REGARDS, a new system under development, will bring new possibilities to improve CNES data inventory and access in a technological and organizational point of view. Thanks to the rise in web technologies, data access catalogs flourished and the CNES has regularly updated its systems. Constant evolution of web technologies, increase in data volume and new interoperability needs, lead to the replacement of old tools by new and more performant ones. The current systems providing data access services will be replaced during the years 2019-2020 by a unique system, REGARDS. In the next few years, the organization of data inventory and data access activities will evolve around the REGARDS development. These technological evolutions will go with updates of the CNES process to take into account the data long-term preservation as soon as possible in the project lifecycle.

Introduction

CNES manages a great variety of space missions addressing various topics from Earth Observation to Astronomy, including physical sciences and technology. Data produced during those missions can be processed, archived and distributed whether by CNES or by scientific laboratories.

CNES is responsible for the inventory of all the data produced by these missions: it relies on the SERAD (Service for Data Archiving and Referencing) inventory and referencing tools.

CNES hosts dedicated data centers for the long-term archiving of space science data, for example:

- The CDPP (French Data Center for Plasma Physics) which is in charge of insuring the long-term preservation of data obtained primarily from instruments built using French resources, in the Solar System natural plasmas domain,
- The SERAD which is in charge of the long-term archiving of space science data issued from missions involving CNES, in a large variety of domains: Astronomy (HIPPARCOS, COROT), Earth Observation (POLDER, SPOT,
  …), Space Sciences (PHEBUS, PHOBOS, …),
- AVISO in the Altimetry domain.

CNES also delegates the long-term archiving to certified data centers: MEDOC (Solar Physics), VITO (Earth Observation).

The SERAD context

SERAD missions are to:

- Constitute the inventory of space data issued from missions involving CNES,
- Reference the inventoried data and give an open access to them,
- Archive over the long term, when necessary, the space data that are under CNES responsibilities (heritage mission, earth observation, space sciences, …) and give an access to these data to users.

Concerning the inventory objective

For each space mission, an inventory is established. A form describes the mission characteristics, the location and the content of each dataset that has been archived. It identifies also the documents describing these datasets, the scientific labs and scientists involved. Lastly, it lists the available means to access to these data. Concretely, the

inventory consists in a CNES intranet site: figure 1 gives an extract of the summary table of the inventoried space missions.

missions inventory

Mission ▼▲	Domain ▼▲	Comments ▼▲	Last validation ▼▲	Last update ▼▲
ALICE	MATERIAL_SCIENCES	ALICE experiment, from ANTARES, ALTAIR, CASSIOPEE, PEGASE and PERSEUS missions.	2012/07/02	2012/07/02
ARCAD3	UNIVERSE_SCIENCES	All french experiments	2012/07/26	2012/07/26
CASSINI	UNIVERSE_SCIENCES	CASSINI-HUYGENS : experiment RPWS	2012/09/24	2012/09/24
CASTOR	TECHNOLOGY	CASTOR experiment, from CASSIOPEE, PEGASE and PERSEUS missions.	2012/07/02	2012/07/02
CHAMP	EARTH_SCIENCES		2011/12/21	2011/12/21
CLUSTER_CIS	UNIVERSE_SCIENCES	CODIF and HIA experiments	2009/11/26	2010/10/05

Figure 1: CNES inventory website (intranet only)

About the referencing objective

The information contained in the inventory forms are used as input to feed the CNES referencing tool REFLECS (REFerence catalogue of Long term CNES Scientific data) which is the CNES clearinghouse (http://reflecs.cnes.fr).

The objective is to assist users to efficiently locate information on available data on a specific scientific domain, or through different domains. It also provides a centralized and overall view on the CNES data patrimony on the long term: users can find descriptions on data collected over more than 30 years.

About the long term archive objective

SERAD relies on CNES infrastructure dedicated to long term storage, the STAF (File Archiving and Transfer Service), and on the generic tool SIPAD-NG (Système d'Information, de Préservation et d'Accès aux Données – Nouvelle Génération) to access the data.

Lessons Learned
About the inventory

The inventory is based on old tools which are not up-to-date. The procedures to fill the inventory forms are essentially manual: they are subject to human errors and to duplication of information. The inventory forms are based on XML format and conform to the inventory schema (XSD): the metadata which are generated for the ingestion in REFLECS conform to the ISO 19115 standard, this result in a complex structure of the form.

The structure of the inventory form is 'mission' oriented: it is adapted for a given space mission which is composed of one or more platforms, each platform having on board one or more experiments, each of these experiments generating different datasets. With this structure it is difficult to search for a specific dataset and to know for example which missions generated it (for example: a level 1 HRV (High Visible Resolution) data set can be linked to one of the SPOT 1 - 4 platforms).

The inventory form:

- is not adapted to data produced by a series (ex: SPOT, JASON or Pléiades series in the Earth Observation domain),
- is not adapted to data produced from several experiments or satellites (multi-mission data).

About the referencing tool (REFLECS)

The referencing tool resulted from R&D development and is very rarely used among the scientific community. The main reasons are:

- the MMI is based on old technologies, it is not accessible through a programmatic interface,

- the MMI offers poor search criteria for discovery,
- only the mission and experiment levels are shown, not the dataset level,
- there is no link between data discovery and data access: the access to the data is possible only through a link to the external data server which hosts the data (i.e. all the selection process has to be replayed to access the desired data).

About the long term archive

SERAD created 3 archives:

- one for the CoRoT space telescope (Convection, Rotation and planetary Transits) level 1 data,
- one for the POLDER instrument data (POLarization and Directionality of the Earth's Reflectances),
- one to archive various historical space mission data.

There is no unique access point to all these data and the 'old data' archive should be refactored.

SERAD Refactoring

It is now necessary for the SERAD to change its tools and re-engineer its processes in order to:

- facilitate discovery and access to data,
- facilitate updates of the inventory,
- facilitate data ingestion and transfer to long term archive,
- provide the data catalog with access services in order to avoid the manual copies and duplication of information.

REGARDS

In order to be able to discover, select, retrieve, use or reprocess space data for years to come, data have to come with different elements: metadata, documents, software, quicklooks, … as shown in figure 2.

A Catalog Access System shows relationships between data and its related information. Since the emergence of web technologies, CNES developed several generations of Catalog Access Systems (figure 3). Some of these systems have been used by CNES Mission Centers or Data Centers (SIPAD, SIPAD-NG). Some of them have been used by scientific laboratories to provide access to their own data (SITools, SITools2).

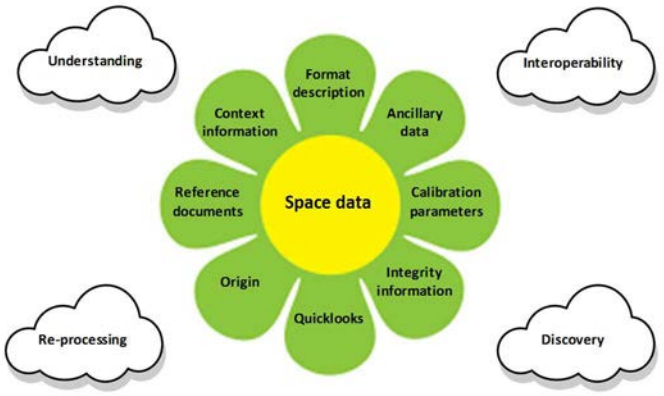


Figure 2: information associated to data systems

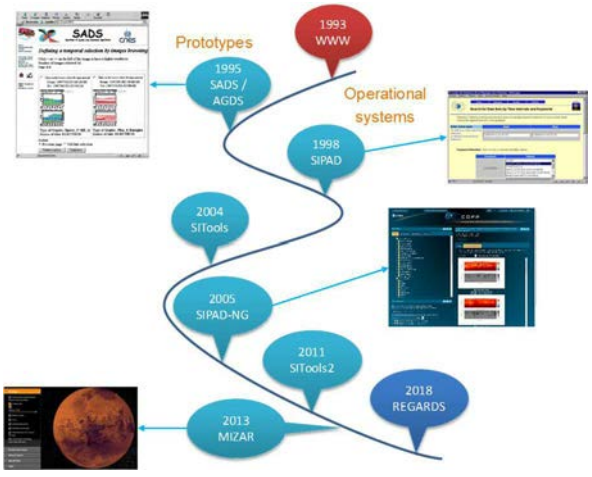


Figure 3: the long history of catalog access developed by CNES

REGARDS ('REnewal of Generic tools to Access and aRchive Space Data') is the next generation of Catalog Access Systems. It will be used to implement Mission or Data Centers located at CNES or in partner laboratories. REGARDS will replace both SITools2 and SIPAD-NG systems currently in use. REGARDS will be able to cope with huge data volumes expected from space missions in the 2020 and beyond.

REGARDS functions and architecture

REGARDS main functionalities rely on the OAIS (Open Archival Information System) model for long-term preservation and access to digital data, as shown in figure 4.

REGARDS relies on a web-oriented architecture. The 'frontend' (user and administrator MMI) is a JavaScript application. The 'backend' is composed of several JAVA micro-services. Each micro-service is a web-server providing REST endpoints (figure 5).

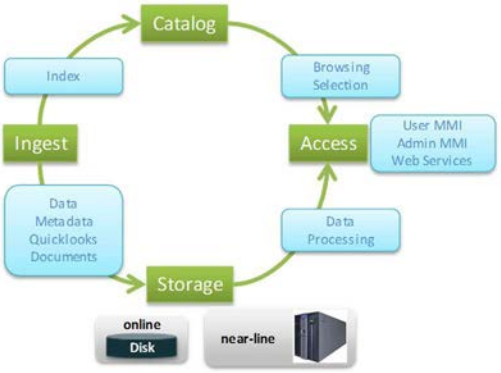


Figure 4: REGARDS functions

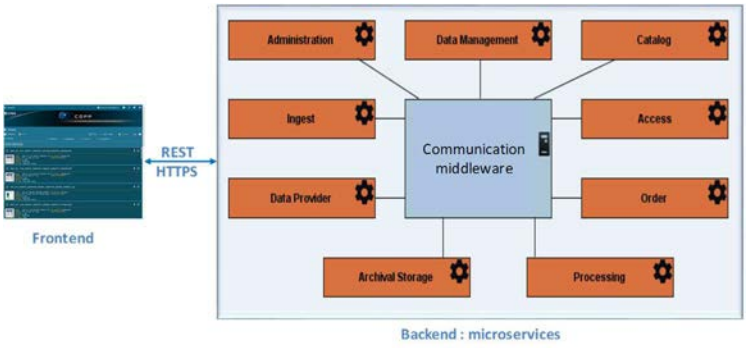


Figure 5: REGARDS architecture

REGARDS is a framework: plugins are used to adapt each component (micro-services or frontend) to the characteristics of the mission or data center.

Use of REGARDS

REGARDS will be the baseline to implement CNES Catalog Access Systems:

- It will replace all SIPAD-NG catalogs,
- It will be used by new Data Centers. The first one will be Microscope Data Center (microsatellite to verify Einstein's Equivalence Principle). The next implementations will be for CFOSat ('China-France Oceanography SATellite'), SWOT (French-US 'Surface Water and Ocean Topography' mission), French-German MERLIN satellite ('Methane Remote Sensing Lidar Mission'). REGARDS will also be used to implement the archive of 30 years of imagery data ('SPOT World Heritage' project).

REGARDS will also be used by scientific laboratories, especially IAS ('Institut d'Astrophysique Spatiale').

REGARDS and SERAD

Through its wide use in many Missions or Data Centers, REGARDS will be able to provide all metadata to automatically feed the inventory of CNES data collections.

On the other hand, REGARDS architecture is flexible enough to implement new software which is going to replace both inventory web server and REFLECS.

CNES Processes

Tools without methodology are not sufficient. It is important to take into account the Long-Term Preservation very early in the development cycle of a Space Project. For this purpose, the CNES applies the PAIMAS methodology (Producer Archive Interface Methodology Abstract Standard) which defines the relation between the Producer (or Project) and the Archive.

The CNES methodology has been updated in order to systematically support the projects in development from the early phases (A & B phases). The main tool is the Preservation Plan: it will guide the Project during all its life, from the early phase (A & B phases) until the end of operation phase (E phase). The Preservation Plan is a subset of the Data Management Plan: it enables the project in development to comply CNES requirements for long-term preservation.

The preservation plan fulfills the following objectives:

- Identification of the long-term preservation needs at the beginning of the project (Mission & System requirements), especially the Archive in charge of long-term preservation,
- List of datasets, documents and software to be preserved,
- Definition of the submission procedures to each Archive.

The different steps of the preservation process are shown in figure 6.

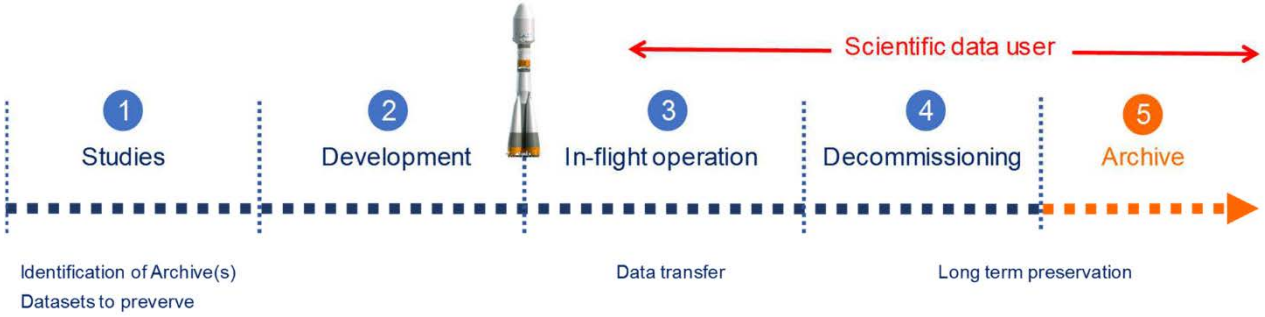


Figure 6: the preservation planning for a space project

Conclusion

Through an inventory of space data produced by space missions, the CNES SERAD service maintain the REFLECS web server to allow discovery of CNES data collections. SERAD is also responsible for long-term preservation of space data which are not preserved by other Data Centers. The new REGARDS system is going to replace these tools in the coming years. The aim is to build a new web portal for discovery and access to CNES data.

In addition, SERAD wants to be involved much earlier in space projects life cycle, for a better preservation planning.

Semantic Framework for Responsible Digital Preservation Policy

Vasily Bunakov
Scientific Computing Department, Science and Technology Facilities Council
Harwell, United Kingdom
Email: vasily.bunakov@stfc.ac.uk

The paper introduces a semantic framework for digital preservation policy definition and validation. Major requirements for responsible digital preservation policy are formulated. A generic semantic wrapper for policy definition is suggested, and a possible approach to the machine-assisted policy validation is indicated.

Keywords: Digital Preservation Policy, Activity Model, Semantic Web

Introduction

Many organizations that embraced advanced digital preservation have, on one hand, a defined preservation policy and on the other hand, software platforms for digital preservation that presumably implement this policy. Conceptual gaps between the policy formulation and its actual implementation, as well as between policy implementation and its validation can be substantial though, and this is not unique to the domain of digital preservation but pertains to any data policy.

This work in progress extends the policy modelling effort by EUDAT project (Bunakov et al. 2017; Bunakov 2017) with the suggestions that specifically suit the case of digital preservation. Semantically clear modelling of preservation policy can help making it responsible in a sense that the policy can be clearly formulated, reasonably implemented in the actual operational environment and sensibly validated.

The work first formulates generic requirements for responsible digital preservation policy, then outlines a modular approach to policy modelling and indicates a possible means of the machine-assisted policy validation.

Requirements for responsible digital preservation policy

Responsible digital preservation policy should cover all stages of data curation, from content ingestion into digital archive, through monitoring digital assets in the actual operational environment, to content dissemination. Even if these stages are covered by different policy artefacts such as separate policies for data ingestion, for data management and for data dissemination, there is a need to make all parts of the policy reasonably interpretable and mutually interoperable. This brings about the following major requirements for a responsible digital preservation policy:

- Policy should be reasonably modelled and allow machine-assisted reasoning
- All data curation actions should be clearly defined and aligned with the policy
- Policy definition (model) should be reasonably shareable and reusable
- Implementation of data policy should be subject to regular audit – with proper means to facilitate audit

There are a few ways to meet these major requirements.

One, possibly most common way is setting up an elaborated policy governance framework, i.e. a set of well-defined processes that allow human agents (policy managers) to look after the policy formulation and implementation, and effectively deliver the policy-related requirements to software developers who implement data policy in a software platform (digital archive).

Another possible way is harnessing a sophisticated policy modelling language: an example of this approach is (RuleML 2018) or skewing workflow engines such as (Taverna 2018) for the purposes of policy definition and execution.

The third possible way is using certain formalism for the expression of policy elements (policy blocks) and of their interconnections, with that formalism being reasonably friendly to machines as well as to humans.

(Bunakov 2017) explains differences amongst these policies approaches for a generic data policy, and casts a vote in favour of the one that is based on policy elements definition and on their assembling in a whole policy model, as this approach seems to have the right balance between the model expressivity and the model simplicity. What is required for a specific case of a digital preservation policy is, as further explained, the specialization (typisation) of policy patterns and reasonable preservation-specific metadata for them.

Modularity as a response to policy evolution

Some data policies, including preservation policies, are developed from scratch, and some are based on the requirements inherited from past considerations, or are imposed by umbrella organizations such as governments or corporations. Also policy may change owing to technological or organizational developments. The modular approach based on Activity Model (Bunakov 2013) allows to effectively and efficiently address the challenge of data policy evolution, with the following major generic "construction blocks" of any data policy:

- **Generic Data Policy Activity**
  This block can be used for various semantic definitions of data handling, e.g. for data characterization or data transformation

- **Logical Switch Activity**
  The block for logical switches of all sorts, e.g. for choosing a certain algorithm depending on data format

- **Control Activity**
  The block for an interface with a particular software platform where policies are being executed and monitored

These generic policy blocks introduced in (Bunakov, 2017) with a suggested RDF serialization allow modelling a wide range of data policies, with the example of a policy model for file characterization presented in Figure 1.

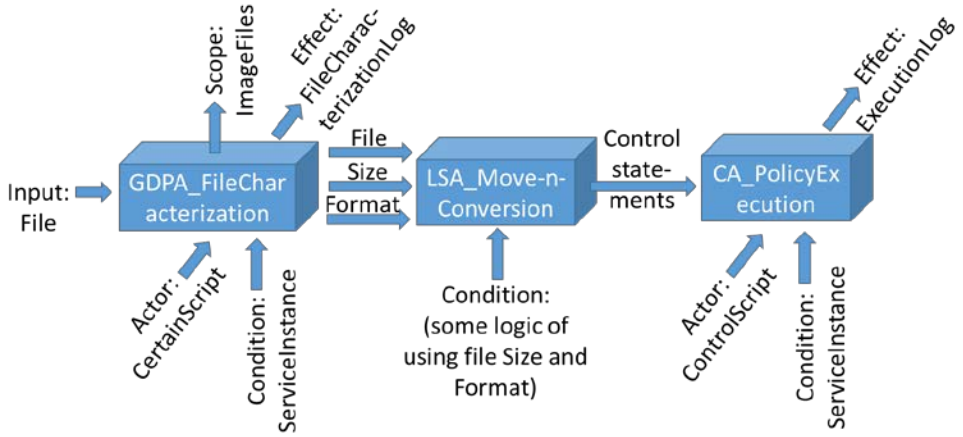


Figure 1: Example of a policy model for file characterization, consisting of three sequential actions. GDPA_FileCharacterization is an instance of General Data Policy Activity, LSA_Move-n-Conversion is an instance of Logical Switch Activity (that can, as an example, migrate formats or move files), CA_PolicyExecution is an instance of Control Activity (that uses an instance of a particular IT service). All three policy blocks, as well as their inputs and outputs allow RDF serialization as per (Bunakov, 2017).

Policy blocks like those presented in Figure 1 may not require further specialization for the case of long-term digital preservation (which, if the model were serialized in RDF, would require RDFS subclasses for each block), but the chain / network of blocks does require specialization to better suit the digital preservation case. The reason for this is that similar chains of actions can be used for different purposes in digital preservation context; as an example, file characterization may happen at the stage of data ingestion in the archive, but it may be required later on, too, e.g. for a semantically clear data checks after migration

into a new file format. So having ability to define types for the chains / networks of actions will be good for their multiple reuses in the archive.

Also, the chains / networks of policy blocks might have specific metadata assigned. One example of such metadata would be having a “purpose” attribute – to indicate what purpose the particular chain / network of actions serves, like “data characterization for ingest” or “data characterization for migration checks”.

Other useful metadata attribute would be a “policy clause”, in order to better relate Control Activity block to specific policy statements (in a granular textual form). Policy clauses can be referred to in Scope aspect of the Logical Switch Activity; this aspect is not shown for the instance of Logical Switch Activity in Figure 1, but can be introduced similarly to Scope aspect as illustrated for GDPA_FileCharacterization activity, as Scope aspect can be defined for any Activity irrespective of its nature (Bunakov 2013).

Yet another metadata particularly beneficial for the case of long-term digital preservation policy will be key-value pairs for significant properties (Wilson 2007; Knight & Pennock 2009; Giaretta 2011) that should be preserved through all content migrations.

One framework for policy formulation and policy validation

The outlined approach can allow sensible formulation of digital preservation policy, but it is also important to ensure the archive compliance to a certain data policy, so that policy checks can be traced back to the policy definition.

If policy is formulated as a textual document, then implemented in a software platform, the evidence will be required for good IT governance in order to support the claim that the policy has been actually implemented. When certain information artefacts are encountered, say in system logs, they can be reasonable indicators of the platform behaviour, yet a well-structured discussion between archive auditors and archive owners should happen every time when the archive compliance to a certain policy is in question. This way of policy compliance checks is relatively easy on the policy “formulation” end but is harder to implement on the “evidence” end.

If the policy is implemented as formal machine-executable rules or workflows, the discussion about evidence of the policy implementation is less required, apart from an additional assurance that the rules engine or workflows engine have not been tampered with. However, formulation (or any change) of data policy in a highly formal notation may present a challenge of its own; also the acceptance of a highly formal machine-executable policy definition as a “reference” policy representation may be challenging for some organizations. This way of policy compliance checks may be easier on the “evidence” end but certainly heavier on the policy “formulation” end.

The modular approach to policy modelling that we promote is only moderately challenging on the “formulation” and on the “evidence” ends. The machine-interpretable formalism that can be employed, such as RDF, will serve as just a semantic “wrapper” for policy statements (in a granular textual form) and for software components that execute policy blocks.

For the suggested modular approach, the same modelling means can be potentially used for the definition of policy patterns and for the patterns validation against policy execution results. This can be achieved by using the existing frameworks for RDF validation, such as Shape Expressions (ShEx) or Shapes Constraint Language (SHACL), see in (Gayo et al. 2017) or on the Web. Unlike other aforementioned approaches, the modular activity-based policy representation can conceptually rely on the same means of modelling: on policy definitions as RDF patterns and on validation against these patterns of the actual RDF graph resulted from policy execution.

Conclusion

This work argues that there should be both flexibility and clearness about how the preservation policy is being defined, applied, and validated. A modular approach to preservation policy modelling allows flexible, reusable and semantically clear policy definitions. The same Semantic Web frameworks can be potentially used for the formulation of preservation policy and for its validation.

References

Bunakov, V (2017). Data policy as activity network. In XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017). Open Access version: http://purl.org/net/epubs/work/35227831

Bunakov, V et al. (2017). Data curation policies for EUDAT collaborative data infrastructure. In XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017). Open Access version: http://purl.org/net/epubs/work/35227914

Bunakov, V (2013). Core semantic model for generic research activity. In XV All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" (RCDL 2013). CEUR Workshop Proceedings (ISSN 1613-0073) 1108 (2013): 79-84. Open Access version: http://purl.org/net/epubs/work/10938342

Gayo, J E L et al (2017). Validating RDF Data (Synthesis Lectures on the Semantic Web: Theory and Technology). Morgan & Claypool Publishers.

Giaretta, D (2011). Advanced Digital Preservation. Berlin Heidelberg: Springer.

Knight, G and **Pennock, M** (2009). Data Without Meaning: Establishing the Significant Properties of Digital Research. International Journal of Digital Curation, issue 1, Volume 4.

RuleML Wiki pages (2018). http://wiki.ruleml.org/index.php/RuleML_Home (accessed in April 2018)

Taverna workflow system (2018). https://taverna.incubator.apache.org/ (accessed in April 2018)

Wilson, A (2007). Significant Properties report. InSPECT Work Package 2.2, Version 2. April 2007.

Database Archiving and Big Data Techniques from the E-ARK project

Janet Anderson[1], Miguel Ferreira[2], Richard Healey [3], Zoltán Lux[4], and Sven Schlarb[5]

[1]University of Brighton, Brighton, United Kingdom

[2]KEEP Solutions LDA, Braga, Portugal

[3]University of Portsmouth, Portsmouth, United Kingdom

[4]National Archives of Hungary, Budapest, Hungary

[5]AIT Austrian Institute of Technology GmbH, Vienna, Austria

In recent years, there has been a fundamental change in the notions surrounding what constitutes archiving. E-government legislation across Europe and beyond has brought about a situation whereby archives are obliged to accept, store, and provide access to digital data on an ongoing basis. However, relatively few memory organisations have the sophisticated digital archiving infrastructure required to handle all aspects of these activities.

Addressing these issues formed part of the grand challenge posed by the EC, answered by the E-ARK consortium which included five national archives. E-ARK drew up metadata specifications for the preparation, ingest, and transfer of digital content and developed a rich set of software tools supporting the day-to-day operations in digital archives.

In this paper, the relevance of Big Data for Archives will be outlined against the background of exemplary use cases in the areas of data warehousing, full-text search, natural language processing, and visualisation.

Introduction

In recent years, there has been a fundamental change in the notions surrounding what consti- tutes archiving. E-government legislation across Europe and beyond has brought about a situation whereby archives are obliged to accept, store, and provide access to digital data on an ongoing basis. However, relatively few memory organisations have the sophisticated digital archiving infrastructure required to handle all aspects of these activities. In particular, there are implications brought about by the scale of operations involved. The vast quantities of data of ever increasing complexity are potentially overwhelming [Hey et al., 2009]. The rapid influx of material poses real challenges for archivists and administrators managing the process, as well as for the businesses, researchers, and citizens who use them. Big Data is often hailed as a solution, but the underlying mechanics of Big Data are generally not well understood by end users.

Addressing these issues formed part of the grand challenge posed by the EC, answered by the E-ARK consortium which included five national archives.[1] E-ARK drew up metadata specifications for the preparation, ingest, and transfer of digital content into archives, as well as for maintenance and digital preservation of the content to guarantee continued access to this material.

Although geared towards national archives, the E-ARK methods, tools and infrastructure are of real interest regarding higher education, as well as scientific and research data centres.

The paper is structured as follows: section 2 describes E-ARK's principles of database archiving and discovery which represent the basis for employing advanced data mining techniques in the Big Data context of an archive. Sections 3, 4, and 5 give details about several use cases of the E-ARK project with reference to Big Data analysis. Finally, section 6 provides the conclusions of the paper.

[1]http://www.eark-project.eu

1Database Archiving and Recovery

The E-ARK metadata specifications for submission, archival, and dissemination packages (SIP, AIP, DIP) laid the foundations for preparation, ingest, and transfer of information packages covering the whole life-cycle of information stored in the repository.

According to E-ARK's layered data model [Kärberg et al., 2016, p. 9] the Common IP Specifica- tion

for information packages forms the outermost layer. The general SIP, AIP and DIP specifications add, respectively, submission, archiving and dissemination information to the Common IP specifi- cation. And the third layer of the model represents specific content type specifications. One of the specific content types are databases for which the SIARD 2.0 specification [Bruggisser et al., 2016] defines the relevant standard.

Alongside these standards E-ARK implemented prototypical archival workflows for database archiving and recovery. Basically, databases were extracted and converted to SIARD using the Database Preservation Toolkit[2] (DBPTK), archived as AIPs, and again recovered into one of the supported Relational Database Management Systems (RDBMS) by using the DBPTK.

The ability to recover database records including contained files (Binary Large Objects, BLOBs) into an RDBMS is a necessary requirement to make data available for content aware information retrieval and data analysis. In the following sections the use cases investigated in the E-ARK project will be described.

Data warehousing
The US historical census use case

One of the use cases in E-ARK related to the creation of a data warehouse (DW) based on archival resources, was a case study about historical census data and their linkage to other datasets from archival sources. The dataset used was derived from one of the largest and most important datasets provided by the Minnesota Population Center (MPC) which contains more than 50 million individual person level records from the US 1880 census, each with approximately 90 variables attached. Some 5.27 million of these records were used as the test dataset.[3]

At first sight, this would appear to be a great boon to intending DW analysts, because the full dataset is very large (nearly 5 billion data items) and they are already checked and coded, ready for loading. Experience showed, however, that substantial work needed to be done to address issues, such as incomplete coding classifications, anachronistic characterisations of geography, and substantially inaccurate imputation of industrial sectors from occupational data.

The case study showed that it is entirely feasible to convert large digital datasets into a DW structure, though significant resources may need to be devoted to design and implementation of the most appropriate dimensional structures. The DW/OLAP infrastructure as described in [Thirifays et al., 2016, p. 36] provides highly effective storage and analytical capabilities for long-term database archiving, but this must be coupled with domain expertise in the archival team, to ensure that every opportunity for adding value to the overall data corpus is taken.

3.2 The Hungarian Prosecution Offices use case

The first relational database which was submitted to the National Archives of Hungary (NAH) and which needed to be archived as a record was the database of the Registration and Case Management System of the Hungarian Prosecution Service from the years 1993-1994.

For displaying and browsing individual tables, it was possible to either use the SIARD Suite[4] or the DBPTK developed by the E-ARK project. However, these software tools did not provide an integrated view of the database nor did they show how tables actually relate to each other. In contrast, a DW can provide an environment in which an archived database can be presented to users, enabling them to run new queries which have never been thought of when the database was still in use. During the E-ARK project, an application based on Oracle Application Express (APEX) was therefore developed to provide a

presentation layer for the users of the archive.

[2] http://www.database-preservation.com [3] https://pop.umn.edu

[4] https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html

As the preparation and pre-processing of data required significant effort, it is important for an archival information system to permit reusing a prepared Dissemination Information Package (DIP) in case other users request access to the same archival data (the underlying Archival Information Packages (AIPs)). In this way, the DW can be considered a means to deliver a database as a DIP which offers additional functionality and added-value for end users by supporting advanced use cases for OLAP or Data Mining.

Information Retrieval and Natural Language Processing

Large-scale full-text indexing

The large-scale full-text indexing in the E-ARK project enabled faceted querying of document col- lections in addition to the rather limited search mechanisms of database indexing [Schmidt and Karl, 2016]. The full-text search functionality was provided through an Apache Solr search server, which relies on a previously generated full-text index created by Apache Lucene.

Depending on the volume of content, indexing as well as ingestion can become very resource and time consuming processes. Both processes have therefore been implemented as parallel applications that can take advantage of a computer cluster to scale out for large data sets.

The generated index is made available by the search server as a query interface enabling a client to formulate and execute queries against the index, compose complex queries based on facets, and rank them based on different characteristics.

Named Entity Recognition

One of the specific Natural Language Processing (NLP) tasks was Named Entity Recognition using archived textual data sources. The goal was to improve and add value to the Solr[5] index created from the ingested data, as well as show possible options in terms of access methods.

For carrying out NER, the Stanford CRF-NER[6] parser was used in combination with the NLTK[7] (Natural Language Toolkit) Python library.

Topic Modeling

An experiment on topic modeling has been carried out using the Python machine learning library scikit-learn [Pedregosa et al., 2011]. Using a pre-trained model based on newspaper categories (“Sports”, “Culture”, “Politics”, etc.), abstract categories of documents were identified and added to the Solr index. The idea of this approach is to enable users to quickly identify and select or exclude documents, and to narrow down search queries on large datasets. However, an important pre-requisite for topic modelling is the creation of a training model which is capable of identifying relevant categories in the corresponding data collection.

Visualization of geographic data

The change of Slovenian borders over time

The visualization of extracted geographical data was undertaken as part of the DIP creation for information packages containing geographic data. In a first step, the data was normalised into XML files of the Geography Markup Language (GML) Encoding Standard.[8] Examples of these files are GML representations of the regions of Slovenia in a time series of the years 1994, 1995, 1998, 2002, 2006, 2010, and 2015.[9]

The Peripleo[10] tool from the Pelagios project[11] was chosen as the environment for visualizing the geographical data because it is especially designed to visualize, search, and discover sets of

[5] http://lucene.apache.org/solr [6] http://nlp.stanford.edu/software/CRF-NER.shtml
[7] http://www.nltk.org

[8] GML XML Schema files: http://schemas.opengis.net/gml [9] https://github.com/eark-project/earkweb/tree/master/earkresources/geodata [10] https://github.com/eark-project/peripleo

[11] http://commons.pelagios.org

geographical data which are created over time.

Peripleo used an RDF based format where the geographical properties are available as one feature amongst other properties of a gazetteer entity, such as name, and year of a region or point on a map.

The location of recognised named entities

As an extension of the NER use case described in section 4.2, the recognized location entities were enriched with geographic information, i.e. the geographic coordinates of the corresponding locations. The coordinates of locations were retrieved using the publicly available Nominatim[12] database.

By combining the NER results with the additional geographic information, novel ways of brows- ing document collections by visualising location entities and their corresponding frequency were investigated. As in the use case described in section 5.1, Peripleo was used to display the results on a map. By that way it was possible to visualize the geographic focus of documents, and to access the corresponding documents within the same interface.

In addition, using metadata associated with the documents, available date-time information was taken into account. Using this information it was possible to get insights about the distribution and frequency of geo-locations in a document collection over time.

Conclusion

The number of digital publications, governmental records, or digitized materials that memory in- stitutions need to cope with is growing at a rapid pace. While this holds great potential for big data analysts, it brings about a whole new set of requirements for memory institutions regarding the scalability, interoperability, and security of the services they offer.

Without doubt, this represents nowadays a big challenge for the development of information systems in archival institutions. In this paper, selected use cases of the E-ARK project have been described to demonstrate the relevance of Big Data in digital archives.

On the one hand, it is clear that databases need specific methods to allow storing and preserving information originating from databases. The SIARD standard represents a cornerstone in this sense. On the other hand, as explained on the basis of the above use cases, ensuring access requires the – potentially labour-intensive – preparation of dissemination information packages to support advanced content analysis using OLAP and data mining techniques.

References

Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

Tarvo Kärberg, Angela Dappert, Andrew Wilson, Levente Szilágyi, Jože Skofljanec, João Cardoso, and Gregor Zavrsnik. Smurf (semantically marked up record format) specification for erms v1.1. Technical report, DLM Archival Standards Board, 2016.

Hedi Bruggisser, Georg Büchler, Alain Dubois, Martin Kaiser, Lambert Kansy, Markus Lischer, Claire Rothlisberger-Jourdan, Hartwig Thomas, and Andreas Vos. ech-0165 siard format specifi- cation 2.0. Technical report, Verein eCH, 2016.

Alex Thirifays, Zoltan Lux, Jože Škofljanec, Gregor Završnik, Anja Paulič, Anders Bo Nielsen, Phillip Tømmerholt, Janet Anderson, Richard Healey, Kuldar Aas, Andrew Wilson, Jan Aspenfjall, and David Anderson. D5.3 e-ark dissemination information package (dip) final specification. Technical report, E-ARK Project, 2016.

Rainer Schmidt and Roman Karl. D6.1 faceted query interface and api 1.0. Technical report, E-ARK Project, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pretten- hofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Re- search*, 12:2825–2830, 2011.

[12]http://wiki.openstreetmap.org/wiki/Nominatim

Sharing Earth Observation Data on the Web

Uwe Voges[1], Michael Schick[2] and Udo Einspanier[1]

[1]con terra GmbH (Münster, Germany)
u.(voges|einspanier)@conterra.de
[2]EUMETSAT (Darmstadt, Germany)
michael.schick@eumetsat.int

A significant obstacle for easy finding and accessing Earth Observation (EO) data on the web is its provision by specific EO Portals usually offering proprietary or domain specific software interfaces. For mainstream search engines and standard web clients it is hard (even impossible) to get access to it. This paper presents approaches based on linked (open) data and spatial data on the web best practices to improve sharing and finding EO data on the web. Especially the new Open Geospatial Consortium (OGC) OpenSearch-EO extensions and simple REST-API´s are in the focus.

1. Introduction

Currently we are facing a massive increase of EO data by newly emerging satellite programs like EU/ESA Copernicus (Sentinel-1, -2, -3, ...) and significantly improved processing capabilities leading to many new derived EO datasets. In contrast to DataCubes (which can be subset upon request) the datasets are often logically and/or physically managed as:

- EO products: single datasets representing instances of a product type (e.g. a sensor) covering a specific area and time, provided in different formats (including a service based access, e.g. Open Geospatial Consortium (OGC) Web Map Services (WMS))
- EO collections: sets of EO products sharing a common specification

Although the EO products are frequently published as "Open Data", it is hard to find and get access to them. Reasons are e.g. their invisibility to mainstream search engines because of their storage in archives of specific EO Portals providing specific interfaces. EO collection and products (meta)data (e.g. ISO19115) is search- and accessible by domain specific services (e.g. OGC Catalogue Services Web (CSW)). These metadata formats and services are the foundation of classical Spatial Data Infrastructures (SDI´s) (see lower (blue) part of Figure 1). But the access to the (meta)data with standard web technologies (e.g. REST-services providing the (meta)data in well-known formats which can be processed by standard Web/Java-Script clients) is not ensured. Further the metadata cannot be linked with other data on the Web. The latter prevents e.g. simple browsing to the EO products just following links.

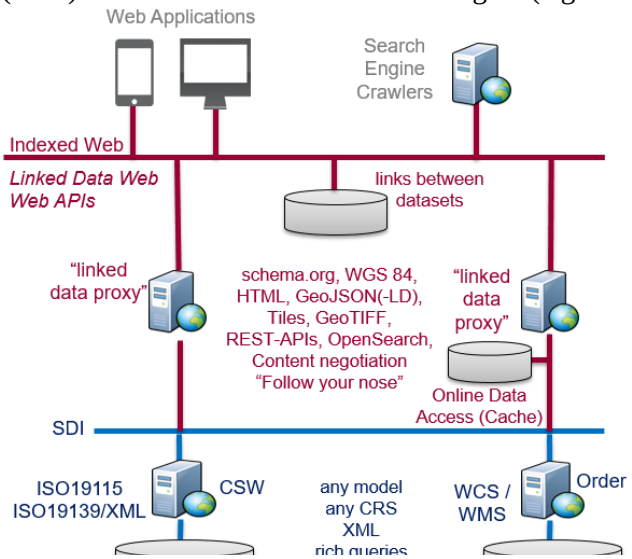


Figure 1: EO Data on the web (here sitting on top of a classical SDI) (modified from Portele et al., 2016)

This situation can be improved using technologies and best practices group-ed under the terms linked (open) data (e.g. Berners-Lee, 2006) and spatial data on the web (OGC / W3C, 2017) implementing (e.g. by a layer sitting on top of an existing SDI, see Figure 1):

- EO (meta)data as HTML, usable e.g. by mainstream search engines for indexing and
- Atom/XML or GeoJSON(-LD) (see below) using standardized vocabularies to better support search and discovery by machine clients and for linking EO metadata with other web resources (enabling navigation, deep queries etc.)

- Tile based data access supporting standard formats like GeoTIFF (e.g. provided by an Online Data Store[24])
- Unique and persistent HTTP URLs for products and collections (with content negotiation)
- Simple web standards based (REST) API´s for search, navigation and data access
- HATEOAS (Hypertext As The Engine Of Application State) where the application status is defined by resource representation,materialized as links giving the options of the client.

2. Providing EO Data for being indexed by Mainstream Web Search Engines

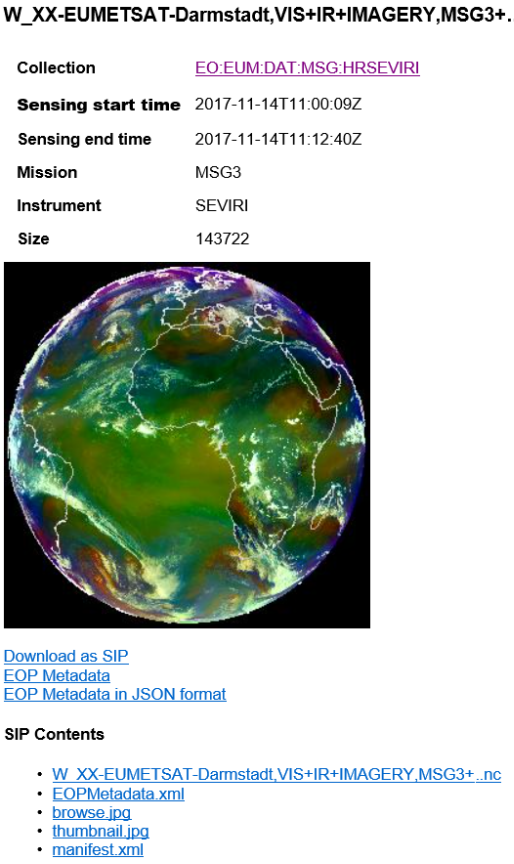


Figure 2: HTML representation of EO product metadata

A typical procedure of getting to the EO products is to search with mainstream search engines like Google. To be indexed here, the EO products must be provided in HTML format offering metadata (e.g. collection, title, spatial extent, previews, links to services for accessing the EO product, see Figure 2). For an improved search experience, it should ideally be annotated with schema.org elements. To enable crawlers to find and index these HTML pages either a sitemap or a landing page (providing the EO collections pointing to the EO products) must be registered with Google & Co.

But this indexing becomes a problem in case of millions of very similar products (permanently increasing in number) as it is not possible to force the search engines to index all existing and permanently incoming new products. Another problem with mainstream search engines is that they cannot support more dedicated searches considering specific search parameters (e.g. spatio/temporal extents, platform/instrument, etc., defined by vocabularies) to constrain the resultset. Further they do not support appropriate machine-readable response formats. Both are requirements for processing environments.

3. Finding EO Data on the Web using OpenSearch(-EO)

OpenSearch (2005) is a standard´s based and easy to use web search engine. It is aligned with the technologies and best practices listed above. It provides a simple to use search interface description, called OpenSearch Description document (OSDD). The OSDD includes one URL template per supported response format including key-value-pair parameters to constrain the search (see Figure 3).

```
<Url type="application/atom+xml" template=" http://eoportal.eumetsat.int/eopos?
pI={eop:parentIdentifier}&g={geo:geometry?}&r={geo:relation?}&dtstart={time:start?}
&dtend={time:end?}&iQD={eop:productQualityStatus?}&st={startIndex?}"/>

http://eoportal.eumetsat.int/eopos?pI=acronym:MSG15:satellite:MSG2:fileid:EO:EUM:DA
T:MSG:HRSEVIRI&dtstart=2017-09-09&dtend=2017-09-10
```

Figure 3: Example OSDD URL Template and a corresponding valid request

A parameter extension allows to describe the type and value range of the parameter values. A client (e.g. browser) can use the OSDD to automatically create a search frontend. The resultset includes the search

[24] Note: the online access to the data is not in the primary focus of this paper (as search and discovery is).

entries, information about the current search and means for pagination. The entries represent metadata about the original data item. Usually they include links to external information, e.g. to all metadata details or to access endpoints. The links are typed e.g. using the media types standardized by IANA. The results can be returned in different encodings e.g. Atom, RDF, JSON or HTML and be aggregated by the clients (e.g. browser).

OpenSearch can easily be extended to support new search parameters (see examples with namespaces prefixes in Figure 3) and additional response formats. The OSDD informs clients about the supported extensions. The currently developed OGC specifications OpenSearch-EO (OGC 2018a, 13-026r9), OpenSearch-EO GeoJSON(-LD) Response Encoding (OGC 2018b, 17-047) and EO Dataset Metadata GeoJSON(-LD) Encoding (OGC 2018c, 17-003) standardize extensions to search for EO collections and EO products. They define EO specific query parameters and Atom/XML- / GeoJSON(-LD)-based response formats (providing EO specific details including the support for 2-step-searches where collection metadata points to an OSDD allowing to search for related EO products). The elements are partly defined in RDF Schema using well-known vocabularies (DC, OGC, Atom, IANA).

Figure 4: Snippet from an OpenSearch GeoJSON-LD search response

In contrast to Atom/XML GeoJSON (IETF, 2016) is more suitable for web clients. It encodes simple geographic features including their non-spatial properties using JSON. But the properties are just bare local names, not based on a vocabulary (no URIs assigned) and so not well suited for linked data. Better suited is JSON-LD (W3C, 2014). It is a RDF serialization format and is part of W3C´s Semantic Web Activities. An important concept is the *@context* which defines the properties. Here bare property names are mapped to URI´s with the *@id* keyword. The type can be indicated with the *@type* keyword. The GeoJSON-LD encoding proposed in OGC 17-047 and 17-003 supports a seamless transition from GeoJSON with a minimum number of edits. GeoJSON can be interpreted by JSON-LD clients simply using a normative @context (see Figure 4). But it does not require JSON-LD processing.

4. Simple REST API´s for Browsing and Accessing EO Products

OpenSearch is well suited for arbitrary searches along varying parameters (axes). In addition, a simple to use REST-API for discovery and access is of advantage for humans using a web browser but also for improving the visibility in mainstream search engines (when implementing HTML with HATEOAS). At its basic path the API should provide the available EO collections accompanied with some metadata. From here it must allow to browse (navigate) through the EO products based on a fixed set of axes (e.g. temporal and spatial). During the navigation, the set of remaining EO products is narrowed step by step by drilling into hierarchically organized, named value ranges of the axes. The results of an intermediate step are not hundreds of EO products but either the next finer grained value ranges (e.g. 2018 -> 01-2018 -> 08-01-

2018) or another axis which could be followed (see Figure 5). Just on lower levels a link to the remaining products is provided. Upon following this link, the list of product-links is presented. By following such a link some metadata (e.g. title, sensing time etc), links to other representations and links for accessing the data (e.g. download) is shown (see again Figure 2).

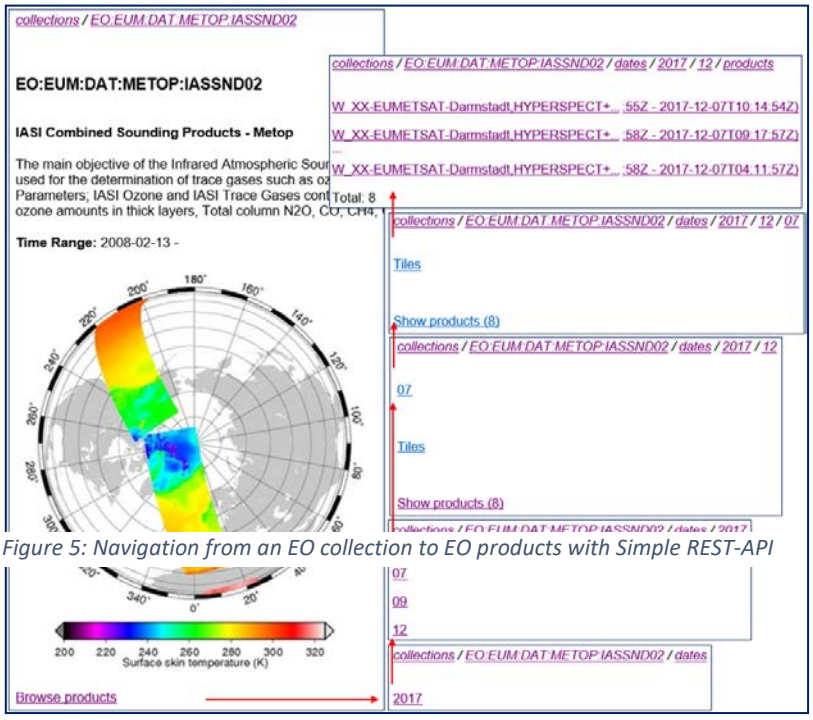


Figure 5: Navigation from an EO collection to EO products with Simple REST-API

In addition to HTML, the results should be returned in a machine-readable format like GeoJSON(-LD). This could be beneficial for use cases of mass processing's where simple iterations (no searches) through "virtual directories" providing huge amounts of EO products are in the focus. This API can then be used within processing environments, Jupyter Notebook etc.

5. Conclusion

The problem of hard to reach (from the web) EO data silos can be overcome using best practices summarized under the terms "spatial data on the web" and "linked open data". Especially the usage of persistent HTTP URL´s, providing data representations in mainstream web formats (e.g. HTML, GeoJSON(-LD)), offering a standard search interface as OpenSearch-EO and providing easy to use REST-based browse API´s (supporting HATEOAS) helps offering, finding and accessing EO data on the web. con terra is currently implementing the described best practices in different projects for EUMETSAT (Figures 2-5).

6. References

- OpenSearch 2005 1.1 Draft 5, http://www.opensearch.org/Specifications/ OpenSearch/1.1
- Berners-Lee, T 2006 Linked Data. https://www.w3.org/ DesignIssues/LinkedData.html.
- W3C World Wide Web Consortium 2014 JSON-LD 1.0, A JSON-based Serialisation for Linked Data, W3C Recommendation, http://www.w3.org/TR/json-ld/
- Internet Engineering Task Force 2016 The GeoJSON Format. RFC 7946. https://tools.ietf.org/html/rfc7946
- Portele C, van Genuchten P, Verhelst L, Zahnen A. 2016 Spatial Data on the Web using the current SDI. https://geo4web-testbed.github.io/topic4/.
- OGC Open Geospatial Consortium & World Wide Web Consortium 2017 Spatial Data on the Web Best Practices. W3C Working Group Note 2017. https://www.w3.org/TR/sdw-bp/
- OGC Open Geospatial Consortium 2018a OpenSearch Extension for Earth Observation (OpenSearch-EO), Doc 13-026r9 (in approval process, final publication expected 2018).
- OGC Open Geospatial Consortium 2018b OpenSearch-EO GeoJSON(-LD) Response Encoding Standard, Doc 17-047 (in approval process, final publication expected 2018).
- OGC Open Geospatial Consortium 2018c EO Dataset Metadata GeoJSON(-LD) Encoding Standard, Doc 17-003) (in approval process, final publication expected 2018).

Sustainable management of agricultural research data: A case for Big Data platform development for climate Smart Agriculture in Kenya

Boniface Akuku, Email: boakuku@gmail.com, Kenya Agricultural Research and Livestock Organization and Professor Robert Oboko, University of Nairobi, School of Computing and Informatics.

The role of research data in agriculture has rapidly gained acceptance among different stakeholders in Kenya. Similarly, a number of platforms for data products and services have been developed for a variety of value chain actors in agricultural in Kenya. This spectacular growth is partly as a result of improved access to agricultural research information and knowledge through the use of Information and Communication Technologies (ICTs) and application of social innovations. However, most of the current platforms supporting management of agricultural research datasets are financially too unsustainable to cope with growing demand for intelligence data products and services. The growing demand has been catalyzed by the aggregate need for real time predictive data analytics as well as discovery of new knowledge from datasets. Specifically, in agricultural sector the problem is exacerbated by the endlessly emerging socio-economic challenges facing the sector such as chronic droughts, hunger, acute water shortage, pestilences and loss of livelihoods. In Kenya, the frequency and intensity of severe weather events has increased and is projected to further worsen due to climate change effects. These extreme weather events' occurrences have caused devastating effects on the agricultural sector's performance despite the performance having a high correlation with Gross Domestic Product. Studies show that these weather events have resulted into numerous problems notably crop losses, death of livestock, high food prices, high poverty levels, food insecurity and malnutrition. This paper underscores that management of agricultural research datasets using evidence based on data solutions is critical for climate smart agricultural (CSA) adaptation and practices in Kenya. The findings show that Big Data solutions provide sufficient prospects to determine previously unconceivable insights from datasets. The technology and ability of data management platform to intelligently extract valuable information from a myriad of datasets, and proactively respond based on the created knowledge from data are important ingredients for consideration for sustainable management of CSA programs. Further, demands for insight and foresight (discovering unknown) from data through automation is growing as a premise to unlock persistent complexities and reactive approaches which has been applied by agricultural stakeholders in Kenya for decades. Theoretical and empirical evidence indicate that these old ways will not cope with emerging challenges facing Kenyan agricultural sector, thus creating the need for mining of intelligence from agricultural data sets. Therefore, appropriate data management solutions such as Big Data platforms must be put in place to innovatively optimize agricultural research data for societal benefits.

Introduction

The role of research data in agriculture has rapidly gained acceptance among different stakeholders in Kenya and this has led to development of different platforms that provide data products and services to different agricultural value chain actors. This spectacular growth is partly as a result of improved access to agricultural research information and knowledge through the use of Information and Communication Technologies (ICTs) and application of social innovations (Celli *et al.*, 2015). Similarly, in recent years, the importance of employing good data management and facilitating data sharing has become increasingly recognized and emphasized in agricultural research domain (Kirub, 2016, p.1-4). Several ICT platforms have been developed to support management and sharing of agricultural data as well as provision of information products and services (Zyl et al., 2014; Oboko *et al.*, 2016). However, most of the current platforms supporting management of agricultural research datasets are financially too unsustainable to cope with growing demand for intelligence data products and services.

A study conducted by the world bank further shows that in Kenya and other developing countries, most of these platforms are initiated and financed from unsustainable and fixed term donor supported projects (Tinsley & Agapitova, 2018). Furthermore, the aggregate need for real time predictive data analytics as well as discovery of new knowledge from agricultural datasets has been identified as a critical factor in climate smart agriculture (CSA) practices (Ekekwe, 2017; Ndemo, 2017a). This demand for intelligence and insights from agricultural research datasets imply that there is need for platforms, methodologies and tools that support deep analytics and learning as core capabilities. Additionally, the rapid growth of data in agricultural sector is fast exceeding the ability of standard Database Management Systems (DMS) and methodologies.

With regard to national development, agriculture remains one of the most important sectors of the economy in Kenya. For instance, the five years' period ending 2012, the sector's annual contribution to Gross Domestic Product (GDP) averaged 27.3% (World Bank report, 2014). It is an important driver of the national economy contributing 26% of the GDP directly and another 25% indirectly annually. In addition, it accounts for 65% of Kenya's total exports and about 75% of industrial raw materials and provides a means of livelihoods to millions of people. Furthermore, it has a high employment multiplier effect with about 18% of Kenya's formal employment and more than 70% of informal employment in the rural areas working in the sector (GoK, 2013). Despite this critical role, daunting challenges remain and the scale of poverty among the communities is staggering, with a majority rural population living in extreme poverty (Kim & O'Brien, 2015). However, an overriding challenge for the agricultural sector in Kenya is increased frequency and intensity of severe weather events and is projected to further worsen due to climate change effects (Ongoma & Shilenje, 2016). These extreme weather events' occurrences have caused devastating effects on the agricultural sector's performance. A study conducted by the World Bank in 2014 shows that these weather events have resulted into numerous problems notably crop losses, death of livestock, high food prices, high poverty levels, food insecurity and malnutrition. The fact that agricultural performance has a high correlation with GDP has resulted in a number of investments to develop platforms that can provide solutions to these problems using agricultural research datasets, mainly to improve productivity and reduce poverty levels. These platforms have attempted to design and support management of agricultural research datasets and are specifically directed towards agriculture, particularly Climate Smart Agriculture (CSA) domain. However, little success and impact has been recorded from the existing platforms, one main challenge in managing agricultural research data is ensuring appropriate use of agro-ecological identifiers (Kirub, 2016, p.4)

Studies show that big data platforms have been useful in designing appropriate techniques for handling data that cannot be handled by the conventional DMS and tools (Chen & Lin, 2014; Protopopa & Shanoyan, 2016; Rao, 2017). The problem is exacerbated by the unending socio-economic challenges facing the agricultural sector such as acute water shortage and hunger, pestilences and loss of livelihood mainly as a result of climate change effects. In this regard, efficient techniques or algorithms to analyze agricultural research data provides useful information for decision making in CSA. Besides, careful mining of these datasets reveals required indicators of socio-economic and climate change events, which can support establishment of effective policies for sustainable management of CSA programme. Difficulties in availability of ICT empowering techniques and tools that have ability to learn and adapt autonomously in obtaining and analyzing data such as big data platforms that can guide policy formulation has been cited as one of the main limiting factors (Shirsath *et al.*, 2017). Furthermore, existing data management and sharing platforms supporting CSA are designed and supported through unsustainable macro-initiatives with majority not able to scale-up beyond the project life. Moreover, not much success has been recorded from these platforms despite some having adopted business models such as social enterprise ideals (Tinsley & Agapitova, 2018). Similarly, there are few studies on prioritization of big data for CSA (Shirsath *et al.*, 2017) as well as what has been done in CSA context using big data platform. This paper underscores that management of agricultural research datasets using evidence based on data is critical for climate smart agricultural (CSA) adaptation and practices in Kenya

Methods

Research methods remains a prerequisite for undertaking a study for data collection, analysis and establishing relationships between variables. A cross-sectional study using multiple approaches was employed to investigate demand for intelligent data products and services as well as knowledge discovery from data among different stakeholders. These included a preliminary baseline study using a questionnaire. Cohort sampling was deliberately chosen to unequivocally evaluate the intervention logic of the different platforms for managing agricultural research data as a matrix and use it (i.e. the matric) as a basis to find out the following:

1. "What are the available platforms employed in agricultural research data management?"
2. "How sustainable are these platforms to support emerging demands for intelligent data products and services as well as discovery of knowledge from data?"

3. "What are the technological options to provide data driven evidence that can impact on agricultural research particularly in the CSA context?"

In evaluating these questions and indicators along the theory of change premises as regards underlying connections between the purpose of these data management platforms and their ultimate goals a selected literature materials were also extensively studied. Subsequently, platforms currently being used for managing agricultural data to provide information products and services were comprehensively reviewed. In this regard, big data platform was identified as a suitable intervention logic (see detailed summary in results section). Table 1 below is a list of reviewed platforms.

Table 1: Data Management platforms supporting agricultural information products and services

Platform	Description	Database	Findings
Esoko	The platform integrates smallholder farmers into a recognized value chain through mobile phone technology	Standard	The platform performs the standard data collection, analysis and output function but lacks ability to generate intelligence and discovery of new knowledge from data
KenCall – M-Kilimo	The platform is a bridge between farmers and agriculture experts through a call management solution.	Standard	The platform provides the standard database query functions with no additional capabilities
NAFIS	The platform is for information storage and allows access via internet enabled devices	Standard	The platform has capacity to holds large amounts of data but does not possess data processing, data mining and analysis capabilities
UjuziKilimo	The platform uses big data and analytic capabilities to provide knowledge-based information and insights	Big Data database	The platform postulate that it has big data capabilities including analytics, predictions, knowledge discovery, as well as addressing CSA concerns. While the platform is a startup with potential to grow, it has not mentioned how data is collected, analyzed and sustained. In addition, there is no evidence of its application in actual settings and field level application. Furthermore, not much mentioned on its data management strategies and business model and approaches that can ensure sustainability given the huge investments that are associated with big data platforms as well as the high demand for intelligent data products and services that are required to support CSA interventions.
FarmDrive	The platform "uses mobile phones, data, and machine learning algorithms to bridge data gaps between financial institutions and	Big Data Database	The platform posits that it uses big data techniques in generating intelligent data products and services. While a closer study of the platform reveals novelty of concept and capabilities, however the data management framework and ecosystem is not provided. On the other hand, critical big data processes and data management protocols are neither provided nor mentioned. Moreover, the provision of intelligent data products and services borrows heavily from standard DBMS approaches and tools such web 2.0 among others. Additionally, data management sustainability

	smallholder farmers through credit worthiness assessment"		model is completely missing and there is no evidence data management strategies and impact assessment to guarantee uptake of the product and services.
M-farm	The platform provides connection intelligent ways of connecting buyers and farmers.	Big Data database	The platform claim it has ability to provide intelligence in linking farmers and the market. A study of the platform reveals a good picture of intellectual innovation and requisite big data capabilities. However, there is no evidence of data management protocols and strategies that demonstrate the platforms ability to handle big data that is data in high volume, velocity, variety and veracity. Since big data platforms are associated with high cost of investments and management, there is no evidence to provide proof for sustainability as well as business model.
DrumNet-Kenya	The platform provides information support services to smallholder farmers in Kenya	Standard	A study of some literature mentioned the platform as one of those categorized as big data platform. A closer review and study of the platform indicates the platform uses a standard DBMS and web-based technologies. There is no evidence of big data capabilities, data management sustainability and provision of intelligent product and services. Moreover, the platform is premised on a macro initiated project.

In addition, quantitative approaches were used to find answers as to why there is a need or otherwise for sustainable management of agricultural research data in the CSA context. In this study, predetermined sample size from agricultural research data users was used, thus the pre-selected respondents who were in the cohort were given questionnaires using a cohort tracking sheet with unique identifiers to ensure consistency. To ensure enough statistical power to identify viability of a big data platform, results from the feasibility study were analyzed and sufficiency test was carried out on the sample size after attrition. The choice of sample size for the study was guided by the fundamental questions the study had to answer in the data collection exercise: "What has been the situation in managing agricultural research data over time? How are data products and services organized? For example, are the desired significant impact for data products and services realized? How sustainable are the platforms? What are the abilities to extract intelligence and knowledge from agricultural research data?" Data collected during the study has been used to explain the reasons for big data platform suitability for sustainable management of agricultural research for CSA in Kenya. Table 2 and Table 3 below information products and services.

Table 2: A summary of CSA products and services

Type of Product and service products	**(%)**
Early warning	80
Climate predictions	78
Weather Forecasts	55
Agro-weather advisories	70
Government policies	45
Insurance derivations	45
Climate projections	30
Transport safety advisories	30
Airspace forecasting	21
others	14

Source: World Bank Group Field Survey Report Number 103186-Ke (2016)

The survey report shown tables 1 and 2 indicates that there is limited products and services being provided by the existing platforms. Also, it is evidenced that the platforms do not have effective mechanisms with which to track their users.

Table 3: The different types of data used for provision of CSA products and services

Type Data used in CSA product and service provision	%
Manually collected data	59
Automatically collected data	7
Data collected from remote sensing	34

Source: World Bank Group Field Survey Report Number 103186-Ke (2016)

The survey report shown tables 3 indicates that majority of the data is collected manually which is an expensive undertaking and difficult to sustain. Also very little data management ecosystem is automated. These findings re-enforce the need for a big data platform.

Table 4: The different approaches used in data processing

Approaches used in data processing	%
Expert consultation	79
Stakeholder consultation	66
Focus group discussion	55
Modeling	52
Others	17

Source: World Bank Group Field Survey Report Number 103186-Ke (2016)

Results and Discussions

The findings shown in table 4 further justify why Big Data Analytics and Intelligence (BDA&I) platforms have become increasingly important for many industries and development initiatives (Manyika *et al.*, 2011; Gokhale, 2011; Chen *et al.,* 2012). One of the main challenges is mechanisms to manage historical data, for instance time-series data, as well as continuous generation and provision of accurate and timely data.

Conclusions

Results indicate that there is no evidence to show that the existing data management platforms have sufficient capabilities to provide for the rapidly growing demand for intelligent data products and services as well as discovery of knowledge from agricultural research datasets. Also, it clear that there is adequate agricultural research data available that can be analyzed, mined and turned into insight using big data techniques and strategies. Similarly, the findings indicate that there are limited platforms that can meet the aggregate need for real time predictive data analytics from agricultural datasets. In addition, the findings shown in the above tables indicate that the existing platforms do not have deep analytics and learning abilities which is a critical limitation in addressing these endlessly emerging socio-economic challenges facing agriculture sector such climate change effects. Moreover, the demand for climate information products and services is quite high as shown in table 2 above. Of interest to note is that, these challenges are persistent and complex and as such processing data using expert consultation, stakeholder consultation or focus group discussion will not be able to effectively respond to meet the expected required levels of performance the sector is expected to achieve. Following the designed big data framework in this study, the authors assert that big data solutions provide sufficient prospects to determine previously unconceivable insights from datasets. Further, to sustainably manage agricultural research datasets for CSA practices and programs data-driven evidence is critical to achieve desired success. From theoretical and empirical evidence established in this study, there are sufficient indicators that current products, services and practices will not cope with emerging challenges facing Kenyan agricultural sector. Therefore, appropriate data management solutions through Big Data platforms must be put in place to innovatively optimize agricultural research data for societal benefits as well as appropriate technological options to provide data driven evidence leading to future research.

References

Government of Kenya. (2010). Agriculture Sector Development Strategy (ASDS), 2010 – 2020. Government Printers. Nairobi, Kenya.

Celli, F., Malapela, T., Wegner, K., Subirats, I., Kokoliou, E., & Keizer, J. (2015). AGRIS: providing access to agricultural research data exploiting open data on the web. F1000Research, 4.

Chen, X. W., & Lin, X. (2014). Big data deep learning: challenges and perspectives. IEEE access, 2, 514-525.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. MIS quarterly, 1165-1188

Ekekwe, N. (2017). How digital technology is changing farming in Africa. Harvard Business Review, 18.

Gokhale, V. (2011). The 2011 IBM Tech Trends Report: The Clouds Are Rolling In... Is Your Business Ready.

Government of Kenya. (2013). Vision 2030: Second Medium Term Plan (2013 - 2017). Transforming Kenya: *Pathways to Devolution, Socio-economic Development, Equity and National Unity.* Government Printers. Nairobi, Kenya.

Kim, H., & O'Brien, T. (2015). Tackling Inequality: Part and Parcel of Kenya's Fight Against Poverty. Georgetown Journal of International Affairs, 16(1), 16-23.

Kirub, A. (2016). Agricultural Research Data Management.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Ndemo, B. (2017a). http://www.businessdailyafrica.com/analysis/Let-Africa-start-predicting-crop-yields/539548-3912692-r3cbmfz/index.html

Rao, N. H. (2017). Big Data and Climate Smart Agriculture-Review of Current Status and Implications for Agricultural Research and Innovation in India.

Shirsath, P. B., Aggarwal, P. K., Thornton, P. K., & Dunnett, A. (2017). Prioritizing climate-smart agricultural land use options at a regional scale. Agricultural systems, 151, 174-183.

Tinsley, E., & Agapitova, N. (2018). Private Sector Solutions to Helping Smallholders Succeed.

Oboko, R., Kimani, J., Kante, M., Chepken, C., Wario, R., & Kiai, R. N. (2016). A knowledge management system for indigenous crops production: case of sorghum farming in south Tharaka, Kenya. AIMS Agriculture and Food, 1(4), 439-454.

Protopop①a, I., & Shanoyan, A. (2016). Big Data and Smallholder Farmers: Big Data Applications in the Agri-Food Supply Chain in Developing Countries.

Ongoma, V., & Shilenje, Z. W. (2016). The effectiveness of agrometeorological information in the realization of Kenya's Vision 2030; lessons learnt from China. ITALIAN JOURNAL OF AGROMETEOROLOGY-RIVISTA ITALIANA DI AGROMETEOROLOGIA, 21(1), 67-72.

Zyl, O. V., Alexander, T., Graaf, L. D., & Mukherjee, K. (2014). ICTs for agriculture in Africa.

World Bank Group Report Number 103186-Ke (2016)

COLLABORATIVE LONG-TERM DATA PRESERVATION: FROM HUNDREDS OF PB TO TENS OF EB

Jamie Shiers

CERN, 1211 Geneva 23, Switzerland

In 2012, the Study Group on Data Preservation in High Energy Physics (HEP) for Long-Term Analysis, more frequently know as DPHEP, published a Blueprint report detailing the motivation for, problems with and situation of data preservation across all of the main HEP laboratories worldwide. In September of that year, an open workshop was held in Krakow to prepare an update to the European Strategy for Particle Physics (ESPP) that was formally adopted by a special session of CERN's Council in May 2013 in Brussels and key elements from the Blueprint were input to that discussion. A new update round to the ESPP has recently been launched and it is timely to consider the progress made since 2012/2013, list the outstanding problems and possible future directions for this work.

INTRODUCTION

The current strategy for European Particle Physics, the main emphasis of which is clearly on physics, makes only a brief mention of data preservation:

*"… as well as infrastructures for data analysis, **data preservation** and distributed data-intensive computing **should be maintained and further developed**."*

We first describe how these somewhat enigmatic targets have been met in both a quantitative and qualitative fashion and then describe our goals for the up-coming revision of the strategy, including outstanding issues and concerns.

In addition, there have been a number of key developments both within and outside HEP that change the environment in which we work. These include:

1. Production services – at least at CERN – now exist for all of the main areas identified in the DPHEP Blueprint. These are described in an iPRES 2016 paper;
2. Significant additional emphasis is now placed on the need for data preservation, sharing and re-use through FAIR data management principles and so forth, both by funding agencies as well as research teams;
3. Cross-disciplinary collaboration continues to develop, following on from the Alliance for Permanent Access to Working and Interest Groups in the context of the Research Data Alliance;
4. The importance of Certification – a *sine qua non* – according to many viewpoints, continues to grow. At a January 2018 Science Europe workshop it was announced that a policy for listing only certified repositories – at least to the level of CoreTrustSeal – could be expected within 5 years (this is expected to be one of the recommendations of the FAIR Data Action Plan;
5. The volume of data to be preserved, as well as the "business cases" for so doing, have become much more quantifiable.

However, new challenges have arisen, emphasizing the fact that "data preservation is a journey – not a destination".

At the time of the Blueprint, it was not uncommon for media migration to be under the full control and responsibility of t0he experiment / project in question. In other words, "bit preservation" as a service was not the norm and this is still the case at some laboratories. Indeed, one of the inputs to the 2012/13 ESPP update dismissed bit preservation as simply "copying the file a couple of times".

Although the use of Invenio-based services, such as INSPIREHEP, or more recently Zenodo or B2SHARE, for storing documentation was a fairly common practice, this covers only a small fraction of the experiment-specific documentation and "knowledge".

Validation of software was an area of concern with some promising work on (semi-)automatic validation frameworks (versus "museum systems").

FROM A STUDY GROUP TO A (HEP) COLLABORATION

In order to implement the recommendations of the Blueprint, the Study Group became a Collaboration with a Collaboration Board and an Agreement signed by most of the key laboratories and some of the funding agencies. A "2020 Vision" was presented in February 2013 to the International Committee for Future Accelerators (ICFA) – a high-level body formed by Directors for HEP institutes worldwide to whom DPHEP reports.

The most recent report – in March 2018 in Cambridge – highlighted a specific problem whereby the 2PB of data collected (and still actively analysed) by the BaBar collaboration at Stanford Linear Accelerator Center may have to find a new home in the relatively short term. Through the DPHEP Collaboration, the possibility of storing the 2PB of data in CERN's tape robots (and/or at an institute that is part of the BaBar collaboration) is being investigated. This has been given Director-level approval but needs further discussion, including implications such as making at least some of the data available through CERN's Open Data portal. (It is the author's strong opinion that the data should be copied to CERN during LHC LS2 – see below – and before the CERN Directorate changes in 2021!)

Other examples of "internal HEP" collaboration is the establishment of a CVMFS repository at CERN for non-CERN experiments where the necessary (modest) resources can no longer be provided by the former host laboratory.

PROGRESS SINCE THE DPHEP BLUEPRINT

As documented in the iPRES 2016 paper, CERN now offers production services in all of the above three areas. There are constant improvements to "bit preservation", even as the data volumes continue to grow and as additional concerns arise, such as the reduction of enterprise tape vendors from two (Oracle and IBM) to one (just IBM). Clearly, alternative but nevertheless cost effective solutions are being investigated.

The use of CernVM and CVMFS (to snap-shot all the different s/w versions and environment needed by the preserved data) has become almost a de-facto standard across HEP data preservation (and also production) efforts and has been a significant success story that was not predicted by the Blueprint. (A paper on its potential was, however, submitted to the ESPP update).

Open Data releases, most notably by the CMS experiment who have now released over 1PB of data, together with the associated documentation and software, has been another big win, with successful re-use of the data by a variety of communities (including leading to a publication not involving any members of the CMS collaboration).

As has been described elsewhere, CERN is pursuing ISO 16363 for its range of data preservation activities, including Scientific Data (e.g. that from the LHC and former LEP collider), its "digital memory" (videos, photographs, minutes etc.) as well as papers, reports, proceedings, circulars and so forth. Many of the metrics in ISO 16363 are matched by existing CERN (documented) practices, whereas in a small number of cases, e.g. business continuity and disaster preparedness, there is still work to do (the latter being done in coordination with other EIROforum institutes). Given the cost and value of CERN's data (together with the fact that we believe we are already quite closely aligned

with this "most thorough" standard), we believe that this is the most suitable approach. We fully understand that it might not be appropriate for all projects / institutes / communities but a "lighter-weight" approach simply does not give the same level of assurance. (Think of the standards that you would like to see applied before blasting off in a unique space rocket). We are clearly indebted to others who have trod the certification path before us and the best advice that we could probably give is the same as for answering examination questions – *"read the question before you try to answer it"*.

THE HEP COMMUNITY WHITE PAPER

Published at the beginning of 2018, the Data Preservation chapter of the HEP Community White Paper that focuses on the challenges of the next decade or so, confirm the above achievements. For example, it notes ***"bit preservation with an acceptably low error rate can now be considered a solved problem"***.

However, it lists other areas of future work that match closely with "the new world order", where increasing emphasis is placed on reproducibility of results (as well as re-use of the data for new analyses and purposes).

ANALYSIS CAPTURE AND PRESERVATION

One of the main new areas of work since the publication of the Blueprint has been on tools / services to capture enough information, including work-flows, to enable key analyses to be repeated and acceptably similar (not identical) results to be obtained. There is no space to describe this work in detail but further information can be found via https://analysispreservation.cern.ch/welcome and https://github.com/reanahub.

OPEN DATA AT THE MULTI-PB SCALE

Given that the LHC has just completed its 2nd "run1", with several more multi-year runs (and significantly more data to be acquired and analysed) before it, we can expect Open Data volumes well in excess of 10PB, possibly in excess of 100PB and conceivably even more.

Some issues around Open Data are largely independent of volume: for example, if an experiment releases 1% 0or 50% of a given dataset, the same amount of documentation and software will still be needed.

However, many people assume that Open Data implies "zero or low latency" that can have a significant resource impact on large data volumes. Is this required? Is it even useful? We know that recently released data is frequently accessed but this then falls away over a period of weeks. Interest may be re-kindled by future events and it may be more appropriate to have a more modestly sized cache for recent data, together with "featured data", that may include "data on request". This will require further discussion and clarification between data producers, consumers, service providers and of course funding bodies.

From a CERN viewpoint, the first data that has been successfully preserved is that from the 4 LEP experiments that took data from 1989 to 2000. During this period, the way in which data has been "findable" by the experiments, as well as the protocols through which it was "accessed", have changed considerably. These changes continue to take place during the LHC era every 3 – 5 years or so. This is at least partly due to the demanding performance requirements of HEP production and analysis which mean that specialised protocols and / or tools are often strongly favoured. Thus, it is important to consider also these needs – including the very large volumes of data and numbers of objects – before concluding that FAIR is fully understood and implemented. The debate will no doubt continue.

MULTI-DISCIPLINARY COLLABORATION

Multi-disciplinary collaboration has existed from many years through bodies such as the Alliance for Permanent Access, the Preservation and Archiving Special Interest Group, the Digital Preservation Coalition and of course this conference series, to name but a few.
Following on from an initial visit from experts at ESRIN to CERN in February 2018, what we are now proposing is more hands-on / technical work and information exchange, initially involving those EIROForum institutes involved in LTDP but potentially expanding to include other relevant parties. The information exchange that has already taken place in the EIROForum IT Working Group on Business Continuity is an example of the level of information exchange and we are proposing to hold possibly bi-annual technical working meetings (not conferences, not workshops – no suits, no ties) starting from later in 2018. CERN would be happy to host the first such meeting and subsequent events – if deemed successful – could rotate round other EIROForum members and / or interested parties.

ARCHIVING AND PRESERVATION IN THE CLOUD

Can we do it? Not in 4 pages, but both technically and financially this may be attractive. See the 4C project post-cards for some background hints.

INPUT TO THE NEXT UPDATE OF THE EUROPEAN STRATEGY FOR HEP

One of the key inputs is expected to be the successful certification of CERN as a Trustworthy Digital Repository. This, together with the associated Preservation Policy, Strategic Plan and surveillance audits will help ensure that preservation is "written into the fabric" of the organisation and that the necessary resources are provided long into the future.
Other inputs are expected to cover the new or re-enforced requirements regarding Open Data and reproducibility of results. Whilst the latter was already of concern at the time of the Blueprint it had not yet become a requirement from funding bodies. In addition, the "designated communities" were almost always the same as the producers: no Open Data policy, let alone release, had been made at that time. (Some collaborations / experiments were relatively flexible about who could become a collaboration member but this was still far from what today is understood by Open Data).

OUTLOOK AND CONCLUSIONS

We are well on the way to implementing our 2020 Vision, where all HEP archived data is easily findable and fully usable by the designated communities with clear (Open) access policies and possibilities to annotate it further. This will be built using best practices, tools and services that are well run-in, fully documented and sustainable, built in common with other disciplines and based on standards.

However, as has been previously noted, constant effort will still be required, in particular to handle the inevitable migrations, changes in technology and updated policies from funding agencies that will occur not only between now and then but also over the multi-decade period for which the data needs to be preserved.

[1] LHC Run1 took place from 2009 to early 2013 and was followed by a "long shutdown" (LS1). Run2 started in mid-2015 and will continue until late 2018 when LS2 starts. Run3 should begin in mid-2021 and so on

20-YEARS OF ESA SPACESCIENCE DATA ARCHIVES MANAGEMENT

Christophe Arviset[1], Deborah Baines[1], Isa Barbarisi[1], Sebastien Besse[1], Guido de Marchi[2], Beatriz Martinez[1], Arnaud Masson[1], Bruno Merin[1], Jesús Salgado[1]

[1]ESAC Science Data Centre, ESA, Madrid, Spain

[2]ESAC Science Data Centre, ESA, Noordwijk, The Netherlands

In the mid-90s, ESA decided to change its data management strategy and started to build at ESAC (European Space Astronomy Centre), data archives for its space science missions, initially for the Infrared Space Observatory and then expanding through other astronomy missions and later on, to planetary and solar helio physics missions. The ESAC Science Data Centre now hosts more than 15 science archives, with various others in preparation.

Technology has evolved a lot through this period, from the simple web pages towards rich thin layer web applications, inter-operable and VO (Virtual Observatory) built-in archives. Maintaining old legacy archives while building new and state of the art ones (eg Gaia), managing people and preserving expertise over many years, offering innovative multi missions services and tools to enable new science (ESASky) have been some of the many challenges that had to be dealt with along these years.

Future prospects ahead of us are also looking exciting with the advent of the "Archives 2.0" concept, where scientists will be able to work "within" the archive itself, bringing their own software to the data, sharing their data, code and results with others. Data Archives have been and continue to be in constant transformation and they are now evolving towards open and collaborative science exploitation platforms..

ESAC SPACE SCIENCE ARCHIVES: AN EVER GROWING FAMILY

In the past, ESA was leaving to the scientific community the role of archiving its data holdings (for example for IUE or EXOSAT), In the mid-1990s and with the advent of the WWW offering new possibilities for data searches and dissemination, ESA changed its strategy and decided to host on-line archives with for its space science missions. This started with the ISO Data Archive in 1998.

Based on its success, it was decided to re-use the existing expertise and to develop the XMM-Newton Science Archive (released in 2002). Herschel and Planck archives were ready by launch (2009), representing an important part of the missions' Science Ground Segment. EXOSAT and SOHO archives were also added in 2009, using new Java development standards.

Together with the expansion of ESAC activities towards ESA planetary missions, the Planetary Science Archive (PSA) was built, hosting all ESA planetary data (initially Giotto (2004), Mars Express (2005) and Huygens (2006)). As all these missions were using the same data format (PDS for Planetary Data System), it was decided from the start to consolidate them all into a multi mission archive. The PSA was later on nenriched with data from Venus Express (2009), SMART-1 (2010), Rosetta (2010) and now Exomars TGO (2016). Consolidation of ESA space science archives to ESAC continued with the migration to ESAC of the European HST Archive from ESO (2012) and the Ulysses and Cluster archives from ESTEC (2013). Building on the wide variety of astronomy archives at ESAC, a major milestone was reached in 2015 with the release of ESASky, bringing all ESA astronomical data holdings (plus some others) through an innovative and user friendly all sky viewer and exploratory tool. In 2016, Gaia Data Release 1 was made public, including most of the Hipparcos catalogues, as well as the archive from LisaPF.

Overall today, ESAC Science Data Centre (http://archives.esac.esa.int/) hosts a dozen of archives containing science data from over 20 space science missions.

ESAC SCIENCE ARCHIVES STRATEGY

In 2012, taking into account the vast and wide variety of ESA scientific data holdings available, we defined

the ESAC Science Archives long term strategy articulated around three main pillars described below.

Enable Maximum Science Exploitation. The science data is the ultimate delivery of any ESA space science mission. One of the metrics used to determine the success of a mission is the number of refereed scientific publications in the literature. Therefore, the archive must provide the best science data together with all the necessary services and tools to maximize its science exploitation. ESA has the responsibility to ensure that the data hosted in its archive are of the highest quality, scientifically validated or even peer-reviewed in some cases, hence fully reliable and with the associated level of documentation, to allow scientists to do their science and then write their scientific papers.

Enable efficient long-term preservation. The archives must remain available for a long time, much longer than the current mission lifetime which already typically spans over 15 years. ESA has a commitment to preserve in the long term not only the data and associated services to access these, but as well the knowledge about this data. By consolidating all ESA space science archives in one place under the umbrella of the ESAC Science Data Centre, we can ensure strong re-use of technology and people expertise across archive projects. People working on active archives can also maintain the legacy archives, which also brings cost savings. Additionally, recognizing that the IT technology evolves rapidly while archive systems must perdure for many years, technology migration for archives will be required every five to seven years to ensure state of the art services.

Enable cost-effective archive production by integration in projects. ESA Science Operations have traditionally been organized by individual missions whereas the archive development, operations and maintenance is a transversal service to all missions. In the past, archives were often developed only in the final stage of the operations of a mission. Nowadays, with even more distributed and complex Science Ground Segment systems, the archive becomes sometimes the heart of the overall system (eg for Euclid) and therefore needs to be developed in the very early phases of the missions.

ARCHIVES DESIGN TECHNOLOGY EVOLUTION

Architectural design does matter a great deal when building an archive system that must be preserved for decades. Modularity and flexibility are key concepts in archive systems architecture, to facilitate technology evolution through time. The right balance needs to be found between providing state of the art services with newer technology options (including migration of existing systems) and avoiding the technology buzzes that won't survive long.

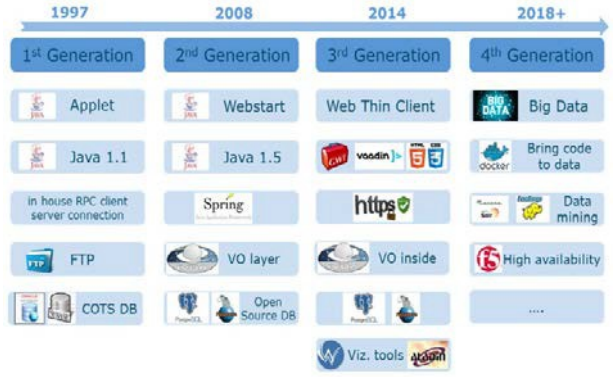


Figure 1: ESAC Archives Technology Evolution

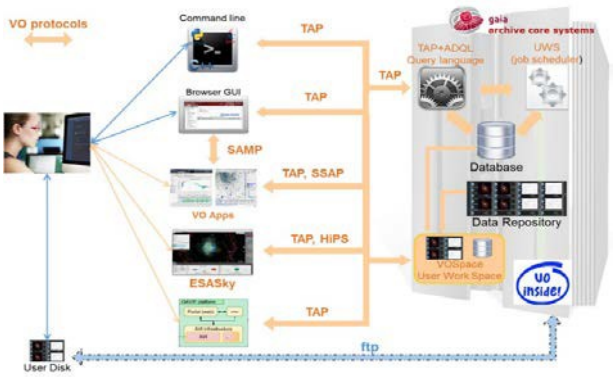


Figure 2: Gaia Archive Architecture

Over the last 20 years, ESAC archives have gone through important technological migrations. RDBMS systems went from COTS (Sybase, Oracle) towards open source solutions (PostgreSQL), enriched by discipline specific excellent plugins (eg pgSphere and PostGIS), and new databases systems are now being investigated (PostgresXL, CitusData, [2]) to address big data challenges brought by new missions (Gaia [3] and Euclid).

An application server allows to separate the data and metadata from its presentation as well as ensuring many other functionalities (caching, security, activity login, etc...). While we had to develop our own software in the early days, we then used existing frameworks, such as Spring and Hibernate which are IT industry standards and therefore facilitate greatly the development and maintenance.

On the GUI side, while Java was the best choice in the late 90s to build rich GUI archive interfaces, its support became poorer and poorer, while other web technologies become more advanced. In the late 2000s, we decided to migrate all our archives GUIs towards web thin clients, providing faster loading, no Java installation and overall better browser support. We chose GWT to continue benefitting from our experience Java software developers.

To ensure interoperability with other archives, we had been developing a VO (Virtual Observatory) layer on top of the existing archive APIs, building VO services through standard protocols (in particular VOTable, Simple Image Access Protocol, Simple Spectra Access Protocol, Simple Line Access Protocol) and connecting to external VO Tools through SAMP (Simple Application Messaging Protocol). With more advanced archives such as Gaia and Euclid, we started to directly use the VO protocols (eg Table Access Protocol, Universal Worker Service, VOSpace, Datalink) to other synchronous and asynchronous archive services. In this context, Gaia is definitely the first VO-built-in archive. It is also interesting to note that some of the VO protocols (eg TAP, SAMP), initially designed for astronomical data are being used for archives in other scientific disciplines (planetary and solar heliophysics).

TOWARDS ARCHIVES 2.0 PARADIGM

The traditional way scientists usually interact with the archives can be characterized by the "bring the data to the user" concept. Scientists go to the on-line archive, perform queries to determine which data they want, usually supported by some light weight visualization tools, and then download the data to theircomputer. From there, they use standard data analysis packages or their own scripts to analyze the data further and then later on write their scientific papers.

But new missions are bringing unprecedented amount of data, in the order of hundreds of GBytes or even PBytes and this calls for new models to access and interact with the data. First, querying billions of data holdings and cross matching them with other catalogues might require longer than what is expected for an interactive query session, hence the need to provide asynchronous services where complex queries can be queued, executed in the background and then provide results to the scientist after a few minutes. Results of such queries can still contain hundreds of millions of results and might be better stored (and indexed for performance) into user tables remaining into the archive database itself so the scientist can use it for further queries' refinement. Second, the scientist cannot download anymore all the data to her computer, as this would take too long and she probably would not have enough disk space anyway. It is up to the archive to provide user workspaces both for database (for user tables as seen above) and for data storage (done through VOSpace for example), so the data does not need to be transferred over the network.

When the data reside in the user workspace in the archive itself, the scientist wants to run standard data analysis package or her own software and scripts onto her data. This is the new archive concept "bring the code to the data". Most probably, archive data centres will also have to provide computing facilities next to their data so archive users can work with the data where the data actually resides. This could be done through dedicated cloud hardware infrastructure at the data centre itself (or eventually an hybrid solution involving external clouds if some data can also be copied there). New technologies (eg Docker containers, Jupyter notebook) should facilitate this implementation and initial examples look very promising.

This new concept of "Archive 2.0" would also allow scientists to collaborate much more easily. Users could share their workspaces (database table, data storage, but as well their own software and scripts) with other archive users. We think that from the original metadata and data repositories, the archives are evolving

towards open and collaborative science exploitation platforms.

ESASKY: BIG DATA VIZUALIZATION

As part of ESA strategy to increase science exploitation of its data holdings, the ESAC Science Data Centre built a completely new tool, called ESASky (http://sky.esa.int/). ESASky ([1]) is a new science-driven discovery portal for most ESA astronomical missions that gives users worldwide a simplified access to high level science ready products from ESA and other data providers (ie NASA, JAXA). The tool features a sky exploration interface and a single/multiple target search interface. It does not require any prior knowledge of the specific details of each mission. Users can explore the sky in multiple wavelengths, quickly see the data available for their targets and retrieve transparently the relevant science products from the corresponding archives.

ESASky is making full use of protocols that have been developed within the IVOA (International Virtual Observatory Alliance), which enable interoperability between astronomical data. On the client side, visualization is made fast by using Hierarchical Progressive Survey (HiPS), which splits the sky into various levels of tiles (depending of the mission) to minimize data transfer from the server to the client. Mission coverage (footprints) is described with MOC (Multi-Order Coverage). On the server side, other techniques

like TAP services on common data models for fast and performant searches, database geometrical indexes, internal connections between databases and wrappers around the individual mission archives to download the final science data have been setup to allow the handling of big amounts of data in a simplified way.

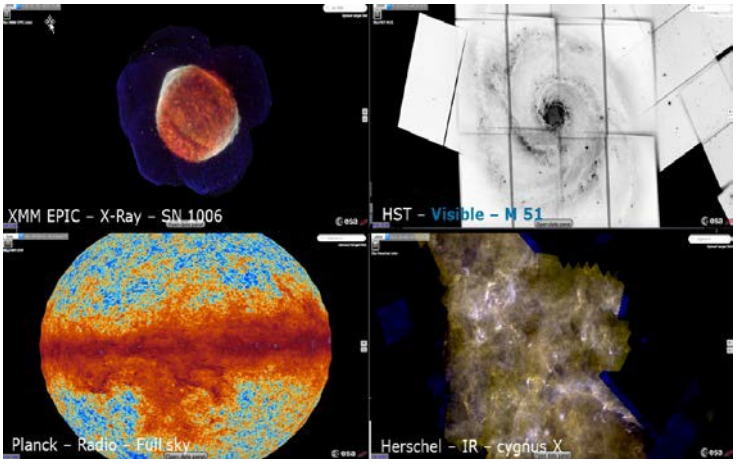


Figure 3: ESASky, see the sky with different "eyes"

CONCLUSIONS

Since the first public release of the ISO Data Archive in 1998 to the most recent Gaia archive release in September 2018, the ESAC Science Data Centre has converted itself into ESA's digital library of the universe, presenting and preserving reliable space science data for over twenty scientific missions. ESA Space Science Archives strategy is clearly articulated towards maximizing the science exploitation of data, ensuring long term preservation of data, knowledge and software, while supporting the development and operations of the Science Ground Segments.

This can be achieved through very close integration of scientists and software engineers, ensuring archives are science driven, and supported by strong IT expertise that need regular technology migration through time. To cope with new archive challenges (open data, big data volume, need to bring the code to the data, open and collaborative archives), a new paradigm for archive development and archive users is ahead of us that will bring the archives towards an exciting era that will revolutionize the way scientists interact with data..

REFERENCES

[1] B. López Martí, B. Merín, F. Giordano, D. Baines, E. Racero, J. Salgago et al., “ESASky: The whole of space Astronomy at your fingertips”, Proceedings of the XII Scientific Meeting of the Spanish Astronomical Society, July 2016, arXiv:1610.09826.

[2] P. de Teodoro, S. Nieto, J. Salgado, C. Arviset, “Considering scale out alternatives for big data volume databases with PostgreSQL” BIDS 2017 conference.

[3] A. Mora, J. Gónzalez-Nuñez, D. Baines, J. Durán, R. Gutiérrez-Sanchéz, E. Racero, J. Salgado, JC Segovia, “The Gaia Archive” Proceedings IAU Symposium No. 330, 2017, arXiv:1706.09954.

The Norwegian National Ground Segment; Preservation, Distribution and Exploitation of Sentinel data.

Halsne, Trygve, Ferrighi, Lara Saadatnejad, Bard, Budewitz, Nico, Dinessen, Frode, Breivik, Lars-Anders Godøy, Øystein
Norwegian Meteorological Institute (MET Norway), Henrik Mohns Plass 1,
0313 Oslo, Norway
trygve.halsne@met.no

In order to take advantage of the Sentinel program, the Norwegian Space Centre decided to establish a national collaborative ground segment with the purpose of simplifying data access and ensure support for operational national services. This is the NBS where MET Norway has the technical responsibility in terms of providing the infrastructure and storage capacity for data management. Serving the data through two separate platforms (i.e. DHuS and satellittdata.no), the end users have access to the data in its origin format in addition to Sentinel-2 products in
NetCDF-4/CF. Using the latter format, services like regridding, subsetting, visualisation and aggregation are integrated utilizing OPeNDAP in combination with WMS and WPS. In addition, data uploading and retrieving operations are simplified for an end user since streaming of data by means of OPeNDAP is supported in multiple programming languages. Due to the strong coupling between earth observation and non-earth observation products, NetCDF/CF is convenient for seamless integration across branches and well suited for data distribution. However, the current CF version is not mature for handling all parts of the Sentinel data but future development looks very promising.

1. Introduction

The European Space Agency (ESA) has developed a new series of satellites which are dedicated to specific missions. These are the Sentinels which are focusing on the operational needs of the Copernicus programme. A number of satellites have been planned and some have been launched, producing data which can be used for both operational and scientific purposes.

Data access is a crucial issue for exploiting the potential of such a large satellite constellation. Ever since the first satellite programs were launched, the access to remote sensing products has been limited in terms of costs and restrictions [1]. Hence, only agencies and institutions with necessary resources and special interests (e.g. security instances and meteorological communities) have been able to retrieve and work with these types of geophysical data. The paradigm shift introduced by NASA in 2008 with Landsat and later through the Sentinel program, where remote sensing products are available for the public under a free data policy, enables the exploitation of data for a broader set of institutions, for SMEs, but also for private persons, thus also enhancing data value [2].

On behalf of the Norwegian Space Centre, the Norwegian Meteorological Institute (MET Norway) is developing and implementing the National Ground Segment ("Nasjonalt Bakkesegment", NBS) for satellite data, following the FAIR (findable, accessible, interoperable and reusable) principles of data management [3]. Currently, data are served through two separate platforms: colhub.met.no, using the DHuS[1] software suite and satellittdata.no, a system focusing on an open data space prepared for integration of non Earth Observation (EO) data as well. Targeting both expert and non-expert satellite users, the NBS setup in satellittdata.no is designed using lessons learned in data management efforts like the International Polar Year (IPY) as well as a number of national geoscientific e-infrastructure projects supported by the Research Council of Norway (including e.g. Norwegian Scientific Data Network - NorDataNet and Norwegian Marine Data Centre - NMDC). This implies that the system is metadata driven following the same approach as the World Meteorological Organization (WMO) Information System and INSPIRE[2], but emphasising the need for semantic translations and dynamic transformation of datasets upon user request. In this article the preservation, distribution and exploitation of Sentinel data will be discussed by means of these two portals with emphasis on the NBS setup.

2. Methods

2.1 ESA Collaborative Ground Segment

MET Norway is participating in ESAs Collaborative Ground Segment (CGS). Sentinel products are retrieved through the ESA distribution node and other CGSs. Data retrieved over the Norwegian area of interest are preserved in a long-time archive implemented using a high performance data storage with integrated integrity checking.

As part of the ESA CGS, the DHuS software suite was integrated allowing for utilisation of the setup in terms of both data retrieval and distribution. In addition to a web GUI, the system provides an API for data access and discovery using Open Data Protocol and OpenSearch. The easy setup of the system allows for serving the data for any user with limited amount of working hours in terms of software development and implementation.

2.2 NBS Distribution Setup

The following parts are wrapped together in the MET Norway Scientific Information System (METSIS), a custom module integrated in the Drupal Content Management System using an open source platform.

2.2.1 Data Format and Distribution

The Sentinel products are packed in the Standard Archive Format for Europe (SAFE)[3], which is compliant with the Open Archival Information System (OAIS) standard. The geophysical measurements, e.g. synthetic aperture radar and multispectral imagery data, are stored in various file formats within the SAFE structure such as GeoTiff and jpeg2000. In the first version of the NBS setup, focus have been put on performing a lossless transformation of data from SAFE to NetCDF-4[4] (Network Common Data Form), following the Climate and Forecasts convention (CF)[5]. NetCDF is an open-source software, developed and supported by UCAR's Unidata Program[6], that allows for creation, access and sharing of scientific data through machine-independent formats. Traditionally, NetCDF has been heavily used in communities working with atmospheric and ocean modelling, but the widespread of use of the format is growing in terms of branches and communities [4]. NetCDF-4 results from a collaboration with the group developing HDF5 (Hierarchical Data Format), which is widely used for remote sensing. The NetCDF-4 data files are self-describing if combined with the CF convention, defining how Earth Science data should be encoded in terms of structure and metadata [5, 6].

Serving the data through a server supporting the Data Access Protocol (DAP), the utilization of OPeNDAP[7] (Open source Project for a Network DAP) gives access to data stored at remote locations through data streams, removing the need for downloading data prior to their usage. Also, creation of virtual datasets by means of aggregation, i.e. combining multiple datasets, and/or subsetting of a potentially large dataset is possible. As a result from collaborative efforts between Unidata and OPeNDAP, a tight fusion of NetCDF with the DAP is achieved. In the setup, data access is provided using Unidata's THREDDS Data Server (TDS). In addition to OPeNDAP, the TDS client-service architecture supports a number of protocols for accessing the information in the NetCDF files.

2.2.2 Data Visualization

For data visualization, the Open Geospatial Consortium (OGC) Web Map Service (WMS) standard for georeferenced scalable data visualization is used. In addition to the ncWMS

client-service on the TDS, a MapServer (version 7.0.0.) has been configured to fit the project purposes. The visualization is carried out in an OpenLayers version 3 client.

2.2.3 Data Transformation

Exploiting OPeNDAP, we have implemented a OGC Web Processing Server (WPS) using the framework of pyWPS. On the server, a transformation process allows for product reprojection, resampling and subsetting in terms of temporal/spatial extent and variables. All target projections supported by Proj[8] can be implemented. The underlying software carrying out the transformation is
FIMEX[9], a C/C++ software developed by MET Norway performing file manipulation, interpolation and extraction on gridded geospatial data. The software is build around the Unidata Common Data Model. Hence, the NetCDF products on the TDS can be accessed directly and transformed according to the user request. Moreover, the processes sent to the WPS runs asynchronously which allows for multiple transformations simultaneously.

2.2.4 Discovery Metadata

The system relies on discovery metadata records storing information about the scientific datasets. This is the cornerstone of the whole system where all the necessary information for carrying out the steps described above are present by means of semantic notation. The metadata records are following the MET Norway MetaData format (MMD) which is compatible with GCMD DIF, ISO19115/ISO19139 - standards imposed by WMO and Norge Digitalt/INSPIRE, but also extends these. The purpose of the format is to document datasets, not web services. Information on the web services for a dataset are provided through a dedicated element. All available metadata records are ingested in an Apache SOLR database. In addition, the records are pushed to an Apache subversion server where an OAI-PMH[10] server makes our metadata available for other institutions to harvest.

3. Results and Discussion

Initially, the primary focus of NBS has been on delivering all Sentinel products through the DHuS while simultaneously making Sentinel-2 level 1C and Sentinel-1 GRD products available in the NBS setup. Wrapping all the parts together, the NBS takes care of discovery, access, visualization, transformation and long time preservation of Sentinel data. There are, however, challenges. Although NetCDF/CF has become the de facto standard for several satellite based level 3 and 4 datasets in some branches [4], some information in the lower level products does not apply in the current CF convention version (i.e. 1.7). An example is auxiliary data for Sentinel-2 stored as geography markup language files containing polygons. Currently, this could be solved by rasterizing the polygon information at the cost of dissolving the original structure. However, there will be support for polygons in CF 1.8 - supporting geometries. In addition, standardized encoding of Earth Science swath data in instrument viewing geometry is under development and a proposal was provided late 2017. An obstacle in the current version is that CF has a flat structure. The intention of CF 2.0 is to support groups and other features of the underlying storage model. This would simplify handling of satellite data, but this is not an option today.

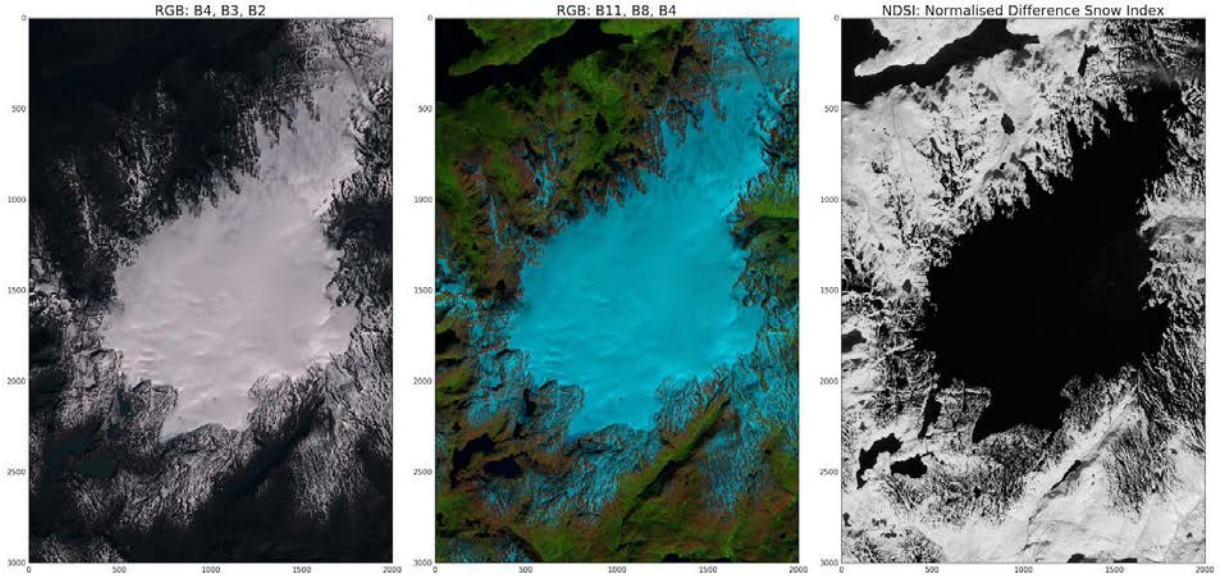
For all geophysical science, researchers rightly concentrate their efforts on dealing with data, both for model development and data processing, rather than on data uploading and retrieving operations, which are nowadays easily achievable by exploiting OPeNDAP. Widely used programming languages (e.g. Python, R, Matlab, C, Fortran) have modules that supports reading of data using OPeNDAP. Hence, distributing data through DAP could potentially reduce time and disk consumption for the user by means of data streaming. In Figure 1, an example of spatial extent and spectral bands subsetting of a Sentinel-2 product accessed and plotted by means of OPeNDAP in Python is presented. The Sentinel-2 products in NetCDF-4 are processed to a higher level in terms of

resampling all spectral bands to highest spatial resolution (i.e. 10x10m pixel values) using a nearest neighbor algorithm to easy retain the original resolution without data loss.

The NBS setup is the same used in other contexts at MET Norway. Hence, integration of other geoscientific products describing e.g. atmospheric conditions and sea ice extent could potentially ease the process chain for a user in terms of combining relevant products.

The first milestones in future development will be to: Finalize and provide a Sentinel-1 GRD product in the system; add Sentinel-3 ocean and land cover imagery products along with sea and land surface temperature radiation products which already are in NetCDF (but not CF); integrate an OpenSearch API in SOLR. Last, identify the need for providing a Virtual Research Environment where end users could upload and run their own algorithms.

Figure 1: Two RGB composites (left) and Normalised Difference Snow Index (right) from Sentinel-2, covering the Folgefonna glacier in Norway. The images are generated in Python accessing the data by means of subsetting through OPeNDAP.



4. Conclusion

The NBS project takes care of discovery, access, visualization, transformation and long time preservation of Sentinel data covering Norwegian areas of interest. Distributing various Sentinel products using NetCDF-4/CF, the data can easily be accessed, combined and subsetted by means of OPeNDAP. Future development of the CF standard makes the format more suitable for EO products.

5. Endnotes

1 https://sentineldatahub.github.io/DataHubSystem/

2 https://inspire.ec.europa.eu

3 http://earth.esa.int/SAFE/

4 https://www.unidata.ucar.edu/software/netcdf/

5 http://cfconventions.org/

6 https://www.unidata.ucar.edu/

7 https://www.opendap.org/

8 http://proj4.org/

9 https://wiki.met.no/fimex/start

10 https://www.openarchives.org/pmh/

6. References

[1] W. Turner, C. et al. (2015) Free and open-access satellite data are key to biodiversity conservation. *Biological Conservation* (182)

[2] Mathae, K.B., Uhlir, P.F. (Eds.), 2012. The case for international sharing of scientific data: a focus on developing countries. In: Proceedings of a Symposium. National Academy of Sciences, Washington, DC

[3] Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* (3), nb 160018

[4] Hankin, S. et al. (2009) NetCDF-CF-OPeNDAP: Standards for ocean data interoperability and object lessons for community data standards processes. *Proceedings of OceanObs'09* (2)

[5] Russ Rew, Glenn Davis, Steve Emmerson, and Harvey Davies (1996), NetCDF User's Guide, Version 2.4 Unidata Program Center, University Corporation for Atmospheric Research, Boulder

B. Eaton, J. Gregory, B. Drach, K. Taylor, S. Hankin, NetCDF climate and forecast (CF) metadata convention, 2018

The Data Distribution Centre of the Intergovernmental Panel on Climate Change: Support for users in challenging IT environments

Charlotte Pascoe[1], Martin Juckes[1], Ag Stephens[1], Martina Stockhause[2], Bob Chen[3], Xiaoshi Xing[3]

1. National Centre for Atmospheric Science – Science and Technology Facilities Council
2. German Centre for Climate Computing, DKRZ
3. CIESIN - Columbia University

Email: charlotte.pascoe@ncas.ac.uk

About the DDC

The Data Distribution Centre of the Intergovernmental Panel on Climate Change (IPCC) was established in 1997 as a distributed data centre with partners in Germany, Great Britain, and the United States. The IPCC is primarily concerned with assessing research literature on climate change in order to provide policy relevant advice to Governments. These assessment reports have substantial impact: they are driving the global transition away from the fossil fuel economy. The IPCC achieves this impact through an exhaustive review process which ensures participation of experts from all 197 parties to the United Nation Framework Convention on Climate Change. The assessment of scientific literature is backed up by illustrations calculated from key underlying datasets.

The Data Distribution Centre (DDC) supports the work of the IPCC by facilitating data sharing between the author teams and by ensuring the long-term preservation of key datasets. It also provides extensive guidance on the use of data for climate analysis and climate impact studies that is prepared by the IPCC Task Group on Data and Scenario Support for Impact and Climate Analysis (TGICA).

The data and guidance documents, like the work of the IPCC, cover all policy relevant aspects of climate change, from the physical basis of climate projections to economic analysis of the options for mitigating both climate change and the impacts of climate change. This diversity of data together with the interdisciplinary data usage creates unique challenges for the small team at the DDC seeking to ensure long-term preservation of data and ongoing support for an increasingly diverse and interdisciplinary user community.

Big Data in Challenging IT Environments

The DDC is a global resource that is used by people all around the world. We receive over 3000 page views per week on average with a roughly even share between users in Europe (30%), Asia (30%) and North America (24%). Smaller percentages of our users are based in Central and South America (8%), Africa (5%) and Oceana (3%) however DDC engagement from African users has seen a dramatic increase in recent years, an additional 15 African countries began using the DDC between 2010 and 1015.

Much important climate change science is done by scientists who are located in places that are not robustly connected to global information technology (IT) infrastructures. Poor infrastructure such as intermittent power supply and internet connectivity, limited processing capacity and storage and lack of sophisticated software analytic tools present significant barriers to scientific experts in developing countries. It is not uncommon for scientists in Africa to travel to a neighbouring country where IT infrastructure is more robust to download climate data. Scientists based on remote Pacific islands rely on satellite data links to access data and have seen their share of the data bandwith reduce as the wider community has begun to use smart phone technologies more and more.

The IPCC-DDC is able to respond to data requests by sending data by post however this route presents it's own challenges: the data can only be sent to a single scientist, it takes time for data to travel by post and navigating customs can add further delays (it has been know for DDC data disks to be destroyed by customs officials). Sending data by post does not address the issues such as limited processing capacity and access to analytic tools that scientists in developing countries can face.

Bringing Scientists to the Data

A new data access paradigm where scientists are brought to the data and are able to interact with data at source has been set up by the DDC team at the Centre for Environmental Data Analysis (CEDA). The IPCC DDC Interactive Server is a computer managed by CEDA on the JASMIN compute platform in the UK. CEDA runs a multi-petabyte data centre that includes climate and earth observation data sets held in the British Atmospheric Data Centre and the NERC Earth Observation Data Centre. These include the CMIP5, CORDEX and ERA-Interim data sets that are known to be of significant interest to scientists using the IPCC DDC. Users of the IPCC DDC Interactive Server will also have timely access to new data from CMIP6, ERA-5 and Sentinel is brought into the CEDA archive.

Users from around the world can connect to the JASMIN environment via a Login gateway server via the SSH protocol. The SSH protocol is of particular use for users in challenging IT environments as it does not rely on having continuous internet connectivity. Once they are logged on to the Interactive Server users can read data from the existing CEDA Archive which is mounted as local disk. Data output can be written to the 20TB Group Workspace which has been reserved specifically for use by the DDC community.

The IPCC DDC Interactive Server has the potential to revolutionise the work of climate scientists working in challenging IT environments and to make a significant impact on their ability to participate in the IPCC assessment process.

PROBA-V MISSION EXPLOITATION PLATFORM AND TERRASCOPE

Martine Paepen (martine.paepen@vito.be**), Erwin Goor (**erwin.goor@vito.be**), Dennis Clarijs (**dennis.clarijs@vito.be**)**

VITO, Boeretang 200, 2400 Mol, Belgium

End 2017 BELSPO (the Belgian Science Policy Office) and VITO Remote Sensing have launched TERRASCOPE, the Belgian collaborative ground segment to maximize the usability and uptake of the Copernicus satellite data by Belgian users. The TERRASCOPE platform takes advantage of existing infrastructure that Belgium and VITO built up for the SPOT-VEGETATION and PROBA-V missions.
The PROBA-V MEP (Mission Exploitation Platform) complements since 2016 the PROBA-V user segment by offering an operational exploitation platform on the PROBA-V data, correlative data and derived products, as well as selected high-resolution data/products. The platform combines scalable processing resources with a large data archive and a rich set of tools. End users can use the time series viewer tool to explore SPOT-VEGETATION and PROBA-V time series, complemented with meteo data and derived indicators for vegetation products. The GeoViewer tool allows the user to view the PROBA-V data in full resolution and is based on OGC standard based web services which can also be accessed using simple web browser or tools as QGIS. By using a Virtual Machine on the PROBA-V MEP, anyone can access the virtual research environment with access to the full archive and a powerful set of tools and libraries to work with the data or to develop-debug-test applications. Additionally, Jupyter notebooks provide a programming interface from a simple web browser for interactive data analytics with rich media output. Moreover, the platform allows users to co-work, share results and relevant documentation.

Key Words: MEP Mission Exploitation Platform, PROBA-V, data analytics on time series, on-demand processing, virtual research environment, TERRASCOPE

INTRODUCTION

In May 2013 the PROBA-V satellite [1] was launched to extend the time series of 15 years SPOT-VEGETATION data. PROBA-V is a micro-satellite aiming at monitoring the Earth's vegetation on a daily basis with a spatial resolution of 1 km and 1/3 km and 100m resolution with a global coverage every five days. VITO developed the PROBA-V user segment and is responsible for the image processing, geometric and radiometric calibration, archiving and distribution of all products as a continuation on the SPOT-VEGETATION time series for which VITO hosted the image processing, archiving and dissemination centre since 1998.
Next to these operational activities for PROBA-V, VITO hosts as well several other processing facilities, which e.g. offer hyperspectral images from the airborne APEX instrument or bio-geophysical parameters in the frame of Copernicus Global Land Service.
To cope with the heterogeneous nature and the huge increase of available EO-data, new approaches and technologies were implemented in the VITO Long Term Archive but certainly also in the approach of providing the data and value-added products to the user community. In that context an operational Mission Exploitation Platform (MEP) was released in September 2016 as a fully operational service to the users. Several applications are provided, e.g. a time series viewer, a full resolution GEO viewer, pre-defined on-demand processing chains, Jupyter notebooks and virtual machines with powerful tools and access to the full data archive. This allows users to design, debug and test applications on the platform. All these services are accessible from the on line PROBA-V MEP portal [2].
Building further on this existing infrastructure, VITO Remote Sensing launched TERRASCOPE, in agreement with ESA and BELSPO, as the Belgian Collaborative Ground Segment to ensure a low-threshold access to the Sentinel satellite images which are freely available for all Belgian users.

PROBA-V MISSION EXPLOITATION PLATFORM

The PROBA-V MEP provides scalable processing facilities with access to the complete data archive and0 a rich set of processing algorithms, models, open source processing libraries/toolboxes and public/collaborative software. The platform became the hub processing infrastructure of the mission by functioning as a powerhouse system and open access development environment.

To realise this the platform consists of the following components:

- The existing Product Distribution Facilities [3] and [4] are serving the access to the data archive, both via a Web portal as well as (OGC) standardised discovery, viewing and data access interfaces.
- Hadoop, as a software framework for data-intensive distributed applications, is designed to process large amounts of data by separating the data into smaller chunks and performing large numbers of small parallel operations on the data. Oozie and Airflow are used to design an EO-application as a workflow of multiple processes. Spark is used intensively on the MEP to allow analytics on large time series of data. The Hadoop ecosystem provides furthermore a rich and still growing set of tools which are used to give fast access to the data in a format needed by the specific application. As an example Accumulo and Geotrellis are used to offer data analytics on the large time series for user-defined polygons or single pixels as part of the Time Series Viewer.
- All EO raster data is accessible via NFS and possibly uploaded to the Hadoop Distributed Filesystem (HDFS).
- Cloud computing technology enables dynamic resource provisioning and is therefore providing a performing and scalable solution. OpenStack is chosen as private cloud middleware. Pre-configured virtual machines are offered and can run on the OpenStack cluster at VITO, providing the environment needed for users to work with the data and develop/deploy applications on the platform, i.e. containing IDE's, a rich set of tools and access to the complete data archive. Several external users are currently performing R&D on these VM's.
- Interactive Web-based geoviewers and dashboards are designed to provide user-tailored information from the EO-data archives of VITO and other providers, by combining existing components such as AngularJS, Javascript libraries and GIS components into one single solution. As an example, the PROBA-V MEP Time Series viewer allows you to view time series for any pixel of user-defined polygon for PROBA-V data, derived vegetation indices and meteo data. Remark that the derived vegetation indices are originating from the Copernicus Global Land Service.

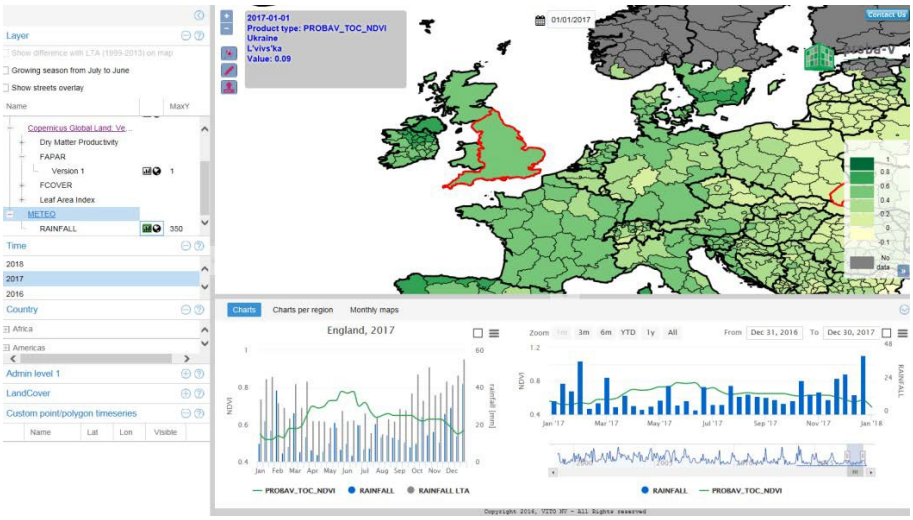


Figure 1: The PROBA-V MEP Time Series Viewer

- The Jupyter Notebooks Web application lets you create and share documents that contain live code,

equations, visualizations and explanatory text. It is based on the Open Source Jupyter notebooks application, and tailored to the needs of remote sensing users. For programming, users can choose between the Python and R programming languages and can work interactively with the full data archive available atthe PROBA-V MEP. An ever growing list of software libraries such as GDAL, rasterio, pandas, numpy, matplotlib and seaborn is included and users can also upload and install their own packages and file. The PROBA-V MEP Web portal provides several example notebooks, showing how to access data, how to use the time series viewer, how to plot charts and maps.

- A Web portal provides access to all applications and tools offered by the PROBA-V MEP and to the cloud consoles. Furthermore the portal provides all information on the data and components available on the platform and offers tools for e-collaboration and knowledge sharing amongst the users. Blogs and tutorials are added regularly to respond to technical questions from users
- A main concern in the architecture was security since we allow user to develop and execute their applications on the platform. Their IPR shall be properly protected and the activities of individual users cannot influence the stability of the system and the work of other users. Single sign-on and proper monitoring of used resources are further requirements.

Since the first pre-operational release, the PROBA-V MEP is used by beta-testers to provide early feedback. Several users are developing a processing workflow or porting an application on the PROBA-V MEP in the frame of the ESA MEP-TPS project. Often the PROBA-V MEP is hosting the data intensive backend service, while the frontend remains at the premises of the user. The users range from universities to SMEs from different European countries.

The usage of the PROBA-V MEP applications is increasing constantly from the early release of the platform. The geoviewer, providing a full-resolution viewing service is as well intensively used in the promotion activities for the PROBA-V user segment, e.g. in the weekly image-of-the-week. The time series viewer is used by several researchers, as well by educational users. E.g. an Earth Observation course was developed by the Belgian office of the European Space Education Resource Office (ESERO) which uses intensively the PROBA- V MEP time series viewer to illustrate long-term changes in vegetation e.g. deforestation or droughts.

The first on-demand processing service on the platform, the N-daily compositor application, is used by several PROBA-V users to compute N-daily composites of PROBA-V 1 km , 300 m or 100 m data with a sliding window for any area of interest. E.g. a user can request to get a 7-daily composite at 100 m spatial resolution every Monday over a given area of interest, and is no longer depending on the standard products offered by the PROBA-V user segment. Furthermore different compositing algorithms are provided.

The PROBA-V MEP is as well used in several R&D and pre-operational projects as a node within a federation. E.g. in the EC H2020 NextGEOSS [5] and DataBio [6] projects, the PROBA-V MEP is used to provide access to the full PROBA-V archive and Sentinel-2 data over limited areas, towards researchers and on-demand pre-operational processing chains. In the ESA TEP Food Security project [7], the PROBA-V MEP is serving the data analytics capabilities of the project, in a federation with public clouds such as IPT Poland.

TERRASCOPE

Building further on the Proba-V MEP solution, VITO was assigned by ESA and BELSPO as designated entity to develop the Belgian Sentinel Collaborative Ground Segment, named TERRASCOPE. The platform was extended to adopt Sentinel-1/2/3 data and derived products into a multi-mission platform, providing easy access to these data via WMS, WMTS and WCS interfaces and a cloud-based processing platform where users can upload-develop-test their own algorithms.
The mission of TERRASCOPE is to maximize and facilitate the uptake of Sentinel data by all Belgian users including industry, government, academia and the private citizens. It provides a flexible and dynamic platform allowing for new developments resulting from collaboration between Belgian partners in dialogue with BELSPO. The TERRASCOPE portal [8] provides the gateway to the Sentinel imagery, higher level

products derived from Sentinel imagery and computing power which third parties can use to generate their own information products. In a first phase, Sentinel 1A and 1B mirror data covering the region of Belgium are available together with the Sentinel 2 data which have been processed using the iCOR [9] algorithm for atmospheric correction. Additionally four derived vegetation parameters NDVI, fAPAR, fCOVER and LAI starting from SENTINEL 2 are made available. The Sentinel 2 data with coverage over Belgium is guaranteed for the archive, while coverage will be expanded to Europe and Africa in a later phase. Additionally the available archive will be enlarged with a subset of the Sentinel 3 data on a global scale.The Terraviewer, an online TERRASCOPE application[10], can be used to display and explore Sentinel collections. The targeted user groups for this portal range a wide spectrum, supporting both expert users with GIS experience and non-expert users mainly interested in viewing the available data. The Terraviewer provides several components to enable the exploration and exploitation of the available data. Next to the general viewing capabilities, it is possible to export an image of the current view, allowing the user to download the data in an image format. In order to support all users and devices, the portal works on both mobile and desktop devices with the added possibility to change the language of the interface to English, French or Dutch.

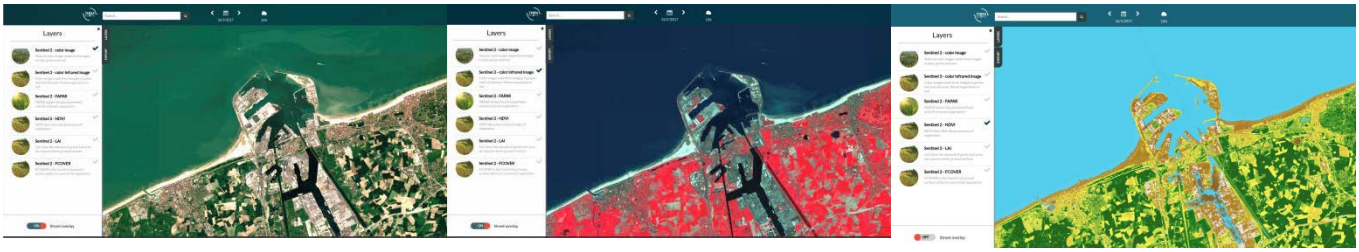


Figure 2: Terraviewer displaying Sentinel 2 natural color image, infrared image and NDVI.

CONCLUSION

To face the challenges of the exponential growth of heterogeneous EO data volumes and to ease the exploitation of these massive amounts of EO-data, the VITO Long Term Archive is integrated in the Mission Exploitation Platform that provides one comprehensive infrastructure offering data distribution, data viewing/analysis and on-demand processing supporting vegetation and agricultural related and wider environmental EO applications. By designing a solution which can be deployed both on a private cloud and on a public cloud hosting relevant data, the solution can address big data which goes beyond the capabilities of a single data centre. Furthermore, expert users can make use of powerful Web based tools and can self-manage virtual machines to perform their work on the infrastructure at VITO with access to the complete data archive. To realise this, private cloud technology (openStack) is used and a distributed processing environment is built based on Hadoop offering a lot of technologies (Spark, Yarn, Accumulo, etc.) integrated with several open-source components.
The impact of this PROBA-V MEP for the user community has been high and has completely changed the way of working with the data. It opened the large amount of time series of valuable EO-data to a larger community of users. With TERRASCOPE, the Belgian Collaborative Ground Segment, based on the PROBA-V platform, the usability and uptake of the Sentinel data is enhanced for the Belgian user community. In the future all PROBA-V and SPOT-VEGETATION data will be integrated in the TERRASCOPE platform to provide one gateway where users can search, view, analyse, download all these data. Furthermore the virtual research environment allows the development of user-driven applications with access to the complete data archive, encouraging the collaboration between Belgian partners.

REFERENCES

[1] http://proba-v.vgt.vito.be/.
[2] https://proba-v-mep.esa.int.
[3] http://www.vito-eodata.be.

[4] http://land.copernicus.vgt.vito.be/PDF/
[5] http://nextgeoss.eu/.
[6] https://www.databio.eu/en/.
[7] https://foodsecurity-tep.eo.esa.int/
[8] https://www.terrascope.be/index.html
[9] https://blog.vito.be/remotesensing/icor_available
[10] https://viewer.terrascope.be/terrascope/
[11] Proba-V Mission Exploitation Platform, Remote Sensing Journal, Technical Note, 2 July 2016, http://www.mdpi.com/2072-4292/8/7/564/pdf.
[12] Proba-V MEP Leaflet for developers, https://proba-v-mep.esa.int/sites/proba-v- mep.esa.int/files/documents/mep_fact.sheets_finalweb.pdf.

Sentinel Data Archiving at ESA

Nigel Houghton

European Space Agency, ESRIN, Italy nigel.houghton@esa.int

AUTHOR INFORMATION

Responsible for all of the Sentinel Long Term Archives at the Copernicus centres for Sentinel- 1,2,3 and 5p. Previously lead the group responsible for all the ERS and Envisat Processing & Archiving Centres.

The European Space Agency at ESRIN is managing the Long Term Archive of the data from the Sentinel satellites on behalf of the European Commission as part of the Copernicus Programme. The data is archived in various data centres across Europe who provide archive as a service with an infrastructure that has been chosen by the providers.

Since the launch of Sentinel-1A in April 2014, the amount of data in the archives has been growing at an ever increasing rate and now stands at over 20 PBytes and 9,000,000 files. The Long Term Archive is growing by 10 PBytes per year and represents the largest data set currently managed within ESA Earth Observation.

This presentation will describe the various diverse hardware solutions that have been chosen for the data sets and the experience gained from this and issues overcome. It also will examine the challenges of an ever increasing data set such as avoiding data loss. It will also raise and attempt to answer the question of what exactly is a long term archive and address the question “has any of the data been lost?”

Finally it will look at some of the particular technological challenges faced in archiving the ever increasing data sets including the infrastructure choices available for the future in a shrinking supplier market.

GENERAL OVERVIEW

Copernicus production is based on a systematic approach whereby all acquired scenes are processed to a pre-defined set of core Level-1 and/or Level-2 products. Newer versions of these standard data products may be generated and archived as a result of reprocessing campaigns. The Long-Term Archive (LTA) therefore not only responds to the need for long-term data preservation but also facilitates rapid availability of the latest product set to users and supports reprocessing. However, the cost of such archiving is a driver for the system operations and its long-term sustainability.

The LTAs for Sentinel data were in place before the launch of Sentinel-1A in April 2014. Procured on behalf of the European Commission through a call to industry, separate LTAs were put in place for each Sentinel mission for Sentinel-1 and Sentinel-2 and for each of the instruments for Sentinel-3. The data from the Sentinel B units is collocated with the data of the Sentinel A units. S5P production and archiving was added later on.

Each LTA stores products ranging from the basic Level-0 and auxiliary data to Level-1 and Level- 2 products. There are two LTAs for Sentinel-1 and for Sentinel-2 at different locations in order to provide a second copy of the archive to minimise the possibility of data loss overall and also the maximise continuity of service. For Sentinel-3, there is only one centre for each instrument whilst Eumetsat also has a copy of the Level-0, Level-1 and auxiliary data. In practice, many of the individual ESA centres have two copies of the data.

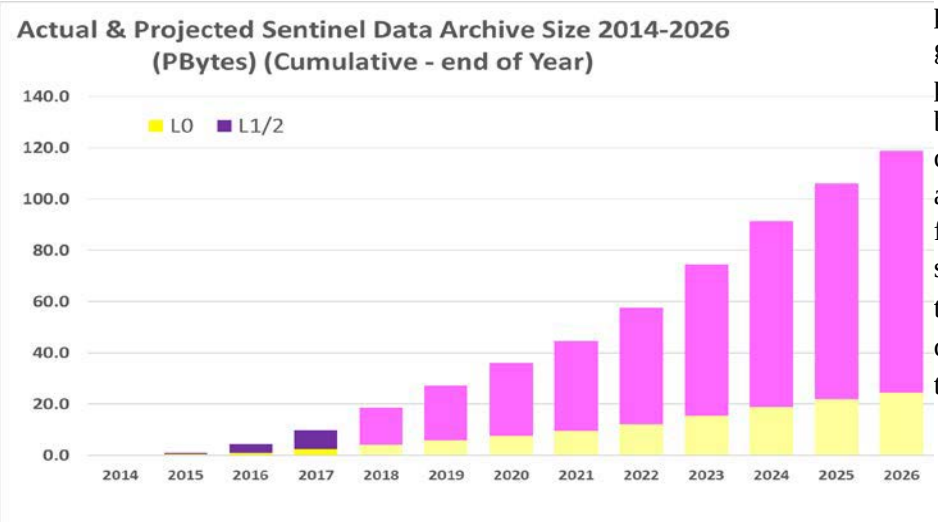
In requesting LTA services, ESA has not specified the hardware solution to be used instead concentrating on the required performance of the solution both from a technical performance standpoint and with an associated set of service performance levels expressed through a series of Service Level Requirements and associated Key Performance Indicators.

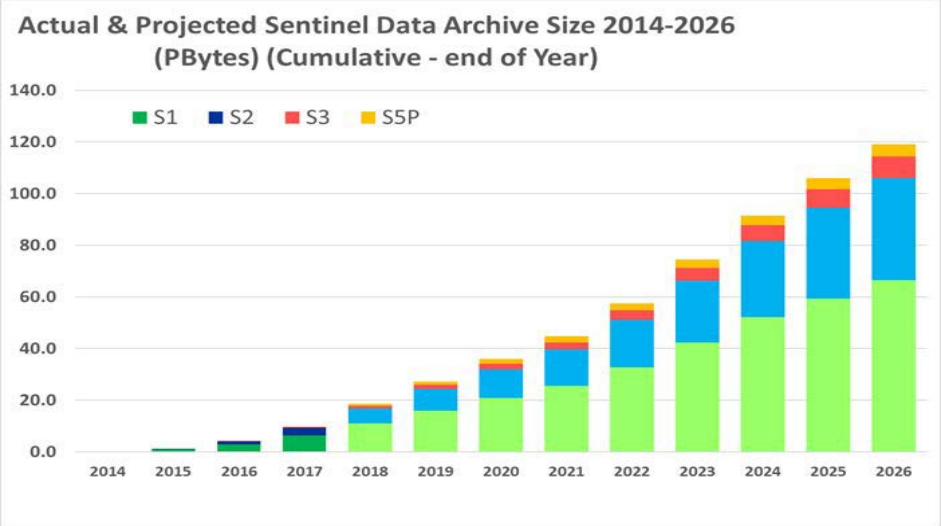
ESA's procurement approach has resulted in different technological solutions from its industrial partners summarised below (figures approximate as of April 2018)

Technical Solution Description	Summary	Size of Archive (TBytes)	Number of Tapes used (approx..)	Sentinel satellite
Quantum I6K robots (LTO-6 and LTO-7) StorNext hierarchical storage system. Cache is supplied by Hitachi G600 Enterprise storage using redundant fibre channel	Tape + Disk Cache	11000	4200	Sentinel-1, Sentinel-2
Oracle SL8500/T10000C (100088 slots)using Oracle SAM-FS; (Backup copy SL3000/LTO7)	Tape + Disk Cache	7500	1200	Sentinel-1, Sentinel-3 OLCI Sentinel-5P
IBM TS3500 Library with 8 TS1140 tape drives (250 Mbps). Tape Cartridges model 3592 (4 TB)	Tape + Disk Cache	3500	1100	Sentinel-2
DELL servers/Solaris OS/ZFS. Each unit is made of one R730 server connected to PowerVault M3060e disks arrays (60 x 8 TB HDD)	Disk	250	N/A	Sentinel-3 SLSTR/Synergy
Scality RING software. 10 data servers + 1 quorum server + 2 dedicated switches + 2 gateway servers	Disk	100	N/A	Sentinel-3 SRAL

DATA SIZING

Even if not massive by 21st Century standards, the Sentinel data set dwarves all other data sets within ESA Earth Observation. Since the first data was received in April 2014, the yearly volume has been increasing so that the archives, including the Sentinel-1 and Sentinel-2 backup archives, holds just over of 9 million products and 20 PBytes. Furthermore, the data archived in the year 2017 was a greater volume than all Sentinel data archived in all previous years combined. This growth is set to continue with a projected volume of 120 PBytes by the end of 2026 assuming the current plans for the Sentinel-C and Sentinel-D units and the plan for parallel operations. This is the size of one copy of the data. For data security, two copies of the data will be kept thus doubling this volume.





DATA INTEGRITY

It is not possible to know if any data has been lost from the LTA without looking. Only the very fact of trying and failing to retrieve data will confirm if that data is readable. The act of trying to read it can cause a failure at that point. Therefore the emphasis has to be on taking reasonable steps to minimise the possibility of data loss, to check systematically but not excessively so that normal operations are affected and to have more than one copy of the data.

As the data is archived, there is an integrity check in the form of a checksum (md5) to ensure correct transfer from the processing systems (PDGS). Once archived, various methods are used to ensure the integrity of the data on the tape, mainly using propriety solutions, with the aim to take a recovery action before a problem has occurred, with different levels of detail in the checks and spanning different time periods. If a tape is suspected of losing reliability, the data is copied onto a new tape before the failure. This process is fully under the implementation and control of the industrial partners.

No data has been lost from the LTAs as a whole. Any data loss that has occurred in an individual LTA has been recovered from the backup LTA.

MOVING DATA

The distribution of the data sets throughout the European Union within industry has provided many advantages including the development of capabilities within those industries for handling large data sets and providing Data Archive as a service. However, there are some challenges as well that will have to be met within the coming years. One of the main challenges is moving these comparatively large data sets from current industrial partners to the new industrial partners that will be chosen when required. This will need to be completed in a limited amount of time whilst still continuing to provide unbroken operational services.

As the technology employed is not chosen by ESA, it is likely that outgoing and incoming technologies will be different from each other. This is not an issue per se but may result in industrial partners needing to deal with technologies with which they are not experienced.

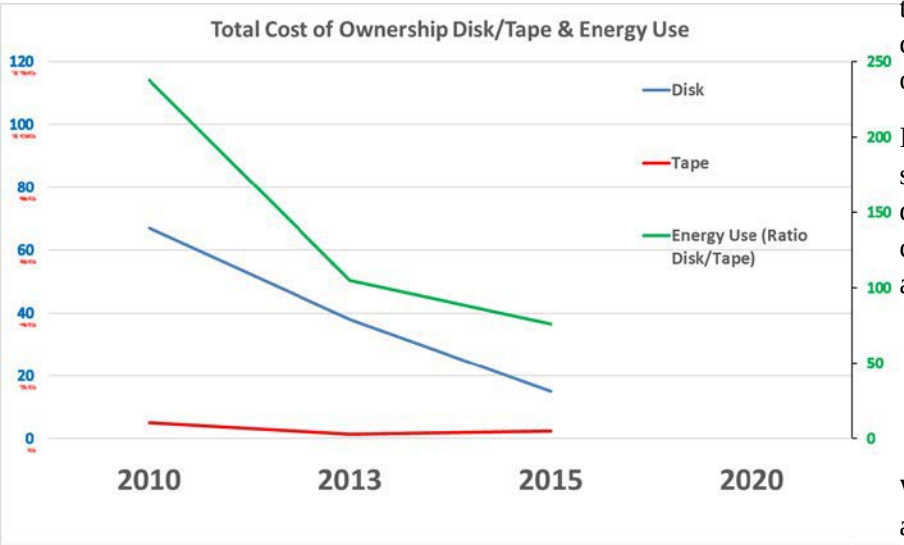
Whilst this challenge is significant, it is one that has to be met anyway. At some point this compaction will have to be faced whether because of technology obsolescence – due to tape drives and tape media versions change – or a simple need to reduce the footprint of the installation. Increasing archive sizes also demand increased throughput offered by newer technologies. Furthermore newer technologies with reduced footprint

may also have associated energy savings in terms of electricity and air-conditioning usage. However, new technology has a cost which needs to be offset against the benefits. This compaction will allow the physical removal of old versions of the products from the archival media. If these old products have been replaced by new products generated from a reprocessing campaign, this could be a significant saving particularly if both versions were intermixed within the same media.

A process has been defined within LTA operations for the deletion of any data from the LTA to guard against accidental deletion of good data. This process requires multiple authorities for the deletion to go ahead. Deletions are currently carried out as the operations progress although it is possible to defer any deletions to the time of compaction.

FUTURE CONSIDERATIONS

The current LTA implementations are a mixture. The smaller data sets are disk only solutions. The larger ones are using tape but with a large disk cache for intermediate storage of the products before streaming to tape. Tape still has a lower Total Cost of Ownership compared to disk only solutions but the costs are converging (Clipper studies). However, the relative energy use is still significantly higher for disk based solutions. This is to be expected where disks are constantly online compared to offline tapes.



It may be possible to envisage a situation where all of Sentinel data is on-line but this requires careful analysis of the advantages and also the risks.

Whilst all Sentinel products still are available on the Data Hubs, analysis of retrievals from the Data Hubs show that 90% of the data retrieved by users is less than two years old. This indicates that a hybrid solution may also be possible where for a certain time period after generation all products are available online thus reducing the load on the tape libraries.
Excessive loads on the tape libraries have not been observed so far but the volume of data being retrieved will increase in volume as more data is archived. Retrieval from tape is relatively slow as retrieval operations are given a lower priority than archive operations.A bigger question is the role of the LTA. The primary aim of the LTA is to ensure safe keeping of the data long term. To meet this role, it can be argued that there should be an emphasis on safe keeping of the basic data – Level-0 and auxiliary – from which all other products can be regenerated if necessary. Access to this data would be severely limited to the case of Disaster Recovery.

There is another role for this “cold backup” LTA – to act as a source of data for reprocessing campaigns. The current LTAs do not make provision for these issues. No policy for the segregation of data was defined and therefore Level-0 data could be intermixed with all other data. This can be rectified at the time of the next compaction of data. Notwithstanding this, the decision would have to be made as to what is the fundamental data that needs to be archived with some elements in the scientific community arguing that this should be Level-1. Clearly this is a decision that impacts the future LTA requirements and hence design.

The overall cost of the LTA operations represents today a large part of the overall cost of the Copernicus

Ground Segment operations. A challenge for the future is to contain the costs of archive operations whilst maintaining efficient operations. This will mean further understanding of the drivers for costs and ways in which they can be minimised. The trade-off between the cost of introducing new technology against the costs of that technology needs to be evaluated. As well at this, there is a need to have a cost model to understand the drivers and to be able to compare costs across the industrial partners.

A 40 Year Review of the Dundee Satellite Receiving Station Earth Observation Data Archive

Neil Lonie, Paul Crawford and Andrew Brooks Dundee Satellite Receiving Station, University of Dundee
email: ntl@sat.dundee.ac.uk, psc@sat.dundee.ac.uk, arb@sat.dundee.ac.uk

Dundee Satellite Receiving Station is one of the main stations in the UK receiving data from Earth Observation satellites. It has been operational since 1975, the primary purpose being to receive, process and disseminate data to UK environmental scientists, and to curate the data for future studies. However, the user base extends far beyond the UK science community with thousands of other users. Major satellite data sets collected include AVHRR, CZCS, SeaWiFS, MODIS and VIIRS.

The focus of this paper is work over the past 4 decades to maintain and preserve the archive of polar orbiting satellite data in particular. Storage and migration of the archive across a range of media is discussed. We highlight examples of significant projects that benefit from the long term management and continued availability of the archive.

The longest continuous data set is AVHRR. Projects that reprocessed this have required conversion of stored formats and curation of satellite orbital elements to deal with gaps and anomalies in Two Line Elements. Additionally, accurate navigation often requires adjustment of satellite attitude parameters and for such a large data set the development of automatic landmark-based registration was needed. This work is also discussed.
Keywords: satellite, data, archive, tape, preservation

Introduction

Dundee Satellite Receiving Station (DSRS) is one of the main stations in the UK receiving data from Earth Observation (EO) satellites. Currently part of the NERC-NCEO funded NEODAAS facility, it originally developed from projects in electronics and communications undertaken during the 1960s and 70s, which was pioneering work at university level. Reception capabilities developed to the extent that they could be provided as a service to support environmental scientists, and so DSRS has received data on a daily basis since 1975. The existing archive extends back to the first digital polar weather satellite in 1978. The Station's primary purpose is to receive, process and disseminate data to UK scientists for studies in meteorology and marine science, for example, and to archive and preserve the data for future studies. However, the user base now extends far beyond the UK science community with thousands of other users globally. In over 40 years of operations major data sets collected from polar orbiting satellites include AVHRR, CZCS, SeaWiFS, MODIS and VIIRS, all providing coverage of Europe, North Atlantic and Arctic regions. In addition, data from geostationary satellites provides users with access to global coverage imagery.

Data Archiving and Preservation

As a facility supporting environmental research, a central element of the operational strategy has been to ensure all data are securely archived and preserved. The significance of this approach is emphasised by the fact ESA and its user communities recently identified the archive as a key long term data record for various ECVs and CCI projects. Here we provide an overview of storage and migration of the archive across the range of media that have been used.

High Density Digital Tape (1978 - 1991)

AVHRR data have been collected at Dundee since 1978 following launch of the TIROS-N satellite. High Density Digital Tape (HDDT) was used to archive data as it was the most appropriate way to capture the live stream received and store relatively high data volumes at that time. These were multi-track tapes that ran on

large reel-to-reel recorders. Tapes were 1 inch wide, 4600 feet long with 14 tracks recorded across the width. Data were recorded serially on a single track with two passes along the length, giving 28 passes per tape. Recording equipment was configured manually for each pass, so data were mainly acquired for day and evening periods when staff were present. Hardcopy photo images were produced for browsing. Data dissemination to users was in the form of photo prints or data transcribed to Computer Compatible Tape delivered by post. Tape playback was needed to recover data each time a scene was requested so physical tape damage and data loss due to repeated playback was a possibility. This represented a risk to the archive as only one copy was made due to high media cost of around £130 per HDDT. Towards the 1990s, problems were experienced when recovering data from some tapes. It was an industry wide issue caused by moisture absorption by adhesive binder between the magnetic oxide and tape. The surface of affected tapes would become progressively stickier, making playback and data recovery increasingly difficult, and heightening the risk of damage and data loss.

VHS Cassette Tape (1991 - 1997)

Given the issues experienced with HDDT, other data storage options were investigated. Digital recording systems based on VHS cassette tapes were selected and used for all new data received from 1991. These could record data rates exceeding those received at the time and again the data was recorded live on tape as it was received. Professional grade E120 cassette tapes were used and allowed up to 14 passes of AVHRR and SeaWiFS data to be recorded on each one. The new systems offered benefits such as error correction information being added to and recorded with incoming data. This provided almost error free playback of data and protection against possible degradation due to tape quality, handling, and archiving. Computer remote control was possible so the system could be set up to automatically record a full cassette of passes without manual intervention. This overcame the limitation of most data having to be received when staff were present so more passes could be acquired overnight, for example. It was also a cost effective solution for archiving compared to the previous approach because of significant reduction in media cost at approximately £5 per VHS cassette. This enabled two recorded copies of each pass to be made for backup purposes for the first time, although second copies were overwritten after a short period once the master copy was verified.

Compact Disk (1997 – 2016)

A range of improvements to the infrastructure during the 1990s benefitted all aspects of operations from satellite tracking to user services. Data were now captured live by dedicated computers and immediately transferred to servers holding a small rolling number of passes. Browse products were created from all new data for website access. Near-real time and archive data could be disseminated by Internet. Data were still recorded on VHS cassette and this continued to be the archive copy until 1997. By then, compact disk technology and media were readily available and affordable, although the price of a CD-R disc was around £5. The decision was taken to begin archiving all new data on CD-R and 8 passes could be stored on a disc. Two copies of each CD-R were produced with the intention to store the second copy off-site to guard against a disaster situation. We continued to record new data on VHS cassettes until verified on CD-R.

Archiving on CD-R provided an opportunity to address ageing and degrading HDDT tapes and consolidate the entire archive to one media type. The decision was taken to transcribe all tape based archive data, HDDTs and VHS, to disc. This was a significant project involving around 35,000 passes, each restored from tape one at a time. A major effort was made to prepare HDDTs for optimum data recovery by two methods. Initially, HDDTs were cleaned by running them over an alcohol impregnated swab on an adapted recorder to reduce stickiness and remove impurities. Later in the project, a technique of baking tapes was used to drive out moisture before cleaning. The transfer of historical data to CD-R was completed in 2002 and off-site backup copies of all discs delivered. Imagery of the entire archive was also produced so that it could be remotely viewed and searched for the first time via the website.

Data Cartridges (1999 – 2018)

In 1999 DSRS started to receive higher rate data than previously, specifically MODIS data. The storage capacity provided by CD-R was insufficient so DLT tape was selected for archiving. DLT IV tapes had a capacity of 35 GB, had been standardised five years previously, and fast/wide SCSI drives had been available for three years. LTO tape drives were not available at this time. The drives were self-cleaning in that a LED indicated when cleaning was required and a cleaning operation simply involved inserting a cleaning tape.

Initially 36 passes of MODIS data were stored per DLT, whilst AVHRR continued to be archived on CD-ROM. The tape drives were expensive so only one was purchased but archive policy required two copies so identical tapes were exchanged daily. Custom software wrote an identification header on the tape to prevent accidental use of an incorrect tape. A randomly-selected track was restored after archiving to check for errors. Restoring a track did occasionally reveal a problem which required the effected passes to be written again. The problem was only ever on one tape so the data could be recovered from the second tape copy if no longer on disk. However the policy of identical tapes meant that both had to be updated when re-archiving a pass. There was no cause identified for the badly archived pass, neither the tape nor the drive were systematically at fault.Most problems that occurred were due to the tape drive failing. Often the failure would cause the tape to be stuck irretrievably inside the drive so a fresh copy had to be made from the other tape. One particularly frustrating 'feature' was drives which would lose the End-Of-Tape (EOT) marker if they were power cycled whilst a tape was in the drive at a position other than fully rewound, as this meant the tape could no longer have any data appended so had to be completely re-written. Having a tape drive on support was beneficial in the early years but proved troublesome later on. The drive was required for archive restore operations after the last ship date and was fully supported by Premier/Platinum Support but when spares were required the company were unable to supply them as there were none left in their global inventory.

A higher capacity Super-DLT drive was purchased specifically to supply data for a customer but to maintain backwards compatibility the drive was not used routinely for archiving. The next change to archiving medium came in 2005 when a LTO-3 drive was purchased. LTO appeared to have better future prospects than DLT so it was considered worthwhile to migrate the archive from DLT. In hindsight this was the correct decision. LTO has maintained a certain amount of backwards compatibility as LTO-(x) drives can write to LTO-(x-1) tapes and read from LTO-(x-2) tapes, although this capability will be more restricted when LTO-8 is released. This allowed the Station to maintain an archive on LTO-3 tapes knowing they could be read on future drives when LTO-3 drives failed and were replaced. The huge increase in capacity with LTO-3 meant that the policy of a fixed 36 tracks per tape was abandoned in favour of maximising their usable capacity. Typically about 700 passes could be stored on one tape.

The large tape capacity of LTO brought with it some concerns. One concern was the amount of data that would be lost if a tape failed. A second was the amount of time taken to fill it would result in a large number of tape load/unload cycles, a parameter which was limited to the low thousands in LTO-3. A third was the difficulty in creating a fresh copy from the second tape if one failed.

Every time there was a change in hardware, firmware or software a thorough set of tests had to be performed using vendor drive tools and custom software. These tests uncovered problems in almost all aspects, including: drives not operating to specification; SAS adapters giving subtle errors or silently writing bad data; a selection of kernel device drivers which behave differently; inconsistent support for drive features such as density; operating system differences such as reporting of tape position, errors due to file markers, and so on.

Disk Storage (2003 to date)

Our first home-built NAS supported 11TB of data spread over 3 servers. At the time it exceeded all of the rest of the University's storage capacity but managing the storage over different mount points was an unpleasant task. This system used hardware RAID cards and lacked the advanced data integrity attributes of file systems like ZFS so if an uncorrectable sector error occurred there was no simple way to determine what file(s) had

been corrupted.

A major leap in the affordability of disk storage allowed the Station to purchase a NAS in 2009 providing over 90TB of usable capacity. This was sufficient to store the whole archive currently on tape plus enable growth over 3 to 5 years by adding more disks. Two main benefits were an 'archive' on disk plus the ability to perform instant data processing on any pass in the archive without restoring from tape. As initially configured the system had enough redundancy to be considered archival but of course at least two tape copies continue to be made.

With the increase in external network bandwidth and a low-cost cloud storage contract the Station started archiving data to the cloud provider Box in 2015. This provides an additional copy which can be accessed at relatively high speed. It currently consumes over 22TB.

We continue to archive all data in a raw format, close to that received from the satellite. The raw format is often encoded or internally compressed which prevents the use of compression within the tape drive so we only obtain the native tape capacity even though the compressed capacity is more often quoted in advertising. MODIS data is archived at level-0 which has been error-corrected using software that previously took 15 minutes. With the increase in CPU speed it would, in hindsight, have been better to archive raw and perform error-correction when restoring from tape, as the error correction codes stored with the data could have been used to recover from tape errors. The theoretical benefit of archiving higher product levels, eg. level-1, is quicker access to useful data after a tape restore, but in practice these higher levels always have to be converted back to level-0 and reprocessed due to improvements in calibration and navigation software and ancillary data. There is also the risk involved in any delay due to data processing between the reception of the data and it being archived onto tape.

Ancillary data for reprocessing

The successful processing, or re-processing, of EO satellite data requires reasonably accurate navigation of the sensor to allow the ground based latitude and longitude of each pixel to be computed so that data may be references to standard maps, etc. This process typically requires the following inputs:

- Orbital elements to determine the satellite's position as a function of time
- Spacecraft Attitude to allow orientation of the sensor to be determined
- Time of each pixel to perform the computation using the above

Each of these has problems with historic data sets and it is sometimes necessary to apply corrections to the data sets. For the orbital elements we now make use of the Two Line Elements (TLE) that are distributed by the US Air Force for most (non-classified) objects in orbit around the Earth.

When attempting to use the old TLE there are numerous issues: some are unwanted additions where data for another satellite has erroneously been added to the identifier for a satellite of interest (e.g. NOAA-10 has spurious element inserted in year 1992 day 260.87302325 with a GPS-like orbit) but more often the problem is the lack of data for days at a time (e.g. NOAA-6 where there is a gap of 27.841 days at the start of 1981 thought to be due to lost/damaged tapes) leading to high propagation errors. Even when elements are more or less "nominal" they can occasionally be ill-fitted and this leads to serious errors after a short time (e.g. NOAA-18 for year 2018 day 114.832462).

To address both of these issues, bad TLE and gaps in good TLE, we wrote some software to compare forwards and backwards propagation errors. For the initial results we could identify and edit out obviously bad data points manually. After that was done the software would adjust the orbital model's drag factor in an attempt to reduce propagation errors. Largely this is a great success, with peak errors being reduced by almost an order of magnitude in tests where we removed known good elements in order to compare the performance at those

removed times.

The issue of ‘time’ for the NOAA spacecraft data has two specific issues: the time code in the telemetry lacks the year; and the clock in kept nominally to +/-1 second, but very occasionally is set incorrectly.

While some reception sites had the ability to time-stamp the received data, we did not and the majority of other reception site did not. Hence processing older data needs some other means of determining the year, and also occasionally correcting for major clock errors. Our approach is to use existing knowledge of the data file from metadata held in our archive, or from a guess for data in other archives, and to compare the computed visibility of the satellite with the data’s time-range. This has proven to be a fairly robust approach.

Finally we have the issue of accurate geolocation of the data sets, and in this case the 1 second clock error and typical propagation errors for the TLE can lead to errors of the order of a few to tens of km. In addition there was never any accurate attitude data, various attempts to read out the on-board gyro information failed, or simply retuned values indicating the spa craft was keeping pointing to near zero gyro error.

With tens of thousands of files to re-process the use of manual adjustment was simply infeasible we developed an image-based correction system to solve for a robust least-squares fit between the images and landmarks automatically generated from the World Vector Shoreline database. That system was supplied commercially by SCISYS in Germany for many other AVHRR users to provide rapid real-time image navigation.

Conclusion

When managing data sets over long periods of time the technology involved will change almost beyond recognition and no physical storage format will remain supportable. Hence there is an ever present need to migrate data from storage system to system, but it is not only the data that needs to be maintained, also the metadata and know-how of using the data needs preservation.

The question of whether to archive on tape, on disk or in the cloud in future remains unresolved due to the balance between cost, speed of access, and the unknown quantity of longevity, something that applies equally to cloud companies as it does to physical media. Out-sourcing to the cloud delegates physical maintenance to others but arguably introduces greater non-physical risks such as ongoing costs, trust and long-term viability of their business model.

Virtual European Solar & Planetary Access (VESPA): a Virtual Observatory in Planetary Science.

S. Erard[1], B. Cecconi[1], P. Le Sidaner[2], A. P. Rossi[3], T. Capria[4], B. Schmitt[5], V. Génot[6], N. André[6], J.-M. Glorian[6], A. C. Vandaele[7], M. Scherf[8], R. Hueso[9], A. Määttänen[10], B. Carry[11], N. Achilleos[12], C. Marmo[13], O. Santolik[14], J. Soucek[14], K. Benson[12], P. Fernique[15]

[1]LESIA, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Univ. Paris Diderot, Sorbonne Paris Cité, 5 place Jules Janssen, 92195 Meudon, France [2]DIO-VO/UMS2201 Observatoire de Paris/CNRS, Fr, [3]Jacobs University, Bremen, Ge [4]INAF/IAPS, Rome, It [5]IPAG UGA/CNRS, Grenoble, Fr [6]IRAP/CNRS, Toulouse, Fr [7]IASB/BIRA, Brussels, Be [8]OeAW, Graz, Aut [9]UPV/EHU, Bilbao, Sp [10]LATMOS/CNRS, Guyancourt, Fr [11]OCA, Nice & IMCCE/Obs. Paris/CNRS, Fr [12]University College London, UK [13]GEOPS/CNRS/U. Paris-Sud, Fr [14]IAP, Prague, Cz R. [15]Observatoire de Strasbourg/UMR 7550, Fr

The Europlanet2020 program, started Sept 1st, 2015 for 4 years, includes an activity called VESPA which fouses on adapting Virtual Observatory (VO) techniques to handle Planetary Science data. The objective of VESPA is to facilitate searches in big archives as well as sparse databases, to provide simple data access and on-line visualization, and to allow small data providers to make their data available in an interoperable environment with minimum effort. This system relies in particular on standards and tools developed for the Astronomy VO (IVOA) and enlarges them where required to handle specificities of Solar System studies.

Introduction: Modern space borne instruments produce huge datasets, especially on long-lived missions. This calls for new ways to handle the data, not only to perform mass processing, but also more basically to access them easily and efficiently. Virtual Observatory (VO) techniques developed in Astronomy during the past 15 years can be adapted to address this problem, provided they are enlarged to handle specificities of Solar System studies. The VESPA data access system focuses on applying VO techniques and tools to Planetary Science data, and supports all aspects of Solar System science (Erard et al. 2018). VESPA (Virtual European Solar and Planetary Access) is developed in the framework of the EU-funded Europlanet-2020 program started Sept 1st, 2015 for 4 years. The objective of this activity is to facilitate searches in big archives as well as in sparse databases, to provide simple data access and on-line visualization tools, and to allow small data providers to make their data available in an interoperable environment with minimum effort. This system makes intensive use of studies and developments led in Astronomy (International Virtual Observatory Alliance, IVOA), Solar Physics (HELIO), and space data archive (International Planetary Data Alliance, IPDA).

Data services: the VESPA architecture (Erard et al. 2018) is based on a new data access protocol, a specific user interface to query the available services, and intensive usage of tools and standards developed for the Astronomy VO (Erard et al. 2014a). The Europlanet data access protocol, EPN-TAP, relies on the general TAP (Table Access Protocol) mechanism associated to a set of parameters that describe the content of a data service (Erard et al. 2014b). These parameters are defined to enable queries on quantities relevant to the scientific user, in particular observational and instrumental conditions. Data services are required to return the metadata of matching results formatted as VOtables, which are handled by all standard VO tools.

Data services are installed at their respective provider institutes and are declared in the standard IVOA registries, so that they are always visible and reachable from query interfaces. At the time of writting, 39 data services are publicly open, and about 15 more are being finalized. They encompass a wide scope, including surfaces, atmospheres, magnetospheres and planetary plasmas, small bodies, experimental data such as spectroscopy in solid phase, heliophysics, and exoplanets. VESPA focuses mostly on derived data, typically associated to publications. To favor the emergence of this kind of material, VESPA organizes a yearly call to the community to select projects of interest; 4 or 5 selected teams are invited to a 1 week workshop to design and install the service in their institute. Some large data archives are also targeted; in particular, ESA's Planetary Science Archive (PSA) will get an EPN-TAP interface in 2018, and bridges with PDS4 are being studied. Several amateur data services were selected at the onset of the program for implementation in research

institutes, including PVOL in Bilbao (planetary images) and RadioJove (Jupiter radio measurements) at Paris Observatory. Finally, a special type of services will gather tables of VOevents produced by alert systems in various fields (Gangloff et al. 2018).

Data access: Although accessible in many ways, EPN-TAP data services are best queried from an optimized user interface, the VESPA portal (Fig. 1).

In the frame of TAP, all data services present a list of granules (usually data files) described by a series of parameters. The Europlanet data access protocol, EPN-TAP, defines a set of mandatory parameters introducing metadata that describe all granules; this is similar to the ObsTAP protocol from IVOA, which describes observational datasets in Astronomy. EPN-TAP metadata introduce both observational and instrumental conditions and are defined to handle the specific diversity and complexity of Planetary Science: ranges on several axes (spatial, temporal, spectral, photometric), measurement type, origin of data, and various references. Location is provided in the most appropriate coordinate system (e.g., sky or planetary coordinates); target-related time (local time and season, through Ls) can be provided when relevant. The VESPA portal uses the mandatory parameters to search for individual granules in all registered data services at once, allowing for discovery of data content unknown to the user. In addition, specific parameters may also be used to describe individual services in more details, and can be used to identify granules more precisely when querying a single service.

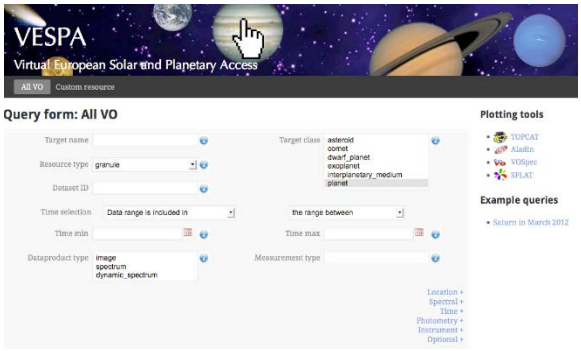


Fig. 1: The VESPA search interface: http://vespa.obspm.fr

An EPN-TAP library was also developed and included in several tools (3Dview, CASSIS, and AMDA) for direct queries from these environments. Since EPN-TAP relies on the more general TAP mechanism, individual EPN-TAP data services can also be accessed via standard TAP clients; these include general query interfaces (e.g., TAPhandle, TAPsh) as well as standard tools (e.g., TOPCAT, Aladin, etc).

All granules provide a link to a data file, or include the data itself in the table when possible (e.g., for scalar quantities). Data description parameters are used to identify adequate VO tools to access, plot and handle the data correctly. They not only provide a description of the file format, but also specify the dimensions, units, and physical quantities, relying on IVOA data models extended for VESPA. For instance, spectra and images are handled in different tools, and spectra measured in radiance or in reflectance are handled differently by the spectral tools.

Tools: Metadata are smoothly transferred from the VESPA portal to VO tools according to the IVOA SAMP protocol. Standard VO tools are connected to the VESPA portal so that they readily display metadata, e. g., spatial footprints are plotted on a 3D sphere in Aladin or Mizar; other metadata such as local time, instrument modes, etc… can be plotted in 2D or 3D with TOPCAT (Fig. 2).

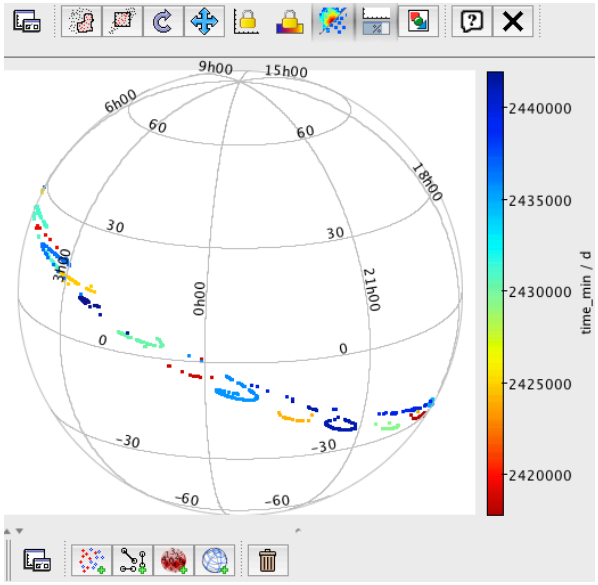


Fig. 2: Location of Mars in the sky from historical telescopic images (1905 to 1976) displayed in TOCAT, evidencing retrograde motion around each opposition.

The data themselves can be transferred in a similar way for display and standard analyses. Data description is used to select appropriate tools, e.g. TOPCAT handles all types of tabular data, Aladin most images and spectral cubes, CASSIS and SPLAT-VO spectra in general, 3Dview can plot measurements along a spacecraft trajectory, Autoplot is dedicated to extracting data from long time series, etc.

Most of these tools have been updated to support Planetary Science and to handle specificities of Solar System data, e.g., measurements in reflected light (Fig. 3), coordinate systems on surfaces and in magnetospheres, etc. Other, non-VO tools have been provided with a SAMP interface so that they can be included in workflows (e. g., ImageJ which now provides conversions for many formats, as well as image processing functions to the VO). In some cases new applications have been developed for VESPA, e. g., to handle georeferenced fits images of planetary surfaces (Marmo et al. 2016, Rosi et al. 2016), or PDS3 spectral cubes (Savalle et al. 2016). Finally, specific web tools developed in support of larger data services are made accessible for use with external data, e.g. AMDA for planetary plasmas at CDDP, or the new SSHADE service for lab spectroscopy in IRAP (Schmitt et al. 2015).

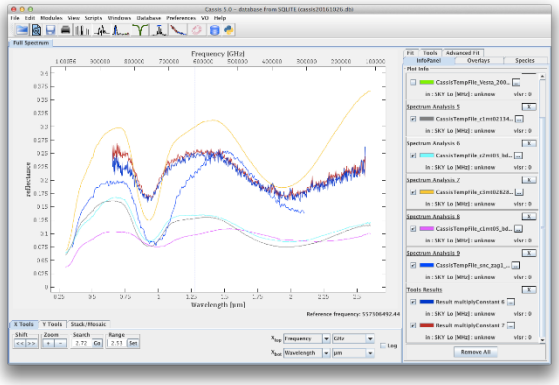


Fig. 3: NIR telescopic spectra of 4 Vesta compared to basaltic meteorites from the PDS spectral library in CASSIS.

TOPCAT can easily integrate sparse surface observations (e.g. from a point spectrometer) using the healpix tesselation system, while Aladin can produce multiresolution maps (HiPS) from large datasets, which allow for very fast change of scale in the client. Currently, 45 planetary maps from USGS have been converted to HiPS and are available from the Aladin data tree. The same technique applied to large panoramas from planetary landers provides a very exciting way to navigate such images, by smoothly changing from the global picture to the highest local details.

A significant on-going activity is the development of a connection between the VO world and Geographic Information Systems (GIS). In a first step, EPN-TAP services are used to provide links as queries to WMS or similar services, i.e. using different, non-VO, access protocols. Tradionally, such links are only handled in GIS applications such as the open source QGIS. While the intermediate VO layer allows for powerful search functions in the data, cross-examinations with other datasets is difficult because of the variety of query systems and image formats. In a second step, the goal is therefore to provide bridges between these two worlds, so that VO (e.g., fits) and GIS (e.g., geotiff) images can be displayed in all applications (Fig. 4). This is done by providing improved georeferentiation support in fits headers and conversion routines in GDAL (Marmo et al. 2016), and with new QGIS plug-ins to add SAMP connnexion and to handle georeferenced fits images directly.

A similar situation applies to time series depicting radio emission of the planets. A protocol of choice in this case is das2server that allows the distribution of data with adjustable temporal resolution. Data services are responsive to EPN-TAP but provide requests to such servers, the results of which can be fetched to the Autoplot tool for display (Cecconi et al. 2018).

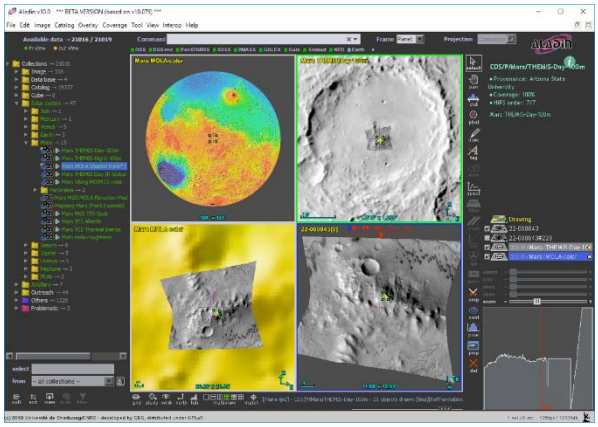


Fig. 4: CRISM spectral cube (georeferenced version converted to geofits) overplotted on MOLA and THEMIS multiresolution maps in Aladin.

As far as spatial data are concerned, VESPA makes use of two IVOA protocols to handle footprints. The first one is the pgsphere s_region standard (used in particular in ObsTAP services) which provides oriented contours; the second one is the Multi-Order Coverage (MOC, healpix based) used e.g. in Aladin, TOPCAT, and Mizar. Both standards can be used to issue powerful searches on intersections or inclusions, and to select objects within arbitrary footprints (Fig. 5).

Fig. 5: Selection of CRISM spectral cubes located in large Martian craters, based on MOC in Aladin (over THEMIS HiPS).

Simulation services: another important goal is to connect on-line computation services with interface similar to that of data services, so as to compare observations and simulations more routinely. This activity has obvious applications, e. g., for radiative transfer in planetary atmospheres or for magnetospheres, but also to connect ephemeris systems (e.g. Miriade) with data services. A most promising solution is to use the datalink protocol of IVOA to call remote services with parameters retrieved from existing data services.

An additional aspect is to provide low level computation functions on-line, e.g., averages, resampling, deconvolution, etc of actual data. This is currently supported only to some extent by standard VO tools and ImageJ; in addition, higher level processing such as retrieval of Hapke parameters from surface spectra, multivariate analyses, etc, would also be beneficial and are being studied.

Building a community: Hands-on sessions are organized twice a year at EGU and EPSC conferences in Europe to support new users (see VESPA web site), as well as contributions to similar workshops in Astronomy. Besides, a procedure has been identified to install data services with little resources, which is expected to foster the installation of data services by individual research teams, e. g. to distribute derived data related to a published study. In complement, regular discussions are held with big data providers, starting with space agencies in the frame of the IPDA. In parallel, a Solar System Interest Group has been started in the IVOA in 2017, where several VESPA partners contribute; the goal here is to adapt existing astronomy standards to Planetary Science.

Acknowledgements:

The Europlanet 2020 Research Infrastructure project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654208.

Support from Paris Astronomical Data Centre (PADC) is acknowledged.

VESPA web site: http://www.europlanet-vespa.eu

References:

Cecconi B., Le Sidaner P., Savalle R., Bonnin X., Zarka P., Louis C., Coffre A., Aicardi S., Lamy L., Denis L., Grießmeier J.-M., Faden J., Piker C., André N., Génot V., Erard S., Mafi J. N., King T. A., Sharlow M., Sky J., Demleitner M. 2018 MASER: A Toolbox for Low Frequency Radio Astronomy, *this conference.*

Erard S, Cecconi B, Le Sidaner P, Rossi AP, Capria MT, Schmitt B, Génot V, André N, Vandaele AC, Scherf M, Hueso R, Määttänen A, Thuillot W, Carry B, Achilleos N, Marmo C, Santolik O, Benson K, Fernique P, Beigbeder L, Millour E, Rousseau B, Andrieu F, Chauvin C, Minin M, Ivanoski S, Longobardo A, Bollard P, Albert D, Gangloff M, Jourdane N,

Bouchemit M, Glorian J, Trompet L, Al-Ubaidi T, Juaristi J, Desmars J, Guio P, Delaa O, Lagain A, Soucek J, and Pisa D 2018 VESPA: A community-driven Virtual Observatory in Planetary Science. *Planet. Space Sc.*150, 65-85. ArXiv 1705.09727

Erard, S., Cecconi, B., Le Sidaner, P., Berthier, J., Henry, F., Chauvin, C., André, N., Génot, V., Jacquey, C., Gangloff, M., Bourrel, N., Schmitt, B., Capria, M. T., and Chanteur, G. 2014a. Planetary Science Virtual Observatory architecture. *Astron. & Comput.* **7-8**, 71-80. ArXiv 1407.4886

Erard, S., Cecconi, B., Le Sidaner, P., Berthier, J., Henry, F., Molinaro, M., Giardino, M., Bourrel, N., André, N., Gangloff, M., Jacquey, C., and Topf, F. 2014b The EPN-TAP protocol for the Planetary Science Virtual Observatory *Astron & Comput* **7-8**, 52-61. ArXiv 1407.5738

Gangloff M., André N., Génot V., Cecconi B., and Le Sidaner P. 2018 A Space Weather VOEvent service provided by the CDPP in the frame of Europlanet H2020 PSWS, *this conference.*

Marmo C, Hare TM, Erard S, Cecconi B, Costard F, Schmidt F, and Rossi AP 2016 FITS Format for Planetary Surfaces: Bridging the Gap Between FITS World Coordinate Systems and Geographical Information Systems. *Lunar and Planetary Science Conference* 47, 1870.

Rossi A. P., Hare T., Baumann P., Misev D., Marmo C., Erard S., Cecconi B., and Marco Figuera R. 2016 Planetary Coordinate Reference Systems for OGC Web Services. *Lunar and Planetary Science Conference* 47 1422.

Savalle R., Erard S., and Le Sidaner P. 2016 APERICubes: an on-line Astronomical and Planetary Ergonomic Research Interface for spectral Cubes. *ADASS* XXVI , abstract 31316.

Schmitt B, Albert D, Bollard P, Bonal L, Gorbacheva M, Mercier L, and Consortium Partners S 2015 SSHADE in H2020: Development of an European Database Infrastructure in Solid Spectroscopy. *European Planetary Science Congress* 2015, 628.

Virtual Planetary Space Weather Services offered by the Europlanet H2020 Research Infrastructure

N. André[1], M. Grande[2], N. Achilleos[3], M. Barthélémy[4], M. Bouchemit[1], K. Benson[3], P.-L. Blelly[1], E. Budnik[1], S. Caussarieu[5], B. Cecconi[6], T. Cook[2], V. Génot[1], P. Guio[3], A. Goutenoir[1], B. Grison[7], R. Hueso[8], M. Indurain[1], G. H. Jones[9,10], J. Lilensten[4], A. Marchaudon[1], D. Matthiä[11], A. Opitz[12], A. Rouillard[1], I. Stanislawska[13], J. Soucek[7], C. Tao[14], L. Tomasik[13], J. Vaubaillon[6]

[1]Institut de Recherche en Astrophysique et Planétologie, CNRS, Université Paul Sabatier, Toulouse, France ; [2]Department of Physics, Aberystwyth University, Wales, UK ; [3]Department of Physics and Astronomy; [3]University College London, London, UK ; [4]Institut de Planétologie et d'Astrophysique de Grenoble, UGA/CNRS-INSU, Grenoble, France ; [5]GFI Informatique, Toulouse, France ; [6]LESIA, Observatoire de Paris, CNRS, UPMC, University Paris Diderot, Meudon, France ;[7]Institute of Atmospheric Physics, Czech Academy of Science, Prague, Czech Republic ;[8]Física Aplicada I, Escuela de Ingeniería de Bilbao, Universidad del País Vasco, Bilbao, Spain ;[9]Mullard Space Science Laboratory, University College London (UCL), Holmbury Saint Mary, UK ;[10]The Centre for Planetary Sciences at UCL/Birkbeck, London, UK ;[11]German Aerospace Center, Institute of Aerospace Medicine, Linder Höhe, 51147 Cologne, Germany; [12]Wigner Research Centre for Physics, Budapest, Hungary;[13]Space Research Centre, Polish Academy of Sciences, Warsaw, Poland; [14]National Institute of Information and Communications Technology 4-2-1, Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan

Corresponding author: Nicolas André (nicolas.andre@irap.omp.eu)

Under Horizon 2020, the Europlanet 2020 Research Infrastructure (EPN2020-RI) will include an entirely new Virtual Access Service, "Planetary Space Weather Services" (PSWS) that will extend the concepts of space weather and space situational awareness to other planets in our Solar System and in particular to spacecraft that voyage through it. PSWS will make twelve new services accessible to the research community, space agencies, and industrial partners planning for space missions. These services will in particular be dedicated to the following key planetary environments: Mars (in support of the NASA MAVEN and ESA Mars Express and ExoMars missions), comets (building on the outstanding success of the ESA Rosetta mission), and outer planets (in preparation for the ESA JUpiter ICy moon Explorer mission), and one of these services will aim at predicting and detecting planetary events like meteor showers and impacts in the Solar System. This will give the European planetary science community new methods, interfaces, functionalities and/or plugins dedicated to planetary space weather as well as to space situational awareness in the tools and models available within the partner institutes. A variety of tools (in the form of web applications, standalone software, or numerical models in various degrees of implementation) are available for tracing propagation of planetary and/or solar events through the Solar System and modelling the response of the planetary environment (surfaces, atmospheres, ionospheres, and magnetospheres) to those events. But these tools were not originally designed for planetary event prediction and space weather applications. PSWS will provide the additional research and tailoring required to apply them for these purposes. PSWS will be to review, test, improve and adapt methods and tools available within the partner institutes in order to make prototype planetary event and space weather services operational in Europe at the end of 2018. To achieve its objectives PSWS will use a few tools and standards developed for the Astronomy Virtual Observatory (VO). This paper gives an overview of the project together with a few illustrations of prototype services based on VO standards and protocols.

Keywords: Planetary space weather, Tools, Interoperability

1. Introduction

Space Weather – the monitoring and prediction of disturbances in our near-space environment and how they are controlled by the Sun - is now recognised as an important aspect of understanding our Earth and protecting vital assets such as orbiting satellites and power grids. The Europlanet 2020 Research Infrastructure (http://www.europlanet-2020-ri.eu/) aims to transform the science of space weather, by extending its scope throughout the Solar System. An entirely new Virtual Access Service, "Planetary Space Weather Services" (PSWS, http://planetaryspaceweather-europlanet.irap.omp.eu/) has therefore been included in the Europlanet H2020 Research Infrastructure funded by the European Union Framework Programme for Research and Innovation.

Planetary Space Weather Services (PSWS) aims at extending the concept of space weather to other planets in our Solar System and in particular to spacecraft that voyage through it. PSWS will give the European planetary scientists for the first time new methods, interfaces, functionalities and/or plug-ins dedicated to planetary space weather in the form of tools and models available in the partner institutes.

The Planetary Space Weather Services will provide at the end of the Europlanet 2020 Research Infrastructure programme 12 services distributed over 4 different service domains – Prediction, Detection, Modelling, Alerts - having each its specific groups of end users. The PSWS portal (http://planetaryspaceweather-europlanet.irap.omp.eu/) gives access to an initial presentation of PSWS activities. Section 2 gives an overview of the services and their status. Each service will be implemented through a combination of data products, software tools, and tutorials.

2. PSWS Services

The Planetary Space Weather Services will provide 12 services distributed over 4 different service domains – Prediction, Detection, Modelling, Alerts. These services and their status in mid-2018 are detailed in this section.

2.1. Prediction

2.1.1 The Heliopropa service

The *Centre de Données de Physique des Plasmas* (CDPP) within the *Institut de Recherche en Astrophysique et Planétologie* (IRAP/CNRS) provide real time and archive access to propagated solar wind parameters at various planetary bodies (Mercury, Venus, Mars, Jupiter, Saturn, ,…) and spacecraft (Rosetta, Juno, Maven,…) using a 1D magnetohydrodynamic (MHD) code available through the CDPP/AMDA tool (http://amda.cdpp.eu) initially developed by Chihiro Tao [1]. A dedicated tool and interface is operational at http://heliopropa.irap.omp.eu

2.1.2 Extensions of the CDPP Propagation Tool

The *GFI Inf ormatique* (GFI) has extended the Propagation tool [2] available at CDPP (http://propagationtool.cdpp.eu) to the case of comets, giant planet auroral emissions, and catalogues of solar wind disturbances such as the ones defined by the FP7 HELCATS project (http://helcat-fp7.eu).

2.1.3 Meteor showers

The *Observatoire de Paris* (OBSPARIS) will link ephemerides of Solar System objects to predictable meteor showers that impact terrestrial planet surfaces or giant planet atmosphereCometary tail crossings

The *Mullard Space Science Laboratory* (MSSL) within the *University College London* (UCL) will develop and post online a software in order to enable users to predict cometary ion tail crossings by any interplanetary spacecraft including future missions like Solar Orbiter, BepiColombo, and JUICE. The tool and its current status can be found at https://www.ucl.ac.uk/mssl/planetary-science/tailcatcher

2.2 Detection

2.2.1 Lunar impacts

Aberystwyth University (ABER) will upgrade and convert its lunar impact software (https://www.britastro.org/lunar/tlp.htm) and post it online in order to enable users to detect visible flashes in lunar amateur or professional images.

2.2.2 Giant planet fireballs

The *Universidad del Pais Vasco* (UPV/EHU) has upgraded and converted its giant planet fireball detection software (http://pvol2.ehu.eus/psws/jovian_impacts/) and posted it online in order to enable users to detect visible fireballs in giant planet amateur or professional images.

2.2.3.Cometary ion tails

Mullard Space Science Laboratory (MSSL) within *University College London* (UCL) will upgrade and convert its cometary ion tail analysis software and post it online, with the aim of also providing it as an interactive suite. The software will be readily accessible to any users (professional or amateur) who work with comet images and wish to obtain an estimate for the solar wind speed at the comet from their observations. The tool and its current status can be found at https://www.ucl.ac.uk/mssl/planetary-science/Solar_Windsocks

2.3. Modelling

2.3.1 Transplanet – Venus, Earth, Mars, Jupiter

The *Centre de Données de Physique des Plasmas* (CDPP) within the *Institut de Recherche en Astrophysique et Planétologie* (IRAP/CNRS) has developed an online version of the hybrid-fluid TRANSPLANET ionospheric model [5] that will enable users to make runs on request for Venus, Earth, Mars, and Jupiter. Particle precipitation corresponding to particular solar wind conditions can be set by the user. The service is operational and can be accessed at http://transplanet.irap.omp.eu

2.3.2 Mars Radiation Environment

Aberystwyth University (ABER) together with the Institute of Aerospace Medicine (DLR Cologne) will develop a Mars radiation surface environment model. The service will enable in particular estimates of radiation doses in the atmosphere (e.g., for orbiters) and at the surface of the planet (e.g., for rovers like the one of the Exomars mission or astronauts). The prototype service and its status can be found at http://radmaree.irap.omp.eu

2.3.3 Giant planet magnetodiscs

University College London (UCL) will adapt the parametric magnetodisc model for Jupiter and Saturn as well as the resulting magnetic field mapping in their ionospheres in order to take into account realistic, rapid solar wind compressions [7], based on time-dependent predictions of dynamic pressure from the CDPP Propagation Tool and/or observations of solar wind at Jupiter orbit. The prototype service and its status can be found at http://magnetodisc.irap.omp.eu

2.3.4 Jupiter's thermosphere

University College London (UCL) will adapt the 2D thermospheric models available for Jupiter and its space environment in order to take into account realistic, rapid solar wind compressions [8], based on time-dependent predictions of dynamic pressure from the CDPP Propagation Tool and/or observations of solar wind at Jupiter orbit.

2.4 Alerts

OBSPARIS together with UCL, IRAP/CNRS, and the *Space Research Center* (PAS/SRC) will create an Alert service linked to prediction of planetary events of various kinds: solar energetic particles (SEP), solar wind disturbances triggering to magnetospheric or auroral events, planetary meteor showers, cometary tail disconnection events, lunar flashes, giant planet fireballs, radio type III. We propose to broadcast these events with VOEvent, an alert service infrastructure developed in the frame of IVOA. The prototype service at IRAP/CNRS and its current status can be found at http://alerts-psws.irap.omp.eu/

3Conclusions

The PSWS Services within the Europlanet H2020 Research Infrastructure are developed following protocols and standards available in Astrophysical, Solar Physics and Planetary Science Virtual Observatories (VO) as described in [9].

Acknowledgments

The Europlanet H2020 Research Infrastructure project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654208.

References

[1] Tao, C. et al., Magnetic field variations in the Jovian magnetotail induced by solar wind dynamic

pressure,
Journal of Geophysical Research: Space Physics, 110, doi: 10.1029/2004JA010959, 2005

[2] Rouillard, A. et al., A propagation tool to connect remote-sensing observations with in-situ measurements of heliospheric structures, Planetary and Space Science, 147, p. 61-77, doi: 10.1016/j.pss.2017.07.001, 2017

[3] Lamy, L., Prangé, R., Henry, F., Le Sidaner, P., The Auroral Planetary Imaging and Spectroscopy (APIS) service, Astronomy and Computing, Volume 11, 138-145, 10.1016/j.ascom.2015.01.005, 2015

[4] Davis, C.J. et al., A synoptic view of solar transient evolution in the inner heliosphere using the Heliospheric Imagers on STEREO, Geophys. Res. Lett., 36(2), L02102, doi:10.1029/2008GL036182, 2009

[5] Marchaudon, A., and P.-L. Blelly, A new 16-moment interhemispheric model of the ionosphere : IPIM, J. Geophys. Res., 120, doi:10.1002/2015JA021193, 2015.

[6] Guo, J., et al., A generalized approach to model the spectra and radiation dose rate of solar particle events on the surface of Mars, The Astronomical Journal, 155, 1, doi: 10.3847/1538-3881/aaa085, 2018

[7] Achilleos et al., Influence of hot plasma pressure on the global structure of Saturn's magnetodisk, Geophys.
Res. Lett., 37, L20201, doi: 10.1029/2010GL045159, 2010

[8] Yates, J.N., N. Achilleos, and P. Guio, Response of the Jovian thermosphere to a transient 'pulse' in solar wind pressure, Planet. Space Sci., 91, 27-44, doi:10.1016/j.pss.2013.11.009, 2013

[9] André, N. et al., Virtual Planetary Space Weather Services offered by the Europlanet H2020 Research Infrastructure, Planetary and Space Science, Volume 150, p. 50-59, doi: 10.1016/j.pss.2017.04.020, 2017

EVER-EST: THE PLATFORM ALLOWING SCIENTISTS TO CROSS-FERTILIZE AND CROSS-VALIDATE DATA

Mirko Albani[1], Rosemarie Leone[1], Federica Foglini[2], Francesco De Leo[2], Fulvio Marelli[3], Iolanda Maggio[4]

European Space Agency[1], ISMAR[2], Terradue[3], Rhea Group[4]

Over recent decades large amounts of data about our Planet have become available. If this information could be easily discoverable, accessible and properly exploited, preserved and shared, it would potentially represent a wealth of information for a whole spectrum of stakeholders: from scientists and researchers to the highest level of decision and policy makers. By creating a Virtual Research Environment (VRE) using a service oriented architecture (SOA) tailored to the needs of Earth Science (ES) communities, the EVER-EST (http://ever-est.eu) project provides a range of both generic and domain specific data analysis and management services to support a dynamic approach to collaborative research. EVER-EST provides the means to overcome existing barriers to sharing of Earth Science data and information allowing research teams to discover, access, share and process heterogeneous data, algorithms, results and experiences within and across their communities, including those domains beyond Earth Science.

Keywords: Virtual Research Environment, Remote Sensing, Research Object, Cross-fertilization, Data Analysis, Earth Science, and Education.

Introduction

EVER-EST is funded by the European Commission H2020 programme for three years starting in October 2015. The project is led by the European Space Agency (ESA) and involves some of the major European Earth Science data providers/users including NERC, DLR, INGV, CNR and SatCEN.

The paper presents specific aspects of this collaboration platform in terms of infrastructure and implemented paradigms. Some case studies on cross-fertilization analysis are documented in order to show the process for creating knowledge and new data starting from collected data from different sources (e.g. from remote and social sensing). The paper concludes with few future outcomes.

EVEREST PLATFORM

EVER-EST is a research and development platform that offers a framework based on advanced services to support each phase of the Earth Science Research and Information Lifecycle. The project follows a user-centric approach which have produced a wealth of innovative and state-of-the-art technologies, systems and tools for e- collaboration, e-learning, e-research and long term data preservation.

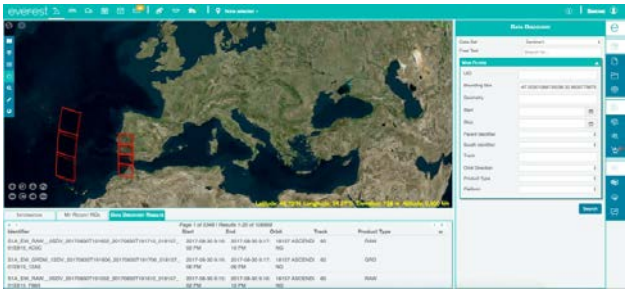


Figure 1 - EVER-EST user interface

It provides innovative services to Earth Science user communities for the generation, communication, cross-validation and sharing of knowledge and science outputs.

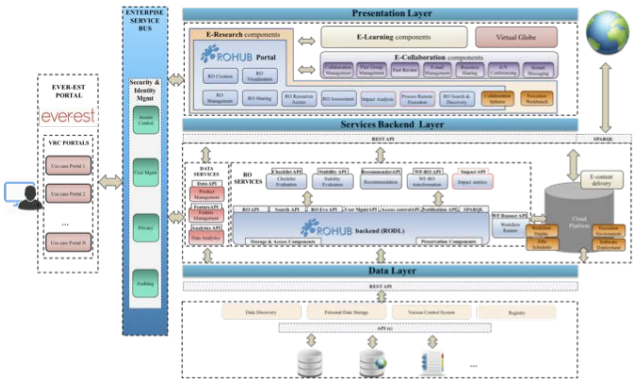


Figure 2 - EVER-EST overall architecture

The architecture reflects the organization of functions across the various layers and with regards to the core element of the design (on the right of the Service Bus vertical component) the following can be identified:

1. Presentation Layer - The EVER-EST VRE offers a web application that includes both the Virtual Research Community portals (Figure 2 – left side) and the ROHub portal (Figure 2 – upper side), the graphic user interfaces to provide the client-side implementation of e-collaboration, e-learning and e-research services along with the mechanisms for Earth Science RO creation.
2. Service Layer - in the central part of the architectural diagram - provides both generic VRE services and Earth Science specific services. These components represent the reasoning engine of the e-infrastructure and actually orchestrate and manage the services available to the VRE final users. The service layer includes the server-side implementation of e-research and processing services, including cloud resources to instantiate virtual machine properly configured according to VRC specifications.
3. Data Layer - bottom part of the design - references the data holdings made available to the VRCs: data is linked and proper means are provided, where feasible, to access it from the VRE. As a default setting, data will not be copied or duplicated, but will continue to reside on the provider's local servers unless it is directly retrieved by the user.

VIRTUAL RESEARCH ENVIRONMENT

The EVER-EST e-infrastructure is validated by four virtual research communities (VRC) covering different multidisciplinary Earth Science domains including: ocean monitoring, natural hazards, land monitoring and risk management (volcanoes and seismicity).

- **Land Monitoring**: Monitoring of urban, built-up and natural environments to identify certain features or changes over areas of interest.
- **Sea Monitoring**: The Sea Monitoring VRC focuses on finding new ways to measure the quality of the maritime environment and it is quite wide and heterogeneous, consisting of multi- disciplinary scientists such as biologists, geologists, oceanographers and GIS experts, as well as agencies and authorities (e.g. ARPA or the Italian Ministry of Environment). The scientific community has the main role of assessing the best criteria and indicators for defining the Good Environmental Status descriptors defined by the Marine Strategy Framework Directive (MSFD).
- **Geohazard Supersites and Natural Laboratories:** is a collaborative initiative supported by GEO (Group on Earth Observations) within the Disasters Resilience Benefit Area. The goal of GSNL is to facilitate a global collaboration between Geohazard monitoring agencies, satellite data providers and the Geohazard scientific community to improve scientific understanding of the processes causing geological disasters and better estimate geological hazards.
- **Natural Hazards Partnership**: is a group of 17 collaborating public sector organisations comprising government departments, agencies and research organisations. The NHP provides a

mechanism for providing co- ordinated advice to government and those agencies responsible for civil contingency and emergency response during natural hazard events.

Each VRC uses the virtual research environment according to its own specific requirements for data, software, best practice and community engagement. This user-centric approach allows an assessment to be made of the capability for the proposed solution to satisfy the heterogeneous needs of a variety of Earth Science communities for more effective collaboration, greater efficiency and innovative research.

RESEARCH OBJECT

Central to the EVEREST approach is the concept of the Research Object (RO), which provides a semantically rich mechanism to aggregate related resources about a scientific investigation so that they can be shared together using a single unique identifier. The original definition of RO is available in Bechhofer et al. [2]. Although several e-laboratories are incorporating the research object concept in their infrastructure, the work done with research objects during EVER-EST, is a novel effort done to adapt the RO model to Earth Science and support automatic generation of research object content-based metadata as presented at the 2017 IEEE 13th International Conference on e-Science [1]. The EVER-EST VRE is the first infrastructure to leverage the concept of Research Objects and their application in observational rather than experimental disciplines.

Research objects aim to account, describe and share everything about your research, including how those things are related.

- To provide a logical organization in a single information unit of the materials, methods and outcomes of an investigation
- To uniquely identify and share your research materials and methods with other scientists at discrete milestones of the investigation
- To be recognized and cited
- To provide evidence to findings claimed in scholarly articles
- To enable reproducibility and reuse
- To preserve scientific results, preventing decay

A Research Object (RO) is defined as a semantically rich aggregation of resources that bundles together essential information relating to experiments and investigations. This information is not limited merely to the data used and the methods employed to produce and analyse that data, but it may also include the people involved in the investigation as well as other important metadata that describe the characteristics, inter- dependencies, context and dynamics of the aggregated resources. As such, a research object can encapsulate scientific knowledge, workflows and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge within and across relevant communities, and in a way that supports reliability and reproducibility of investigation results [4].

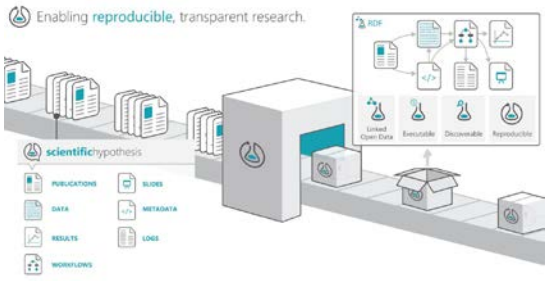


Figure 3 - Research Object

More specifically, by encapsulating workflows, using Apache Taverna, into research objects and accompanying them with the necessary data and metadata needed for their execution and understanding, one makes the latter more (re-)usable and preservable. This metadata can include, among others, details

like authors, versions, citations, etc., and links to other resources, such as the provenance of the results obtained by executing the workflow or datasets used as input. Such additional information enables a comprehensive view of the scientific investigation, encourages inspection of its different elements, and provides the scientist with a clearer picture of the investigation's strengths and weaknesses with respect to decay, adaptability and stability.

The research object recommendation system that shall be used as a basis in this project is again derived from the WF4Ever project and consists of two components:

- The Research Object Recommendation Service API that combines a variety of recommender algorithms each of which implements a different recommendation paradigm.
- The Collaboration Spheres Web Application, a graphical user interface that implements a novel visual metaphor for the more intuitive interaction of the user with the Recommendation Service.

In turn, the visual metaphor implemented by the Collaboration Spheres web application is based on a set of concentric spheres centred around a central point that represents the user. These spheres represent different types of similarity metrics between the context of interest and the results obtained by the recommenders. The context is expressed by the user as a collection of research objects as well as other users that the user finds relevant for a particular purpose. The distance between the center, i.e. the user and the context of interest, and the two external spheres, where recommendation results are displayed, provides a notion of confidence about the recommendations. The closer to the center, the more specific the recommendation result will be with respect to the user and the current context of interest.

CASE STUDY: EVALUATE HOW HUMAN ACTIVITIES CAN CAUSE POSIDONIA MEADOWS REGRESSION

Coastal anthropogenic activities increased worldwide in the last half century, amplifying the pressures on marine coastal ecosystems (Millennium Ecosystem Assessment MEA 2005). The management of those multiple and simultaneous threats requires reliable and precise data on the distribution of the pressures and of the most sensitive ecosystems (Halpern et al. 2008; Micheli et al. 2013). In this case study, starting from historical remote sensing data of Posidonia meadows distribution, the Sea Monitoring (SM) VRC detected Posidonia regression areas off shore the Apulia region in Italy and compared their distribution with the different human activities identified by the Change Detection WPS developed by Land Monitoring (LM) VRC.

LM run the change detection WPS using the EVER-EST VRE service in the Apulia Region and created a RO encapsulating the Taverna workflow and the results as .shp file. In parallel SM run runs a workflow implemented to detect Posidonia regression using the EVER-EST VRE Virtual Machine and created a RO with data, results, and workflows.

Overlaying through the EVER-EST VRE globe the results from the LM and SM research object it was possible to visually identify a correlation visual between the human activities detected by LM and the Posidonia regression off shore Gallipoli detected by SM.

Among the various types of human activities, the mechanical damages resulting from boats anchoring in shallow coastal waters appear to be responsible for localized regressions of Posidonia oceanic meadows.

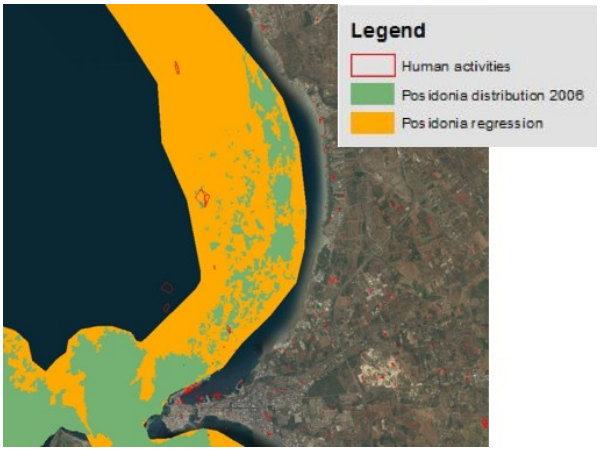


Figure 4 – Overlay between SM and LM Research Object results

CASE STUDY: CROSS-FERTILIZATION BETWEEN JELLYFISH OUTBREAKS & ANOMALIES DETECTION IN THE MEDITERRANEAN SEA.

As mentioned, one of the main feature of EVEREST is to provide Data Scientists the possibility to use already validated RO and/or create a cross validation between RO and between Virtual Research Environments. These characteristics have been used for a cross-fertilization study in synergy with UniSalento biological researchers group located in Lecce. The group is specialized on the quantification of deterministic and stochastic components of environmental change that lead to outbreaks of maritime species: in this specific case, the jellyfish.

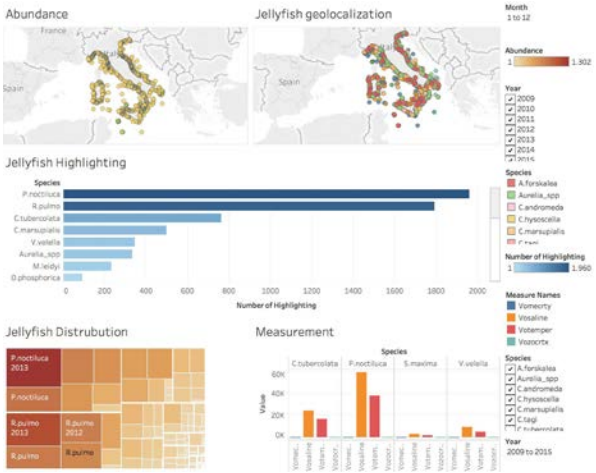


Figure 5 - Jellyfish analysis Dashboard

The Research Objects created by UniSalento have been cross-fertilized with the RO on "Mediterranean Sea Anomalies detection" developed during the Master. This can be considered as a good example of joint work between two communities – Earth Observation researchers and Maritime Biologist – which could be not necessarily strictly linked in their everyday activities and that was de facto facilitated by the common use of RO's and the adoption of the EVER-EST infrastructure as working environment. The analysis led to the successful identification of correlations between the two phenomena over specific areas of the Adriatic Sea.

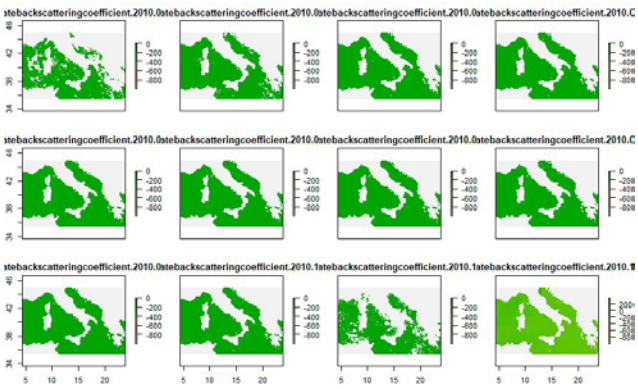


Figure 6 - Example of R plot for Mediterranean Sea Anomalies Detection

Some results have been graphically represented using an EVER-EST GIS tool overlapping all information produced by both studies.

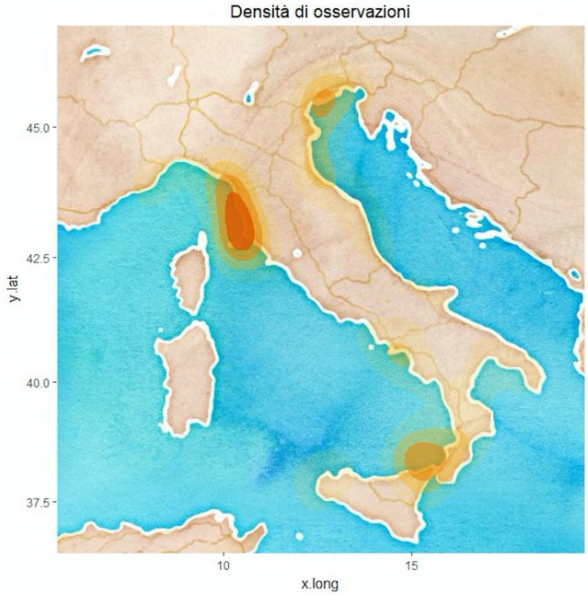


Figure 7 – Observations density

Partial results were collected in terms of light correlations with temperature, chlorophyll and particulate. We have identified some years to be better analysed and the need to create a model to solve the seasonality problem. Further analysis need to be performed applying an ad-hoc non-metric multidimensional scaling (NMDS) analysis for defining principal components. The plots on the density help the further verification in terms of reduction of the pixel polygon for time series definition and reduction in time consuming and performances.

REFERENCES

[1] Gomez-Perez, J.M., Palma, R., Garcia-Silva, A.: Towards a human-machine scientific partnership based on semantically rich research objects. In: 2017 IEEE 13th International Conference on e-Science (e-Science). pp. 266–275 (Oct 2017)

[2] S Bechhofer, I Buchan, D De Roure, P Missier, J Ainsworth, J Bhagat, P Couch, D Cruickshank, M Delderfield, I Dunlop, M Gamble, D Michaelides, S Owen, D Newman, S Sufi, and C Goble. Why linked data is not enough for scientists. Future Generation Computer Systems, 29(2):599 – 611, 2013. Special section: Recent advances in e-Science.

[3] K Belhajjame, O Corcho, D Garijo, J Zhao, P Missier, DR Newman, R Palma, S Bechhofer, E Garcia- Cuesta, JM Gomez-Perez, G Klyne, K Page, M Roos, JE Ruiz, S Soiland-Reyes, L Verdes-Montenegro, D De Roure, and C Goble. Workflow-centric research objects: A first class citizen in

the scholarly discourse. In 2nd Workshop on Semantic Publishing (SePublica), number 903 in CEUR Workshop Proceedings, pages 1–12, Aachen, 2012.

[4] R Palma, P Hołubowicz, O Corcho, JM Gomez-Perez, and C Mazurek. Rohub—a digital library of research objects supporting scientists towards reproducible science. In Semantic Web Evaluation Challenge, pages 77–82. Springer, 2014.

[5] ESA, NERC, INGV, ISMAR, SatCen, "Use Cases Description and User Needs" , EVER-EST DEL WP3-D3.1

[6] ESA, NERC, INGV, ISMAR, SatCen, "Workflows and Research Objects in Earth Science - Concepts and Definitions", EVER-EST DEL WP4-D4.1

[7] ESA, NERC, INGV, ISMAR, SatCen, "VRE Architecture and Interfaces Definition" , EVER-EST DEL WP5-D5.1

[8] Lisandro Benedetti-Cecchi, Antonio Canepa, Veronica Fuentes, Laura Tamburello, Jennifer E. Purcell, Stefano Piraino, Jason Roberts, Ferdinando Boero, Patrick Halpin, "Deterministic Factors Overwhelm Stochastic Environmental Fluctuations as Drivers of Jellyfish Outbreaks", doi:10.1371/journal.pone.0141060

[9] http://www.researchobject.org/

[10] http://www.scidip-es.eu/

[11] http://ceur-ws.org/Vol-679/paper4.pdf

[12] http://www.geowow.eu/

[13] http://www.envriplus.eu/

Building an Infrastructure for Climate Model Archives

Martin Juckes[1], Alison Pamment[1], Charlotte Pascoe[1], Ag Stephens[1]

[1]Centre for Environmental Data Analysis, RAL Space, UKRI STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot, OX11 0QX, United Kingdom.
Email:Alison.Pamment@stfc.ac.uk

This paper describes how several technologies have been drawn together to create an infrastructure for archiving and providing user access to climate model data from the Coupled Model Intercomparison Project (CMIP). The work described here has been carried out at the Centre for Environmental Data Analysis (CEDA) as part of an international collaborative effort among data service providers to organise and provide access to CMIP6 data.

Introduction

The CMIP Project has been coordinating climate model experiments involving multiple international modelling teams since 1995. CMIP is an activity of the Working Group on Coupled Models (WGCM), which in turn is part of the World Climate Research Programme (WCRP). The current sixth phase of the CMIP process (CMIP6) consists of a suite of common experiments and 21 separate CMIP-endorsed Model Intercomparison Projects (MIPs) making a total of 244 separate experiments (Eyring et al., 2016). The combined data produced by all the CMIP6 experiments is anticipated to be in the volume range of 15 - 30 PBytes and will be composed of binary data in NetCDF format (Stockhause and Lautenschlager, 2017).

ES-DOC for CMIP6

CMIP6 introduces an ambitious new organisational structure to cope with the rapidly expanding scope of climate modelling, leading to the creation of 21 different international science consortia. Each consortium is responsible for developing research objectives, experimental designs and data requirements for one of the approved MIPs. Figure 1 shows a schematic representation of all the MIPs; each one focuses on a particular aspect of climate modelling, for example, C4MIP compares models that simulate the global carbon and nitrogen cycles, while ISMIP6 focuses on land based ice sheet modelling. Some experimental designs can be important to more than one MIP, for example the 'piControl' experiment (Figure 1, centre) refers to simulations run under forcing conditions of pre-industrial greenhouse gases. The majority of MIPs use pre-industrial control as a baseline when comparing the behaviour of different models in changing climate conditions.

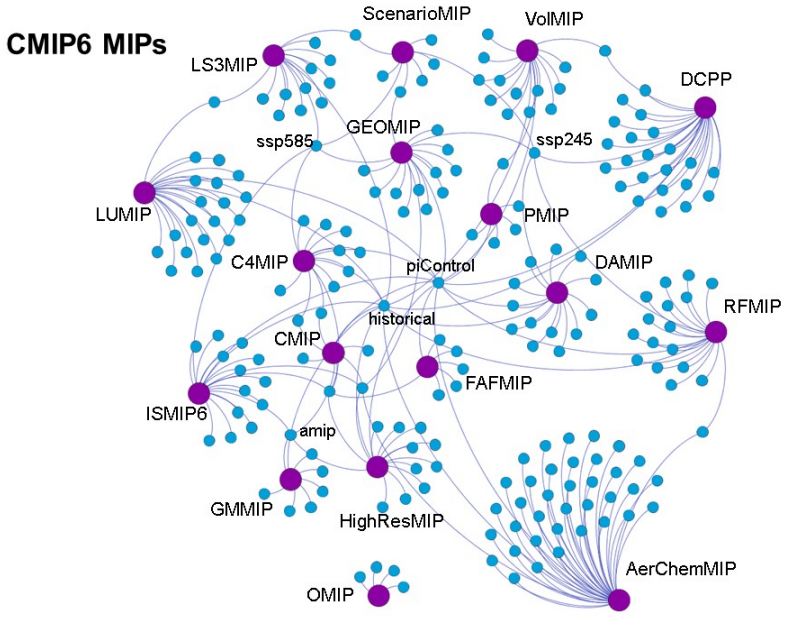


Figure 1. Schematic diagram showing the 21 Model Intercomprison Projects (MIPs) in CMIP6. The large purple dots represent the individual MIPs. Each small blue dot is a climate model experiment and the lines show how the experiments contribute to the MIP

The Earth System Documentation (ES-DOC) project nurtures an eco-system of tools and services in support of Earth System documentation creation, analysis and dissemination. Such an eco- system enables the scientific community to better understand and utilise Earth system model data. ES-DOC is coordinated with other community efforts such as CMIP and the Earth System Grid Federation (ESGF) via WCRP WGCM and its Infrastructure Panel.

ES-DOC model descriptions include scientific domain descriptions which include all the key properties which are likely to be compared between model domains (for example, resolution, grid extent, tuning properties, processes simulated). The depth of the model information collected by ES-DOC is deliberately limited, with a focus on scientific descriptions rather than describing software details, both to simplify the documentation task in terms of scope, but also to force more concise descriptions to allow for more salient comparisons.

ES-DOC experiment descriptions: The large number and variety of model experiments and their interconnected nature necessitates a systematic approach to capturing metadata about the experiments themselves and making this information available to data users. Precise descriptions of the suite of CMIP6 experiments have been captured in a Common Information Model (CIM) database (Lawrence et al., 2012) within ES-DOC. The database contains descriptions of forcings, model configuration requirements, ensemble information and citation links, as well as text descriptions and information about the rationale for each experiment. The database was built from statements about the experiments found in the academic literature, the MIP submissions to the WCRP, WCRP summary tables and correspondence with the principle investigators for each MIP. It allows end users of the climate model archive to ask questions such as 'Which MIPs make use of experiment A?'

ES-DOC ensembles descriptions: When valid CMIP6 datasets are submitted to an ESGF archive, their NetCDF metadata are read on the ESGF node to determine which ensembles and simulations the datasets were created for. This is done automatically as part of the ESGF publication process. The raw descriptions of these ensembles and simulations are then automatically sent to ES-DOC for publication. ESGF is described more fully in the last section of this paper.

Further Info URL: Each CMIP6 NetCDF file contains an attribute called “further_info_URL” which is the URL of a landing page from which all of the documentation relevant to the data may be accessed. This landing page will describe the ensemble for which the simulation was run and will contain links to documentation for the ensemble’s experiment, the simulation itself and any other simulations that the ensemble may contain. Each simulation document will contain a link to the description of the model that ran it.

The CMIP6 Data Request

The WCRP WGCM mandated that the data requirements should be consolidated into a single data request. The primary objective is to enable modelling centres participating in CMIP6 around the world to generate consistently defined data variables, and to do so with as much automation as possible.

The CMIP6 Data Request compiles data requirements from all the MIPs into a consolidated technical document. The request document defines over 2000 parameters with detailed specifications of technical metadata for each parameter, and links them to the experiment specifications provided in ES-DOC. Many of the MIP consortia have overlapping scientific interests, and interoperability of data between the consortia will be crucial to the success of CMIP6. The data request has many uses: it allows modelling groups and data service providers to get an overview of the data volumes expected in the CMIP6 archive, it tells the modelling teams which parameters they should be saving from their numerical experiments, and it tells those interested in analysing the results which variables they should expect.

The data request has been developed as an XML document with a supporting python library and a web site

allowing the contents to be navigated. The python library supports both a command line interface and an API.

Within the data request, individual data variables are defined by metadata attributes from the Climate and Forecast convention (CF) (Eaton et al., 2017), particularly the 'standard name' attribute to identify the parameter and the 'cell methods' attribute to define sub-grid processing (e.g. grid averaging, or masking). Additional information is provided through more than one hundred 'coordinate' parameters, many of which are specifying different land cover and land use types which will be used to analyse a range of land surface processes which are new to CMIP6.

Extending CF Controlled Vocabularies for CMIP6

The CMIP6 data will be archived in the NetCDF (Network Common Data Form), a well-established and widely adopted standard data format in the climate and environmental science research community. NetCDF consists of a file storage format, a library of data access routines (Unidata, 2015) and a metadata convention for use within the files. Other metadata conventions have been developed to extend those defined in the NetCDF interface, the aim being to serve the needs of those working in particular science domains. The CF metadata conventions are one example and were originally developed to enable sharing of large geolocated datasets such as those produced by climate and numerical weather prediction models. CF is a set of conventions entirely driven by the user community and over a period of almost twenty years has undergone a series of extensions to accommodate model developments. CF-NetCDF is also being increasingly used as a format to store in situ and remote sensing observations, such as radiosonde measurements and radar and satellite instrument data.

Within the CF conventions a number of controlled vocabularies are used to aid identification of the parameters in the data request. The largest of these is the 'standard name' vocabulary, used to identify the geophysical variable, and currently containing over 4,000 entries. Approximately 400 of these have been introduced to meet the requirements of CMIP6 and a similar number of names was introduced for the previous CMIP phase, known as CMIP5. The growing number of standard names reflects the increasing complexity of climate and earth system models. The variables they describe allow detailed diagnosis of atmosphere and ocean dynamics, representations of physical processes such as the earth's radiation budget, clouds and convection (collectively known as 'parameterizations') and other parts of the climate system such as ice sheets on land and sea.
Many of the new standard names added for CMIP6 are required for models of the global carbon and nitrogen cycles in the atmosphere, oceans, vegetation and soil. Names have also been added to describe detailed cloud microphysics processes, volcanic aerosol and land surface processes such as human land use changes.

The list of standard names is published as a table to supplement the CF conventions (Eaton et al., 2017) (Chapter 3.3) and also in the NERC Vocabulary Server. Each name is accompanied by a description and a 'canonical' unit of measurement, for example all standard names for velocity variables have canonical units of metres per second. The names themselves, and the corresponding units, are attached as attributes to individual data variables within the netCDF files while the descriptive text can be found in the published standard name table. The primary purpose of the standard name attribute is to help a data user to determine whether variables that have been calculated using different climate models, each with its own formulation, are directly comparable with one another and with observations of the same phenomenon.

The Earth System Grid Federation (ESGF)

ESGF (Williams et al, 2017) is an international collaboration of data centres and providers that develops the software underpinning most global climate change research, notably periodic scientific assessments by the Intergovernmental Panel on Climate Change (IPCC). Through ESGF, petabytes of high-profile climate simulations are archived and replicated across the globe. The federation maintains and develops tools and interfaces for data management, access and analysis. Over 20 organisations around the globe are involved in ESGF with participation ranging from actively managing major software packages through running “nodes” for disseminating climate data.

Whilst ESGF systems are already currently available, the recent focus has been on preparation for the CMIP6 project. This has led to the development of a number of new features intended to aid both data management

and accessibility. The look and feel of the 'CoG' web front-end has been improved and the search interface enables search by project, organisation, experiment (as described in ES-DOC), variable (identified by the CF standard name, or alternatively, a CMIP6 'short name'), data node and a range of other facets. The search results are now more useful because they link to a range of additional resources relevant to the data sets of interest.

The 'Errata' service provides a mechanism for recording and reporting errors and other issues found with data sets published to ESGF. It enables scientists to upload information about such issues so that they are searchable and are accessible alongside the data set records themselves. All traceability of files and data sets will be significantly advanced with CMIP6 due to the 'Handle' service that captures a unique persistent identifier for each object published to ESGF. This acts as a canonical identifier that can be used to connect disparate services and databases so that end- users can track appropriate metadata and provenance information across the federation. Given the estimated size of the CMIP6 archive, much effort has gone into optimising networks and data- replication systems to enable high-bandwidth intercontinental data transfers. These are currently being tested as part of a series of Data Challenges being run by the ESGF collaborators in order to ensure readiness for receipt of the torrent of data that is expected soon as part of CMIP6.

References

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Bentley, P., Blower, J., Caron, J. Höck, H., Juckes, M., Pamment, A., Signell, R., Rappa, G., and Raspaud, M. *2017 NetCDF Climate and Forecast (CF) Metadata Conventions Version 1.7, 17 August 2017. Available at http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html [Last accessed 30 April 2018].*

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. 2016 Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9: 1937-1958. DOI: http:// doi.org/10.5194/gmd-9-1937-2016.

Lawrence, B.N., Balaji, V., Bentley, P., Callaghan, S., DeLuca, C., Denvil, S., Devine, G., Elkington, M., Ford, R.W., Guilyardi, E., Lautenschlager, M., Morgan, M., Moine, M.-P., Murphy, S., Pascoe, C., Ramthun, H., Slavin, P., Steenman-Clark, L., Toussaint, F., Treshansky, A., and Valcke, S. 2012 Describing Earth system simulations with the Metafor CIM. *Geoscientific Model Development*, 5: 1493-1500. DOI: http://doi.org/10.5194/gmd-5-1493-2012

Stockhause, M. and Lautenschlager, M. 2017 CMIP6 Data Citation of Evolving Data. *Data Science Journal*, 16: 30. DOI: http://doi.org/10.5334/dsj-2017-030.

Williams, D N. et al. 2017 U.S. DOE. 6th Annual Earth System Grid Federation Face-to- Face Conference Report. DOE/SC-0188. U.S. Department of Energy Office of Science, March 2017. DOI: https://doi.org/10.2172/1369382.

Unidata 2015 *Network Common Data Form (netCDF) version 4.3.3.1 [software]*. Boulder, CO: UCAR/Unidata. DOI: http://doi.org/10.5065/D6H70CW6.

ESA's Research and Service Support as a Virtual Research Environment for Heritage Mission data valorisation

Paulo Sacramento[1], Giancarlo Rivolta[2] and Joost van Bemmelen[3]

[1] Solenix c/o European Space Agency (ESA-ESRIN), Frascati, Italy

[2] Progressive Systems c/o European Space Agency (ESA-ESRIN), Frascati, Italy

[3] European Space Agency (ESA-ESRIN), Frascati, Italy Corresponding author: Paulo Sacramento (paulo.sacramento@esa.int)

This paper proposes a view of the ESA's Research and Service Support (RSS) service as the prototype of a Virtual Research Environment for the valorisation of data from ESA Heritage Missions like Envisat, ERS-1 and ERS-2. It starts by covering the state of the art of virtual research environments, highlighting related previous work by ESA, including as part of EU framework projects since the late 1990s, and moves on to the rationale, motivation and current status of RSS, which *de facto* constitutes a prototype of a virtual research environment for Heritage Mission data valorisation. The RSS offering is described, from the simple provisioning of Virtual Machines previously prepared with a wealth of tools for EO scientists and researchers, potentially mission- oriented (e.g. to Proba-V), or Virtual Machines that highly simplify the exploitation of large volumes of EO data collocated with Cloud resources (the so-called RSS CloudToolbox), to the operation or support to the operations of full-blown exploitation environments such as the Proba-V Mission Exploitation Platform (PV-MEP) and the GeoHazards Thematic Exploitation Platform (GEP). Specific emphasis is put on the RSS CloudToolbox, which provides the EO Community with resource flexibility (processing power and storage) avoiding unnecessary resource allocation, and seamless access both in terms of geographical location (VMs accessible from anywhere in the world) and in terms of devices used for access (PC, laptop, tablet). The Cloud providers on which such Virtual Machines and Exploitation Platforms rely on provide high-speed network connections and data access to vast catalogues of Heritage Mission data (the full Envisat, ERS-1 and ERS-2 archives), allowing users to autonomously test and run their own algorithms, with the ultimate goal of building long data time series from the early 1990s (and even before if missions like Seasat, JERS-1, etc., are considered) until present time with the Copernicus data. The RSS offering has proven effective in supporting university and workshop courses as well as research institutions, SMEs, technology projects and individual researchers. Examples of such success stories are given in the paper. Finally, the ESA perspective on the evolution of the RSS service is given, to make sure that the decades of relevant work in this domain, which constitute a sound basis for further construction, are leveraged rather than lost.

Keywords: Long Term Data Preservation; Virtual Research Environments; Heritage Missions; Earth Observation; Research; Data Valorisation

Introduction

The term Virtual Research Environment, despite having been coined several years ago, has recently come into mainstream use to refer to very different and heterogeneous sets of systems, services and platforms that support the research community in its endeavour to evolve scientific knowledge, by offering remotely-hosted (hence the word virtual) functionality and resources, that attempt on the one hand to remove traditional hurdles and on the other hand to speed-up and make research work more

effective, allowing scientists to focus larger slices of their time on their core business – the development and continuous improvement of processing algorithms -, rather than on lateral issues such as ICT and data access.

Numerous EU R&D framework projects, as well as national and international industry initiatives, have dealt with these kinds of environment, allowing the build-up and evolution of technology and know-how, but also very importantly promoting community building in various research domains.This, in turn, has led to scientific collaboration and the sharing of experiences, data and results. In fact, early efforts resulting from the proliferation of TCP/IP networks, the Internet, the World Wide Web and Grid computing fostered the creation of the concept of a "virtual organization".

The European Space Agency, within its Earth Observation Programme, has also garnered years of experience in related topics. Starting from the Grid computing paradigm, the Agency moved onto an overarching service model for research and service support, which is the main subject of this paper. In parallel, ESA also laid down the founding stones for its "open science program", which identifies the several phases of the scientific process, from Conceptualization, to Data Gathering, Analysis, Publication and Review, putting them in relation to the most recent trends such as open access publications, alternative reputation systems, citizen science, open data access, online courses, etc..

Other related, and popular, initiatives triggered by ESA in the more recent years are the so-called Exploitation Platforms, both mission-oriented as in the case of the Proba-V Mission Exploitation Platform and theme-oriented as in the case of the Coastal, Urban, Polar or Geohazards Thematic Exploitation Platforms (http://tep.eo.esa.int). Also the Sentinel Application Platform, SNAP, known as the Sentinel Toolbox, is worth mentioning given its relevance and popularity amongst the Copernicus user and scientific community.

Latter initiatives and systems have no doubt been influenced in their foundation and design by the "tsunami" of data that was expected and has started to materialize with the launch and entry into operations of the Sentinel units of the EU Copernicus programme. In the Big Data era, systems have to be designed with the challenges represented by its Vs – Volume, Velocity, Veracity, Variety (and others depending on the source) - in mind. This was not the case, however, for ESA Heritage Missions like ENVISAT, ERS-1 and ERS-2, defined as the missions which have not been in in-orbit operations since at least 5 years. Whilst modern infrastructure and operations concepts have fully embraced the "bringing users to the data" paradigm, Heritage Mission data did not come into being at the same time and therefore there is a gap to fill if its adequate exploitation and valorisation are to be made effective, particularly when contextually to the exploitation and valorisation of Copernicus data and - of utmost importance and as a natural goal
– in an integrated fashion to allow long time series studies from as early as the 1970s to present day. ESA's Research and Service Support service exists to help realise this vision.

ESA's Research and Service Support

As introduced in the previous section, ESA and its industrial partners have, in the last few years, developed an overarching model for research and service support. This model has the research process as its cornerstone and foresees the provision of tools and services in support of the process's several phases and actors, naturally with the scientist in a central role. This is illustrated in Figure 1.

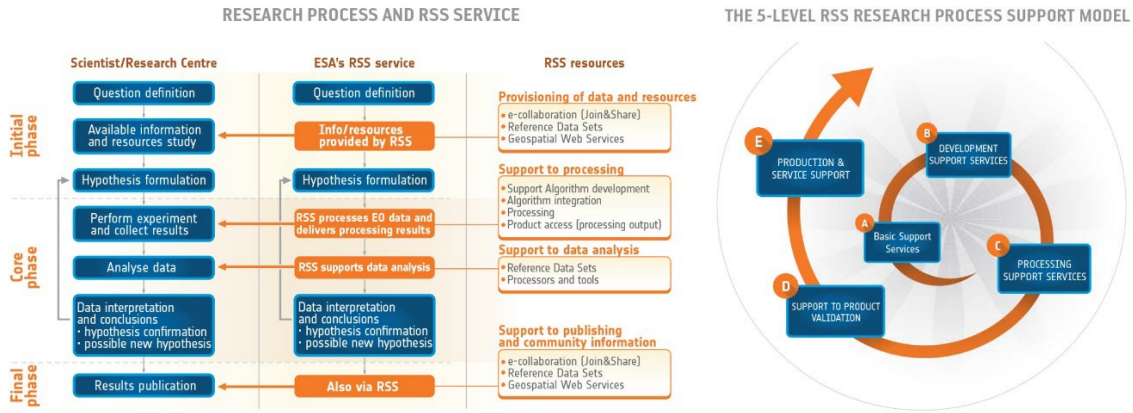


Figure 1: ESA RSS and its resources in relation to the research process; the 5 levels of the support model

As the scientist begins to define his/her question and formulating a hypothesis, he/she collects available information and resources – the state of the art. RSS supports him/her by providing such information and resources, for example through e-collaboration tools and reference data sets. These can be considered part of the basic support services. In the core phase of his/her work, the scientist will perform experiments and collect results, analyse and interpret the data, extracting some conclusions which may confirm/reject the hypothesis or possibly lead to new hypotheses. In this phase, RSS supports by making available processing resources, setting them up (either by helping algorithm development or just its integration) and running processing on Earth Observation data on behalf of the user. It also supports the data analysis, giving access to reference data sets, processors and tools. These constitute development and processing support services, as well as product validation support services when applicable.

Finally, once results have been generated and documented, the scientist will want to publish them. RSS offers utilities that enable this, such as e-collaboration tools and geospatial web services (e.g. when a user wants to publish generated data as WMS layers).

RSS for Heritage Mission data valorisation

One of the tools made available by RSS is the so-called Cloud Toolbox. Through it, researchers can get customised Virtual Machines equipped with some of the latest ESA and third-party tools for Earth Observation data processing (e.g. SNAP, BEAM, PolSARPro). These Virtual Machines come with free/sponsored storage and computing resources on any one of several Cloud providers and, fundamentally, are hosted close to large libraries of Earth Observation data. This maximizes the processing speed that can be achieved and reduces the data management woes.

Figure 2: The RSS Cloud Toolbox

Cloud Toolbox instances (see Figure 2) can be, for example, preloaded with stacks of ENVISAT and Landsat products – two of the most relevant Heritage Missions that ESA deals with -, and these can be freely manipulated and processed by researchers. Such instances provide EO data users with resource flexibility, accessible via their own devices (PC, laptop, tablet) from anywhere in the world.

Looking at RSS from the service perspective instead, a very concrete and relevant example of its contribution to Heritage Mission data valorization is the SBAS interferometric processing offered on 10 years of ENVISAT data over the entire world, implemented using the G-POD infrastructure (see Figure 3).

More generically, through the tools and services offered by RSS, researchers have the possibility of creating their own long time series joining together heritage mission data with current Sentinel data.

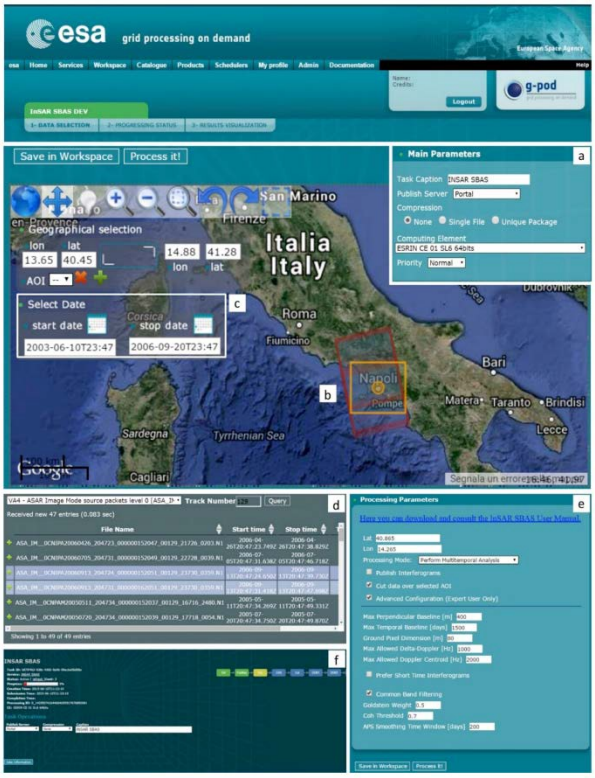


Figure 3: The InSAR SBAS service on G-POD (Grid Processing On-Demand)

RSS as a Virtual Research Environment

The Cloud Toolbox introduced in the previous section, but also other RSS tools and services not addressed in this paper, has been used successfully in the last few years to support, among others:University courses

- Workshop courses
- Research institutions
- Individual researchers

Two examples of such activities are the provisioning of customized toolboxes to Delft University of Technology (TU Delft) students attending Interferometry courses, or to La Sapienza University Data Science students attending lessons on Earth Observation data exploitation. It is also relevant to mention the support to researchers working on various different research topics, ranging from evaluation of the potential of optical time series for improving Land Cover classification (Wageningen University) to algorithm development for generating decorrelation products from pairs of radar images (BRGM).

There is then a clear orientation of the service towards the research and academic community, which brings about its dimension or view as a Virtual Research Environment. In a nutshell, a Virtual Research Environment in this sense is web-based, is tailored to serve the needs of the community, is expected to provide the commodities needed to accomplish the community's goal(s) and is open and flexible with respect to the overall service offering and lifetime.

Conclusion

This paper has described how ESA's Research and Service Support service contributes to the valorisation of ESA Heritage Mission Earth Observation data. It has also presented a view of RSS as a Virtual Research Environment that equips researchers with tools and services that facilitate their work. When such an environment is used to operate on Heritage Mission data, either standalone or in relation to more recent data (e.g. from Copernicus), it becomes a natural part of the valorisation process.

ESA is currently defining use-cases, requirements and an architecture for the evolution in a comprehensive manner of its Heritage Mission valorisation infrastructure. RSS has through the years built-up competence in this area, by focusing predominantly on ENVISAT and ERS data, and will thus have an important role in this strategy. It constitutes a sound basis for further construction and should be leveraged to achieve the goal of maximizing the usage of the ESA EO data legacy.

References

Candela, L, Castelli, D, Pagano, P 2013 *Virtual Research Environments: An Overview and a Research Agenda*. Data Science Journal, Volume 12, 10 August 2013

De Luca, C, Cuccu, R, Elefante, S, Zinno, I, Manunta, M, Rivolta, G, Casola, V, Lanari, R, Casu, F 2015 *Unsupervised on-demand Web Service for DInSAR processing: the P-SBAS implementation within the ESA G-POD Environment*, Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International, Issue Date: 26-31 July 2015

Eberenz, J, Verbesselt, J, Herold, M, Tsendbazar, N, Sabatino, G, Rivolta, G 2016 *Evaluating the Potential of PROBA-V Satellite Image Time Series for Improving LC Classification in Semi-Arid African Landscapes*, Remote Sensing 8 (2016)12. - ISSN 2072-4292 - 11 p.doi:10.3390/rs8120987, 2016

Farres, J, Mathot, E, Pinto S 2010 *G-POD: A Collaborative Environment for Earth Observation at the European Space Agency*, Proceedings of the ESA Living Planet Symposium, 28 June- 2 July 2010, Bergen, Norway, Special Publication SP-686 on CD-ROM, ESA Publications Division, European Space Agency, Noordwijk, The Netherlands, 2010

Foster, I, Kesselman C.O. 1998 *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann

Fusco, L, Gonçalves, P, Brito, F, Cossu, R, Retscher, C 2006 *A new Grid-based system to assist users in ASAR handling and analysis*, European Geoscience Union General Assembly, Vienna, 02-07 April 2006

Manunta, M, et al 2016 *The contribution of the Geohazards Exploitation Platform for the GEO Supersites community*, EGU General Assembly, 2016

Marchetti, P. G., Rivolta, G, D'Elia, S, Farres, J, Mason, G, Gobron, N 2012 *A Model for the Scientific Exploitation of Earth Observation Missions: The ESA Research and Service Support*, IEEE Geoscience and Remote Sensing(162): 10-18, 2012

Mathieu, P.P., Borgeaud, M, Desnos, Y.L., Rast, M, Brockmann, C, See, L, Fritz, S, Kapur, R, Machecha, M, Benz, U 2017 *The Earth Observation Open Science Program*, IEEE Geoscience and remote sensing magazine, pp 86-93, Digital Object Identifier 10.1109/MGRS.2017.2688704, June 2017

VRE for meteorological and climatic processes analysis

Evgeny Gordov
Institute of Monitoring of Climatic and Ecological Systems Siberian Branch of Russian Academy of Sciences, Tomsk, Russian Federation
Institute of Atmospheric Optics, Siberian Branch of Russian Academy of Sciences, Tomsk, Russian Federation
E-mail: gordov@scert.ru (corresponding author)

Igor Okladnikov
Institute of Monitoring of Climatic and Ecological Systems Siberian Branch of Russian Academy of Sciences, Tomsk, Russian Federation
Institute of Atmospheric Optics, Siberian Branch of Russian Academy of Sciences, Tomsk, Russian Federation

Alexander Titov
Institute of Monitoring of Climatic and Ecological Systems Siberian Branch of Russian Academy of Sciences, Tomsk, Russian Federation
Institute of Atmospheric Optics, Siberian Branch of Russian Academy of Sciences, Tomsk, Russian Federation

Alexander Fazliev
Institute of Atmospheric Optics, Siberian Branch of Russian Academy of Sciences, Tomsk, Russian Federation

This paper describes a Virtual Research Environment (VRE) which provides an access to analytic instruments processing 19 collections of meteorological and climate data of several international organizations. This environment provides systematization of spatial data and related climate information and allows a user getting analysis results using geoinformation technologies.

Introduction

On-going climate changes, especially their extreme manifestations, such as heat waves, cold periods, heavy rains or snowfalls, storms, floods or droughts, have an increasing impact on economic, political and social processes (IPCC 2012; IPCC 2013; Sillmann et al. 2014). Reliable assessments of their trends and impacts on these processes are critical for the development of adequate local and / or regional strategies for adapting and mitigating negative effects of climate change, for example, for sustainable agriculture and forestry, or planned infrastructure. But such assessments are still missing for various parts of the world. This circumstance is an essential driver for the development of climatic characteristics' monitoring and climate modeling to assess possible future trends. Local and remote observations, as well as numerical modeling of climatic processes, resulted in an unprecedented growth of data archives. Such an increase in data archive volume makes using the traditional approach to climate information analysis doubtful, and requires new approaches based on distributed networks and usage of modern information technologies. Currently, the main efforts and resources in the world are focused on creating a sustainable distributed cyberinfrastructure for open, permanent, reliable and secure access to high quality Earth observation data and corresponding metadata. According to Candela, Castelli & Pagano (2013), all these efforts can be described as a development of a Virtual Research Environment (VRE) for climate domain. It is a system with following major features: (*i*) a web-based working environment; (*ii*) tailored to serve the needs of a targeted community; (*iii*) expected to provide a community with the whole array of products needed to accomplish the community's goal(s); and (i*v*) promotes sharing of research results. At the same time, the reliable analysis of climate changes, and nature and society's responses them require skills in dealing with big datasets, abilities to interact with powerful computing resources and complex numerical models, knowledge of modern methods of statistical analysis, and usage of high-level programming languages. The skills mentioned are not typical for specialists in the field of economic, political and social sciences, and unfortunately, it is completely uncharacteristic for decision-makers. Therefore, an integration of all those in the Internet-accessible research environment to provide specialists and decision-makers with reliable tools for studying economic, political and social consequences of climate change should be done.

In this work we present a state-of-the-art in a development of VRE based on a combination of web and GIS techniques. It integrates climatic data archives, interactive processing and visualization tools, and computing resources in a distributed Internet-accessible hardware and software complex for meteorological and climatic processes analysis.

Virtual Research Environment

The presented VRE is aimed at 'cloud' processing, analysis and visualization of geospatial gridded datasets in Earth system science using Internet-accessible interactive tools (Gordov et al. 2016; Okladnikov et al. 2015). It's based on a dedicated software framework consisting of three key components: a server-side computational backend; a server-side middleware represented by a geoportal; and a specialized JavaScript library containing typical widgets of web mapping client GUI which is based on AJAX technology. Geospatial datasets are processed by a set of validated software modules running by the backend. Results are represented by overlapped raster and vector cartographical layers accompanied by corresponding binary data files. VRE's functionality includes basic and complex statistical analysis of data, whilst online geo-information system (GIS) instruments give a user an ability to combine and map georeferenced results over a chosen cartographical basis. It provides specialists and users without programming skills with reliable and practical online instruments for integrated research of climate and ecosystems changes through a unified web interface.

Architecture

VRE's simplified architecture is shown in Figure 1. It represents a typical client-server structure, where in general case the server might be a set of geographically distributed standalone nodes providing common (federated) interface (API), and client applications (basically, Web-GIS client)

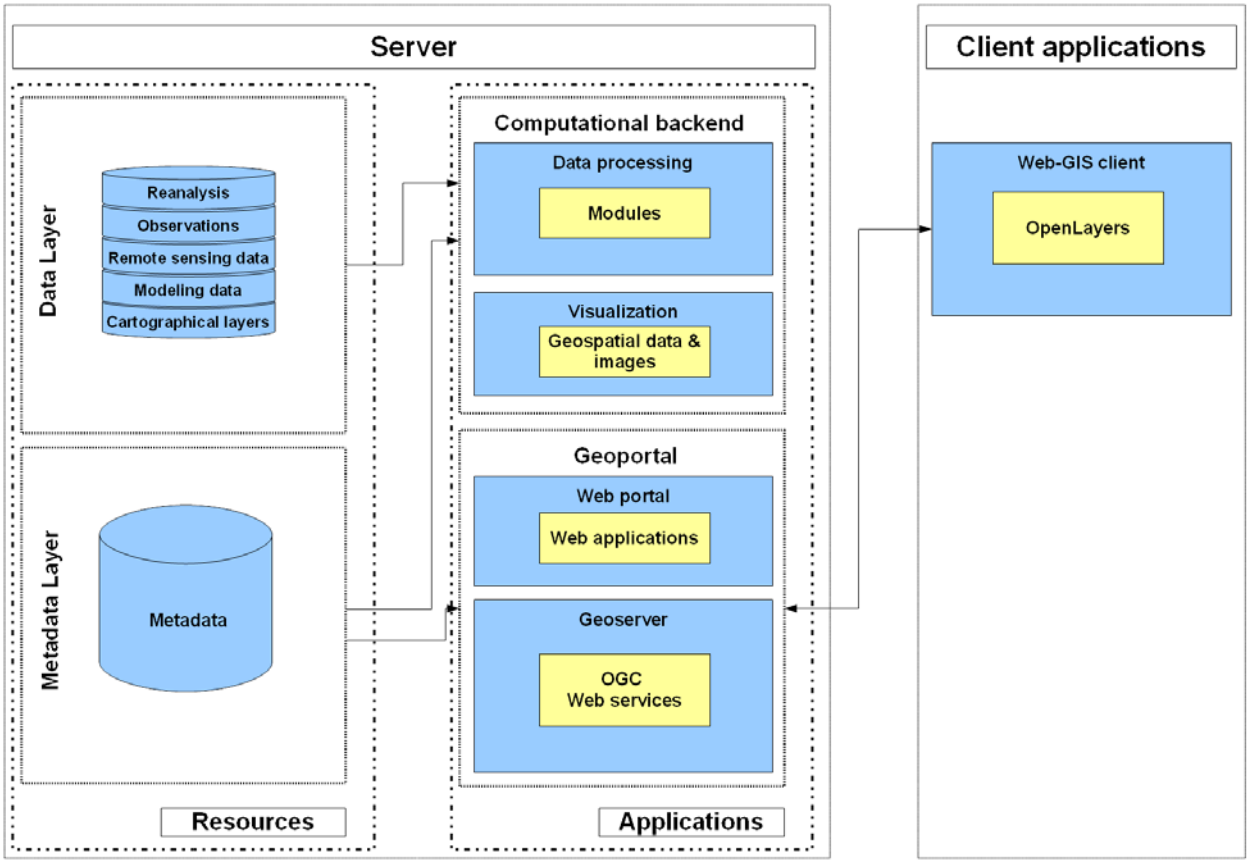


Figure 1: VRE's general architecture outline

The server part of the architecture includes a high-performance computing system with a data storage attached. It is presented by two tiers:

- resources tier, including data and metadata;
- server applications (middleware) tier.

The client part of the architecture is based on a modern graphical web browser. It is presented by a single 'Client applications' tier, respectively.

The ‘Resources’ tier of VRE employs two basic layers, which are Data and Metadata layers. The data layer contains datasets, located on the data storage system either in the form of collections of network Common Data Form (netCDF) files or PostGIS databases. Metadata layer is presented by the Metadata database, which describes geospatial datasets and their processing routines, and provides effective system functioning (Okladnikov, Gordov & Titov 2016). The database contains structured spatial and temporal characteristics of available geospatial datasets, their locations, and configurations of software components for data analysis. According to the chosen data storage model (Okladnikov, Gordov & Titov 2016), spatial datasets are mostly represented by collections of netCDF files grouped by spatio-temporal features and placed in the hierarchy of directories on data storage systems. Each netCDF file stores one or more variables containing values of meteorological parameters on a given spatio-temporal domain, and horizontal, vertical and time domain grids.

The ‘Server applications’ middleware tier consists of two basic software components: computational backend and geoportal.

Computational backend

The computational backend contains data processing and visualization software components. The data processing is a key software component containing computational modules based on GNU Data Language (GDL, http://gnudatalanguage.sourceforge.net/) and Python and providing integral geospatial data statistical processing as well as API to work with netCDF, Hierarchical Data Form (HDF), ESRI Shapefile data files and PostGIS databases. Depending on the result type required visualization component of the backend generates files in the following formats: GeoTIFF, ESRI Shapefile, Encapsulated PostScript, CSV, XML, netCDF, float GeoTIFF.

Computational modules allow calculating basic statistical parameters (mean, standard deviation, maximum and minimum values of meteorological parameter) and indicators of the temporal structure of the meteorological series (repeatability and continuous duration of atmospheric phenomena with values of meteorological parameters above or below the specified limits within the specified time range), reflecting regularities of random variables in time and space. In addition, computational modules for calculating indices of climate change (http://cccma.seos.uvic.ca/ETCCDMI/indices.shtml) were developed. They allow to extract information about extreme values of daily temperature and daily precipitation amount, and their probability characteristics. Some indices are calculated for fixed thresholds related to specific applications. Other indices are based on thresholds, which vary depending on the location of observation posts. In these cases, the threshold values are defined as respective percentiles of data series (Sillmann & Roeckner, 2008). Features of the temporal dynamics of climatic indices are defined by long-term components of time series, trends allowing to assess the change tendency of meteorological values, by assessment of statistical significance of identified trends, as well as by the degree of correlations between weather events. This sequence of procedures, including calculation of climatic parameters and studying their spatial and temporal dynamics, allows one to get the most complete picture of features of occurring fluctuations of the climate system in the studied region. The functionality of the computational backend can be easily extended on-demand by new modules developed by both developers and users.

Geoportal

Spatial Data Infrastructure (SDI) geoportal contains two basic components: web portal and Geoserver (http://geoserver.org). Geoserver provides cartographical web services such as Web Mapping Service (WMS), Web Feature Service (WFS) and Web Processing Service (WPS). In general, Web Processing Service provides standard HTTP interface for remote configuring and launching data processing software modules and presenting results in generic formats. The services can be used by either standard GIS environments or web applications.

The web portal serves as a connection point between different SDI elements (geospatial data, metadata, services and client applications). Its main feature is providing unified API for client web applications which comply with the conventional Boundless / OpenGeo architecture (Becirspahic & Karabegovic 2015). The web portal provides server-side part of the Web-GIS client application which complies with general INSPIRE

(INfrastructure for SPatial InfoRmation in Europe, https://inspire.ec.europa.eu) requirements to geospatial data visualization and implements computational processing services launching to support solving tasks in climate monitoring.

Conclusion

The fact that climate science deals with georeferenced data requires usage of a combination of web and GIS techniques, integrating data, processing and visualization tools and computing resources in a distributed Internet-accessible hardware and software complex. As a response, the development of the thematic VRE for meteorological and climatic processes analysis was initiated. It is a free, cross-platform, composite software complex with an open-source computational backend that complies with common data format conventions for climate data and provides functionality of common desktop software in a window of an Internet browser. It allows smooth adding of new computing nodes, data storage systems as well as provides solid computational infrastructure for regional climate change studies based on modern Web and GIS technologies. Usage of the metadata database improves system functional capabilities in terms of extending geospatial dataset archives and statistical processing routines as well as providing computational resources as web services. This thematic VRE will provide interdisciplinary distributed research groups of non-experts in information technologies (climatologists, ecologists, biologists, and decision makers) with easy-accessible reliable online tools for reliable analysis and visualization of multidimensional heterogeneous climatological datasets obtained from various sources. Possibility to get analysis results in GeoTIFF, ESRI Shapefile, Encapsulated PostScript, CSV, XML, netCDF, float GeoTIFF formats opens a way to practitioners to develop climatic services using their own applied software environment.

Acknowledgment

The authors thank the Russian Science Foundation for the support of this work under the grant No16-19-10257.

References

Becirspahic L and Karabegovic A 2015 Web portals for visualizing and searching spatial data. In: 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia on 25-29 May 2015 pp. 305-311. DOI: 10.1109/MIPRO.2015.7160284

Candela, L, Castelli, D and Pagano, P 2013 Virtual Research Environments: An Overview and a Research Agenda. *Data Science Journal*, 12: GRDI75–GRDI81. DOI: http://doi.org/10.2481/dsj.GRDI-013

Gordov, E, Shiklomanov, A, Okladnikov, I, Prusevich A and Titov A 2016 Development of Distributed Research Center for analysis of regional climatic and environmental changes. *IOP Conf. Series: Earth and Environmental Science*, 48: 012033. DOI: http://dx.doi.org/10.1088/1755-1315/48/1/012033

IPCC 2012 *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press.

IPCC 2013 *Fifth Assessment Report 'Climate Change 2013'*. Cambridge, UK: Cambridge University Press.

Okladnikov, I G, Gordov, E P, Titov, A G and Shulgina T M 2015 Information-computational System for Online Analysis of Georeferenced Climatological Data. In: Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), Obninsk, Russia, 13-16 October 2015. CEUR Workshop Proceedings. Vol. 1536. P. 76-80.

Okladnikov, I G, Gordov, E P and Titov A G 2016 Development of climate data storage and processing model. *IOP Conference Series: Earth and Environmental Science*, 48: 012030. DOI: https://doi.org/10.1088/1755-1315/48/1/012030

Riazanova, A A, Voropay, N N, Okladnikov, I G and Gordov, E P 2016 Development of computational module of regional aridity for web-GIS 'Climate'. *IOP Conference Series: Earth and Environmental Science*, 48: 012032. DOI: https://doi.org/10.1088/1755-1315/48/1/012032

Ryazanova, A A and Voropay, N N 2017 Droughts and Excessive Moisture Events in Southern Siberia in the Late XXth - Early XXIst Centuries. 2017. *IOP Conference Series: Earth and Environmental Science*, 96: 012015. DOI: https://doi.org/10.1088/1755-1315/96/1/012015

Sillmann, J. and Roeckner, E. 2008 Indices for extreme events in projections of anthropogenic climate change. *Climate Change*, 86: 83-104.

Sillmann, J, Donat M G, Fyfe J C and Zwiers F W 2014 Observed and simulated temperature extremes during the recent warming hiatus. *Environmental Research Letters*, 9(6): 064023. DOI: https://doi.org/10.1088/1748-9326/9/6/064023

Adding Value and Facilitating Data Reuse: the Case of the 4TU.Centre for Research Data

Maria Cruz[1], Egbert Gramsbergen[1]

[1]4TU.Centre for Research Data, TU Delft Library, Delft University of Technology, Prometheusplein 1, 2628 ZC Delft, The Netherlands
Corresponding author: Maria Cruz (M.J.MarquesdeBarrosCruz@tudelft.nl)

The history of the 4TU.Centre for Research Data goes back to 2008, when it started as a project of the libraries of three technical universities in the Netherlands. The aim was to serve the data curation needs of heterogeneous research communities. Fast forward ten years, and over 90% of the data stored in the 4TU archive are geoscientific datasets coded in netCDF (Network Common Data Form). This is a data format and model that, although generic, is mostly and widely used in atmospheric sciences and oceanography. As an endeavour to ensure that the 4TU.Centre for Research Data remains relevant and successful in the long term, we are exploring options for expanding the services related to netCDF data and potentially build a community of netCDF data depositors and users. Here we present the results of semi-structured, qualitative interviews with eleven researchers, all based in the Netherlands, who use and produce netCDF data; nine of them deposited netCDF data in the 4TU archive. These researchers represent heterogeneous research communities within the Earth sciences, with different views and attitudes to data archiving and publishing. Any new services or community building attempts will need to take this diversity into account. A common need for training and advice may guide the way forward for the 4TU.Centre for Research Data.

1. Introduction

The 4TU.Centre for Research Data (formerly known as '3TU.DataCentrum') was started in 2008 as a collaboration of the libraries of three universities of technology in the Netherlands: Delft University of Technology (TU Delft), Eindhoven University of Technology, and the University of Twente. The ambition was, and still is, to create and maintain a national state-of-the-art facility for storing and preserving science and engineering research data, and for making those data openly accessible. The 4TU data archive has been fully operational since 2010 and it has evolved to become a trusted and certified repository for science and engineering. As of 30 April 2018, the archive held 7581 datasets, corresponding to about 32.6 TB of data.

The 4TU archive was originally built "as a data curation facility to meet the diverse needs of heterogeneous research communities" (Rombouts & Princic 2010). Although it contains, and still attracts, heterogeneous data types, 90% of the data stored in the 4TU archive are environmental research data coded in netCDF (Cruz et al. 2018). Therefore, the 4TU.Centre for Research Data has a special interest in this area and it offers specific services and tools to enhance the access to and the use of netCDF datasets. In recent work (Cruz 2018, Cruz et al. 2018), we argued that repositories need to have a subject or format focus to remain relevant and successful in the long term. In the case of the 4TU.Centre for Research Data, that means exploring our options for providing further services related to netCDF data, be it technical services or training and guidance.

NetCDF is described by its authors as "a set of software libraries and self-describing, machine- independent data formats that support the creation, access, and sharing of array-oriented scientific data" (Unidata 2018). For netCDF datasets, besides the usual download, the 4TU archive offers access via the OPeNDAP (Open-source Project for a Network Data Access Protocol) protocol, themain advantage of which is the ability to inspect files (and metadata hidden within the files) and retrieve subsets of files without the need to download whole datasets.

2. Methods

To assess what expanded or novel netCDF services the 4TU.Centre for Research Data could potentially provide to its community of data depositors and users, we conducted nine semi- structured qualitative interviews with 11 researchers, all based in the Netherlands, who use and produce netCDF data. Most of these

researchers deposited netCDF datasets in the 4TU archive; only two of them hadn't done so. We interviewed researchers at all careers stages, namely: four PhD students, one post-doc, one senior scientist, three assistant professors, one associate professor, and one full professor. They were all geoscientists, working in areas ranging from atmospheric sciences and remote sensing, to hydrology, oceanography, and coastal engineering. They were mostly affiliated with technical universities in the Netherlands, especially TU Delft, but some were based at national research facilities and industry.

The interviews, which were not recorded, lasted around 60 minutes each and were conducted between November 2017 and April 2018. The interviewers took notes of key points during the conversation and wrote preliminary, more extended reports in the day or so after the interview. All interviewees were informed that the findings were going to be published, but were assured that they wouldn't be named and that no information would be individually attributed to them.

3. Results

3.1An overview of netCDF data stored in the 4TU data archive

It was clear, even before talking with any data depositors or users, that within the scope of netCDF data, which is mainly used in a limited number of geoscience disciplines, the 4TU.Centre for Research Data is still serving heterogeneous research communities, albeit in the Earth sciences.

In terms of volume, most of the netCDF data stored in the 4TU archive originate from one single experiment – the IRCTR Drizzle Radar (IDRA), developed at TU Delft's International Research Centre for Telecommunications and Radar (IRCTR) and installed on top of the Dutch meteorological observatory at Cabauw in the Netherlands (Otto & Russchenberg, 2014). This project has contributed with 2325 datasets to date, corresponding to about 27 TB of data (Otto et al. 2010). This is a growing time series of datasets, updated every few months, providing detailed observations of the spatial and temporal distribution of rainfall and drizzle around the radar's location.

The remaining 4198 netCDF datasets (as of 30 April 2018), corresponding to a total of about 3.2 TB of data, are either part of much smaller collections (less than 20 GB in size), or are individual datasets that range in size from about 600 KB to 136 GB. These datasets can be broadly classified as environmental research data, ranging from river discharge data (Hellebrand 2004) to measurements of aeolian sediment transport (Hoonhout, de Vries and Cohn 2016), and from climate projections (Mezghani, Dobler & Haugen 2016) to local mean sea level models (Gerkema & Duran Matute 2017).

3.2 Main findings from the interviews

3.2.1 Use of the 4TU archive

In addition to the heterogeneous nature of the datasets, the interviews showed that different communities and individual researchers use the 4TU archive in disparate ways. Some projects chose to store only raw data for long-term preservation; processed data for comparative analyses were stored elsewhere. Other projects stored both raw and processed data together with software and scripts used to process the data. The need for long-term preservation was particularly important for researchers dealing with climate data and long-term data series. Although some of the researchers affiliated with these projects mentioned the need and appreciation for processing and visualisation services, they didn't feel this should be a priority or a main role for the 4TU archive. In their opinion, archiving of (mostly) raw data for long-term preservation should be the focus of the 4TU archive.

For many of the researchers we interviewed though, while they appreciated the benefits of long- term preservation and of making their data publicly available, their main motivation to use the 4TU archive was to comply with publisher and journal requirements regarding data availability. In these cases, often only processed, output data used to produce the figures in a journal publication were archived. The potential for data reuse and data citation advantage (Piwowar & Vision 2014) were an important motivation for most of

the data depositors.

Overall, the OPeNDAP services offered by the 4TU archive did not seem to have had much influence in the choice of archive for most of the data depositors we interviewed. A minority of the researchers were not aware of OPeNDAP and its functionalities; many knew about OPeNDAP but were just not fully aware that the 4TU archive provided it as a service; others knew about this service but did not consider it important, mostly because their datasets were not big enough for them to care about retrieving data subsets or metadata without the need to download entire datasets.

Many data depositors chose the 4TU archive because it was locally available at TU Delft; the vast majority of netCDF datasets in the 4TU archive originate from TU Delft. Most data depositors chose the 4TU archive after the recommendation of a colleague, supervisor, or data librarian, suggesting that community building efforts may lead to an increase in the use of the 4TU archive.

3.2.2 Use of netCDF and training

For the majority of researchers we interviewed, netCDF is the standard data format and model adopted by their communities and it's the primary data format they use and handle. For a few researchers, netCDF was not a standard in their community. In some cases, netCDF was used out of choice because of its self-describing properties and interoperability; in other cases, it was simply because it was the output format of commonly used models or software packages.

During the interviews, we noticed that some researchers, who would have benefit from the use of OPeNDAP and its functionalities, were not aware of its existence. With a few notable exceptions, we also noticed a general lack of awareness of the importance of metadata, which can be included in netCDF files, and a lack of attention or adherence to metadata standards and conventions.

None of the researchers had had formal training on the use or production of netCDF files. Most of them started using netCDF during their PhD and learned by reading manuals and documentation, through advice from peers and colleagues, and just simply by trial and error. Asked about receiving formal training, there was a general, but not unanimous recognition that there was a need for it, particularly on the research data management aspects of handling netCDF data (e.g. how to include metadata, what metadata to include, conventions and community standards, etc.). Many of the early career researchers we interviewed, mainly PhD students, where enthusiastic about receiving formal and in-depth training. The more senior researchers recognised the need for training, but mostly for PhD students. At their level, they felt that training put too much of a burden on their already busy schedules. That said, because most researchers didn't learn about netCDF in a structured way, they sometimes had gaps in their knowledge. For example, some researchers noted that it took them a while to learn about useful netCDF tools (e.g. Climate Data Operators) and conventions (Climate and Forecast metadata conventions) that were very useful to them and which they wished they had learned about sooner in their careers. In this sense, short training sessions with high-level information about what is possible and available would be welcome even by busy senior researchers.

4. Conclusions

The netCDF data depositors and users of the 4TU archive represent heterogeneous research communities within the Earth sciences. They have different views and attitudes to data archiving and publishing, and store wide-ranging types of netCDF datasets in the 4TU archive. Ensuring that any new and current netCDF services continue to be relevant to these communities will require taking their diversity of needs and requirements into account. A need for training and guidance –

particularly on data management aspects related to documentation, metadata standards and conventions – may be the common thread uniting these communities. This may provide the way forward for the 4TU.Centre for Research Data to build a community of data depositors and users. Ultimately, as noted by Leonelli (2017), well-informed, inclusive, and participatory development of data infrastructures is expected to lead to an

increase in the quality and re-usability of research data.

Acknowledgement

We are extremely grateful to all the researchers who agreed to speak with us for their time and for their invaluable comments and feedback.

References

Cruz, M J, Böhmer, J K, Gramsbergen, E, Teperek, M, de Smaele, M and **Dunning, A** 2018 From Passive to Active, From Generic to Focused: How Can an Institutional Data Archive Remain Relevant in a Rapidly Evolving Landscape? OSF Preprints. DOI: http://doi.org/10.17605/OSF.IO/JGRKB

Cruz, M J 2018 How does a data archive remain relevant in a rapidly evolving landscape: the case of the 4TU.Centre for Research Data. Zenodo. DOI: http://doi.org/10.5281/zenodo.1175238

Gerkema, T and **Duran Matute, M** 2017 Annual mean sea level in the Dutch Wadden Sea 2009-2011. NIOZ Royal Netherlands Institute for Sea Research. Dataset. DOI: https://doi.org/10.4121/uuid:115ef6c5-8c58-4905-91f5-537985fb3b6f

Hellebrand, H 2004 All data measured by discharge meter in Attert basin. TU Delft. Dataset. https://doi.org/10.4121/uuid:0e38dcc8-d524-4abf-ab59-5c9a38075dc3

Hoonhout, B M, de Vries, S and **Cohn, N** 2016 Field measurements on aeolian sediment transport at the Sand Motor mega nourishment during the MegaPeX field campaign. TU Delft. Dataset. DOI: https://doi.org/10.4121/uuid:3bc3591b-9d9e-4600-8705-5b7eba6aa3ed

Leonelli, S 2017 Towards the European Open Science Cloud: Five Lessons from the Study of Data Journeys. Zenodo. DOI: http://doi.org/10.5281/zenodo.1043154

Mezghani, A, Dobler, A and **Haugen, J H** 2016 CHASE-PL Climate Projections: 5-km Gridded Daily Precipitation & Temperature Dataset (CPLCP-GDPT5). Norwegian Meteorological Institute. Dataset. https://doi.org/10.4121/uuid:e940ec1a-71a0-449e-bbe3-29217f2ba31d

Otto, T and **Russchenberg, H W J** 2014 High-resolution polarimetric X-band weather radar observations at the Cabauw Experimental Site for Atmospheric Research. *Geosci. Data J.*, 1: 7-12. DOI: 10.1002/gdj3.5

Otto, T, Russchenberg, H W J, Reinoso Rondinel, R R, Unal, C M H and **Yin, J** 2010 IDRA weather radar measurements - all data. TU Delft. Dataset. https://doi.org/10.4121/uuid:5f3bcaa2-a456-4a66- a67b-1eec928cae6d

Piwowar, H A and **Vision, T J** 2013 Data reuse and the open data citation advantage. *PeerJ* 1:e175. DOI: https://doi.org/10.7717/peerj.175

Rombouts, J and **Princic, A** 2010 Building a 'data repository' for heterogeneous technical research communities through collaborations. In: International Association of Scientific and Technological University Libraries, 31st Annual Conference. Paper 10.

http://docs.lib.purdue.edu/iatul2010/conf/day2/10

Unidata 2018 NetCDF 4.6.1 Available at

https://www.unidata.ucar.edu/software/netcdf/docs/index.html [Last accessed 30 April 2018].

Audit and Certification of Trustworthy Digital Repositories - lessons learned

David Giaretta[1], J. Steven Hughes[2], John Garrett[3], Mark Conrad[4], Terry Longstreth[5], Bruce Ambacher[6], Barbara Sierman[7]

[1]PTAB Ltd, Dorset, UK, david@giaretta.org
[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA, John.S.Hughes@jpl.nasa.gov
[3]Consultant, Colombia, MD, USA, garrett@his.com
[4]NARA, Washington DC, USA, mark.conrad@nara.gov
[5]Data and Information Standards Consultant, Laurel, MD, USA, terry.longstreth@comcast.net
[6]Digital Curation Innovation Center, University of Maryland, College Park,, MD, USA, bambacher@verizon.net
[7]KB, The Hague, The Netherlands, barbara.sierman@kb.nl

There are digital repositories which are responsible for looking after much of the digitally encoded information on which we all depend. In 1996 the Preserving Digital Information report of the Task Force on Archiving of Digital Information declared that “a process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information”. The Roadmap in the first version of OAIS published in 2002 recognised the need for standard to support such a certification process.

Over the past decade the CCSDS Repository Audit and Certification Working group produced two CCSDS standards which became ISO 16363 and ISO 16919. Their importance is that they provide the basis for the ISO audit and certification of Trustworthy Digital Repositories which has been sought for so long.

The former provides a hierarchical list of metrics concerning which an auditor would need to examine evidence in order to make a judgement about the Trustworthiness of a candidate Digital Repository. The latter provided the standard against which the auditors themselves could be judged. ISO 16919 is an extension of ISO 17021. ISO 17021 is the general ISO standard specifying the processes which must be followed by an organisation which audits and provides certification for management systems such as those of candidate trustworthy digital repositories.

In 2017 ISO audit and certification became possible when the PTAB Ltd became the first organisation to be internationally accredited to be able to perform audit and certification based on ISO 17021, ISO 16363, and ISO 16919. The accreditation involved both an assessment of PTAB procedures and its processes by the accreditation body, which also witnessed an audit PTAB performed.

This paper will describe the lessons learned during the accreditation and audit processes about the advantages of the ISO audit and certification processes of ISO 17021, and why it is ideally suited to be the basis for certifying Trustworthy Digital Repositories, as well as the many other things on which our health, wealth and happiness depend.

Comparisons will be made with other proposed methods for evaluating the trustworthiness of repositories. In addition, this paper will describe some of the ideas about updating both ISO 16363 and ISO 16919 raised in the CCSDS/ISO review process.

Keywords

Digital preservation, Audit and certification, Digital Repository

Introduction

Certification of digital repositories can be done in a variety of ways and for a variety of purposes. For example, one may be interested in knowing how good the food in the cafeteria is, and for this one might look at whether the kitchen has good hygiene or whether customers think the food tastes good. Similarly, one might be interested in how quickly a repository responds to queries.

The aim of ISO 16363 [1] is to judge whether the repository will be able to preserve its digital holdings, following OAIS [2] concepts.

As to who can certify a repository, in principle anyone can examine the repository in some way and give a certificate. But would such a certificate be worth anything? Many questions would need to be answered to determine if that certificate has any value. For example, was the examination carried out in a competent, unbiassed way? What was the repository judged on? Are audits consistent?

When the CCSDS Repository Audit and Certification (CCSDS-RAC) Working Group looked at what was needed, the working group decided that the ISO process should be followed, because (1) we rely on ISO audit and certification in all areas of our lives, from medicine to food to automobiles (2) the ISO process involves checking, at every level, repeatedly and (3) the ISO process, through its multi-national agreements, ensures that certifications granted by any ISO accredited auditor are recognized internationally by any ISO affiliated entities. By choosing to follow the ISO process, all the example questions above are successfully and adequately answered.

The following text tries to give a flavour of the process the PTAB group, the members of which were all part of CCSDS-RAC, has followed to help the take-up of ISO 16363 certification.

The journey to accreditation

ISO 16363 provides a set of metrics that help to document and measure the extent to which a digital repository might be expected to be trustworthy over the long-term. However, in order to have an ISO audit and certification process one needs not just the ISO 16363 standard, but also a standard which says how the audit process is carried out and how to decide whether someone is suitable to conduct audits, i.e. who should be accredited to perform audits, and how should this be done.

The standard CCSDS-RAC created for this is ISO 16919, Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories [3].

The ISO process for assessment

There are basically two types of ISO certification, namely one for products and one for what are termed management systems. ISO 16363 concerns the latter because it involves the whole repository organisation, not just a particular piece of software.

The overall guide to audit and certification of management systems is ISO 17021:2015. ISO 16919 essentially provides a small number of modifications to the requirements of ISO 17021 in order to make it suitable for auditing digital repositories adherence to ISO 16363. In particular ISO 16919 specifies the competences which the repository auditors must possess.

ISO 17021 states that: *The overall aim of certification is to give confidence to all parties that a management system fulfils specified requirements. The value of certification is the degree of public confidence and trust that is established by an impartial and competent assessment by a third-party.*

It specifies that there are a number of steps which must be followed, in particular the organisation which has been asked to perform the audit must

(1) determine whether it is able to perform the audit. If so it must
(2) develop the audit programme which must include
    a. An initial certification with these components
        i. Stage 1 – off-site review ofdocumentation - identifying areas of concern that could be classified as a nonconformity during stage 2.
        ii. Stage 2 – on-site review using a defined process to identify nonconformities
        iii. Repository resolves issues

iv. Certification committee makes decision on whether or not to award certificate

b. Annual surveillance audit in year 1 and year 2 after the initial certification
c. Re-certification audit year 3, to begin the cycle again

Nonconformities are defined as things which represent non-fulfilment of requirements. They are classified as Major nonconformities, which are ones that affect the capability of the management system to achieve the intended results, and Minor nonconformities, which do not affect the capability of the management system to achieve the intended results. Stage 1 seeks to identify any areas of concern that could be classified as a nonconformity during stage 2. Since Stage 1 is often conducted off-site based on repository supplied answers and documentation, it is possible that additional non-conformities could be identified during Stage 2 of the audit.

PTAB's accreditation

The same process is used for accrediting audit organisations, indeed it was applied for the accreditation of PTAB. A National Accreditation Body (NAB) appoints suitable assessors who examined PTAB's procedures and documentation (Stage 1) and then observed PTAB while PTAB conducted an audit (Stage 2, termed a Witnessed Audit).

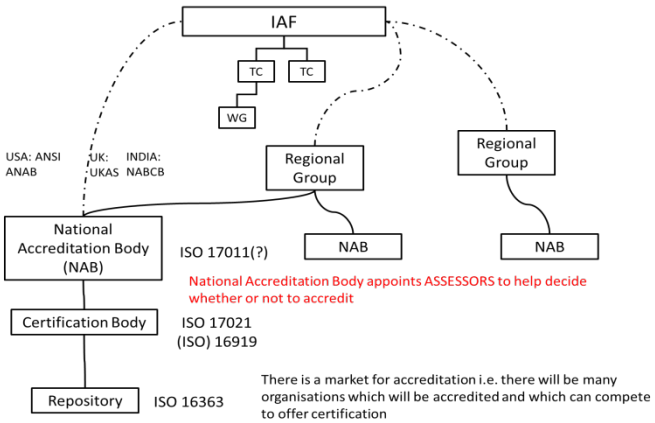


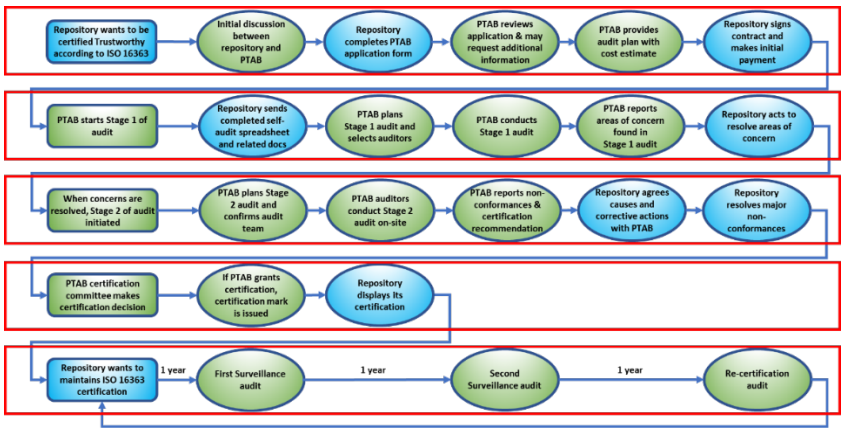
Figure 6 Consistency of ISO accreditation

PTAB's potential auditors were also interviewed in order to assess which should be able to perform and lead audits, and which needed further evidence of their competencies. PTAB members of course have complete command of ISO 16363 but full understanding of ISO 17021 could not be assumed.

Agreements between NABs ensure that they will be consistent in their assessments, illustrated in Figure 1.

PTAB put a great deal of resources into becoming accredited, and at the time of writing is the only organisation which is accredited to perform ISO 16363 audit and certification. The aim in becoming accredited was to prove that there is a significant market for ISO 16363 certification, based on our shared belief that a huge number of organisations have digitally encoded information on which they rely and which must continue to be understandable and usable. ISO 16363 certification is the only way for management and users to be assured that their digital intellectual capital is safe.

PTAB's audit and certification process



To explain the process PTAB has produced a diagram to show the stages of the process, which is essentially a simplified version of a diagram contained in ISO 17021. It is available on the PTAB website http://www.iso16363.org

Figure 7 PTAB process, based on ISO 17021

Lessons learned

During the accreditation process PTAB, guided by the NAB's expert assessors, had to improve its processes and associated documentation, in order to conform to ISO 17021.

In conducting the whole process and particularly the Witnessed Audit it became clear that:

- Stage 1 is vital and must be carried out thoroughly. If a thorough Stage 1 review is conducted, it should be possible to minimise the number of non-conformities found at Stage 2. Possible non-conformities identified during the Stage1 audit can be addressed by the repository prior to the Stage 2 audit. ISO 17021 does not impose a time limit on how long a repository has to address the non-conformities before the initiation of Stage 2 of the audit.
- Stage 2, the on-site part of the audit, is very short. It is constrained by the requirements of ISO 17021 in order to ensure the audit costs are kept under control and that there is consistency between audit organisations. There are a great number of activities to be completed in the time available, so these activities must to be well planned and carried out in an efficient way. ISO 17021 specifies requirements on the communications between the auditors and the repository and the meetings which are to be conducted at the start and end of Stage 2.
- Fixing the issues found in Stage 1 and Stage 2 lead to improvements in the ability of the repository to preserve information, and other opportunities for improvement may be identified.
- The auditors who performed Stage 1 and Stage 2 cannot make the decision as to whether or not to certify the repository. A separate Certification Committee composed of other members of PTAB must do that. In making that decision, the certification committee makes use of reports and evidence collected by the auditors.

While it is clear that perfection is not required, and certainty is hard to achieve, in coming to a judgement the fact that there must be two annual Surveillance audits followed by a re-certification, provides an important context for the decision.

For example, while one cannot be sure of the long term funding of a repository, one can gain reasonable certainty about the funding in the next year. Therefore, in the worst case, scenario, the repository may need to close but at least there should be time to prepare to hand over its holdings, as long as it has prepared its Archival Information Packages (AIPs).

As an added level of reliability, there is also a requirement for a certified repository report to report any material changes in its processes to the auditor and if necessary, the certification can be withdrawn until a follow-on audit verifies continued compliance to the metrics.

Potential updates for ISO 16363 and ISO 16919

At the time of writing OAIS is being undergoing its 5-year review; ISO 16363 will be updated shortly afterwards, in part to ensure consistency between the standards. Since improving the testability of OAIS conformance is a major concern in the OAIS update, this should mean that the updates to ISO 16363 will improve the way in which the audits are performed. ISO 16919 will be updated in the following 2 years.

Benefits of certification

The benefits of conducting audit and certification under ISO accreditation accrue from the fact that the ISO process requires continuous improvements to repositories. The ISO process of accreditation also requires checking everyone and every organisation at every level – repeatedly and consistently around the world.

As such, ISO 16363 certification, by an accredited audit organisation, of a repository is clear evidence that the repository can be trusted to preserve important digital holdings.

References

[5] Audit and Certification of Trustworthy Digital Repositories, 2011, CCSDS 652.0-M-1 and ISO 16363:2012. Available from https://public.ccsds.org/Pubs/652x0m1.pdf

[6] Reference Model for an Open Archival Information System (OAIS), 2012, CCSDS 650.0-M-2 and ISO 14721:2012. Available from https://public.ccsds.org/Pubs/650x0m2.pdf

[7] Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories, 2014, CCSDS 652.1-M-2 and ISO 16919:2014. Available at https://public.ccsds.org/Pubs/652x1m2.pdf
[8] The website for the 5-year review of OAIS and ISO 16363, http://review.oais.info/

Digitizing analogic spectrograms recorded by the Nançay Decameter Array on 35 mm film rolls from 1970 to 1990

Baptiste Cecconi[1,2], Laurent Lamy[1,2], Laurent Denis[2], Philippe Zarka[1,2], Agnès Fave[1], Marie-Pierre Issartel[1], Marie-Agnès Dubos[3] Corentin Louis[1], Pierre Le Sidaner[4] and Véronique Stoll[3]

[1]LESIA, Observatoire de Paris, CNRS, PSL, Sorbonne Université, Meudon, France, [2]Station de Radioastronomie de Nançay, Observatoire de Paris, CNRS, PSL, Université d'Orléans, Nançay, France, [3] Bibliothèque, Observatoire de Paris, CNRS, PSL, Paris, France, [4] DIO, Observatoire de Paris, CNRS, PSL, Paris, France.

Corresponding author: Baptiste Cecconi (baptiste.cecconi@observatoiredeparis.psl.eu)

The Nançay Decameter Array (NDA), which has now passed 40 years old, acquires daily observations of Jovian and Solar low frequency radio emissions over a continuous spectrum ranging from 10 up to 100MHz, forming the largest database of LW radio observations of these two bodies. It also intermittently observed intense radio sources since its opening in 1977. Before that date, decametric observations were conducted on the same site with an interferometer formed of a pair of log-periodic Yagi antennas mounted on mobile booms. These observations have been recorded with a series of analogic recorders (before 1990) and then digital receivers (after 1990), with increasing performances and sensitivities.
The NDA scientific team recently retrieved and inventoried the archives of analogic data (35mm film rolls) covering two decades (1970 to 1990). We now plan to digitize those observations, in order to recover their scientific value and to include them into the currently operational database covering a time span starting in 1990 up to now, still adding new files every day. This modern and interoperable database has virtual observatory interfaces. It is a required element to foster scientific data exploitation, including Jovian and Solar data analysis over long timescales.
We present the status of this project.

Keywords: Data archive; Radio Astronomy; Data at risk

Introduction

The Nançay Decameter Array (NDA), hosted at the *Station de Radioastronomie de Nançay* (Sologne, France), is observing quasi-continuously low frequency radio emissions of the Solar corona and Jupiter (and occasionally other intense radio sources) in the spectral range 10-100 MHz since 1977 [1, 2, 3]. During the preceding decade, decametric observations were already carried on with an interferometer (hereafter referred to as the Nançay Decameter Interferometer, NDI) composed of a pair of log-periodic Yagi antenna mounted on mobile booms [4], whose receivers were originally tested at the Arecibo Observatory [5]. These observations have been acquired with a series of locally developed analogic (NDI/NDA before 1990) and digital (NDA after 1990) receivers with increasing performances. The NDA scientific team recently retrieved the archive of analogic data recorded on a series of 35mm 100ft film rolls covering two decades (1970 to 1990) of observations. We now plan to digitize this data collection in order to favor its scientific exploitation and extend the current NDA database, which already contains all digital data recorded since September 1990. This database is updated daily with new observations. It has been recently reorganized and it now implements modern interoperable interfaces (e.g., virtual observatory standards) [3]. This database will ultimately host the historical digitized data, providing thus a unique data collection spanning on more than 4 decades, which is more than 3 solar cycles and 3 Jovian revolutions around the Sun.

Figure 1. (Left) The Nançay Decameter Array consists in 144 helicoidal antennae divided in two left-handed and right-handed polarized subsets [1]. (Right) Historic log-periodic Yagi antenna in Nançay [2].

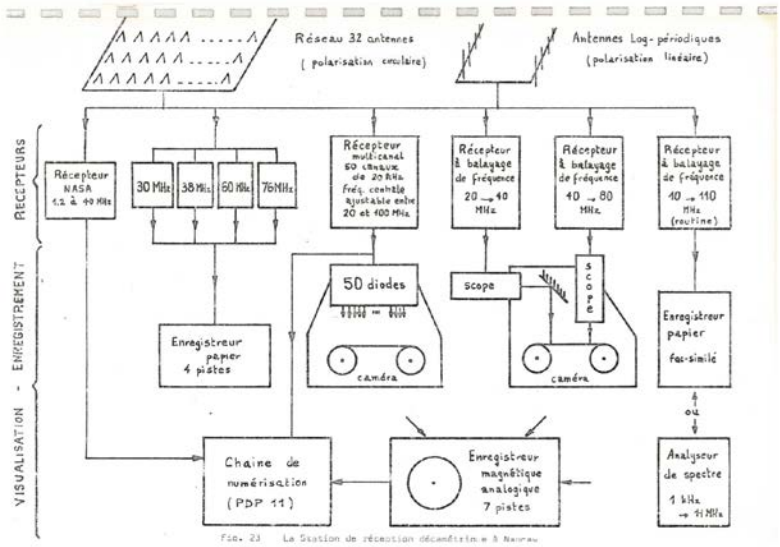


Figure 2. Instrumental setup of NDA and Yagi antennae during the Voyager flybys of Jupiter [3].

Film roll collection

The analogic data collection is composed of 1492 films, which are standard 35mm 100ft rolls. The data was projected in real-time on the roll, with a continuous sliding recording device, usually running at a speed of 3 cm per minutes. Each film thus consists in ~30 meter-length spectrograms recorded continuously on the film roll and generally includes several observing sequences of the Sun and Jupiter. The rolls have been stored in their original metallic canister after chemical development. Most of the film rolls are in healthy conditions, except for a limited series, which have been put in contact with water. This small set shows rusted canisters and glued or damaged photographic gelatin. Most of rolls are numbered with stickers on the film canister. These manual annotations include the name of the receiver, the date(s) of observations and sometimes the observed radio source. Other annotations are reported on the film itself, mostly with hand writing. The film rolls contain time stamps (day of year, hour of day, and decimal number of minutes) and temporal tick marks in the form of 2 dashed lines: a high-resolution line with 1 second consecutive plain and empty segment of equal lengths (30 pairs of plain and empty segments make 1 minutes of data), and a low-resolution line with

1 minute consecutive plain and empty segments of equal lengths (5 pairs of plain and empty segments make 10 minutes of data, i.e., the interval between 2 successive time stamps.

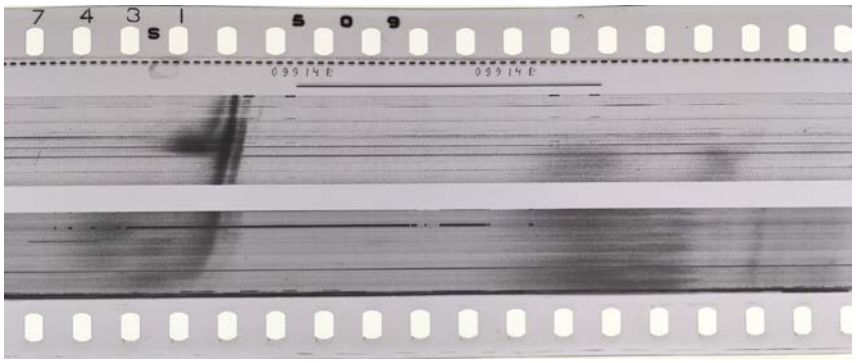


Figure 3. Example of film roll, showing time stamps (day 099 of current year, 14h, start of 10th minute, and long/short 1- minute/1-second ticks. The spectrogram display a solar Type III burst.

Some rolls contain data and annotation recorded on the side perforations, usually due to a recorder misalignment.
Most the data collection contains recording of Sun and Jupiter observations. However, a few early rolls correspond to astrophysical objects observations. A noticeable series concerns solar and pulsar observations led at Arecibo in July and August 1970 [5], including test measurements of the first identified pulsar CP 1919 (PSR B1919+21) discovered by Jocelyn Bell in 1967.

An exhaustive inventory has been built, listing all the metadata available by visual inspection of the canister and roll (by without unrolling it). Precisely, the inventory file so far contains, for each roll: the name of the person filling the file, the date of reporting, the roll number, description and time interval (as written on the canister), its physical location, its preservation state, other observations if any. It will be completed at the digitization step by additional metadata on the roll state and on the carried observations (listing any manual writing and/or markers located on the roll).

Digitization challenges

The continuous recording of the data is a real challenge for the digitization. The companies that are digitizing 35mm films containing moving pictures, i.e., a long series of still images, with a standard known and fixed aspect ratio. The digitization machines have to be adapted for our purpose in two obvious ways: the digitization set up must include a significant overlap between 2 successive snapshots (about 10% of their length); the snapshots must also be “over-scanned” in the direction across the film, in order to digitize annotations and temporal markers. For those two points, it is possible to adjust the existing scanning machines rather easily. Another challenging issue is the geometry of the optical setup. Our data is very sensitive to trapezoidal aberration (i.e. when one side of an image is larger than the other). The trapezoidal aberration is not critical of classical movie scan, as each successive image is scanned with the same aberration, which is barely noticeable if kept small. In our application, a change of a few pixels of length from one end of a single snapshot to the other will prevent us from directly superimposing the successive snapshots during the reconstruction phase. We will need to correct for this aberration and this will result in a blurring of the images due to the resampling process. A last (but not least) challenge is the intensity dynamics of the digitization. The devices used by the industry is limited

to 10 bits recording (e.g., 2^{10} = 1024 levels of gray) of dynamical range. This should be sufficient for our application, if the black and white levels are correctly setup.

The digitizing companies are providing data in movie industry standards: either DPX of TIFF formats. They both contain full resolution uncompressed single images. DPX images can be converted into TIFF with usual image processing software (such as image-magick [4] on Linux). Standard TIFF processing tools are often using 8 or 16 bits of dynamics, so that all files have to be converted (from 10 bits to 16 bits).

The proposed image resolution for digitization is the HD resolution (1920x1080 pixels), with 10 bits per pixel. This results in ~20 MB single files, and each roll is fully scanned with about 1600 snapshots. This results in ~31 GB of disk space for a single film roll, and in ~45 TB for the full collection. The reconstruction of the scientific data will double this data volume. All raw image scans will be archived at the Bibliothèque of Observatoire de Paris (the library department). A collection of quick-look images (JPEG format) will be derived from the raw images, for easier display (e.g., on web pages). The reconstructed scientific datasets will be recorded in a science ready format (such as CDF).

Future steps

We expect to have a first series of digitized rolls by the end of 2018. The assessment of the data quality has to be set up. Concerning the reconstruction algorithm, preliminary testing has been conducted on a series of digitized samples. This first test allowed us to prepare the invitation to tender, reducing the risks. However, we plan to improve the algorithm and test new libraries, such as the photographic stitching software that are commonly used.

The rehabilitation step (transposing digitized raw data into scientific observations accessible to the community at standard format with associated documentation) will require a significant work from the NDA team, which will be carried on in the frame of national observation services activities.

References

[1] Boischot, A, C Rosolen, M G Aubier, G Daigne, F Genova, Y Leblanc, A Lecacheux, J de la Noë, and B M Pedersen. 1980. “A New High-Gain, Broadband, Steerable Array to Study Jovian Decametric Emissions.” *Icarus* 43: 399–407.

[2] Lecacheux, Alain. 2000. “Decameter Array: a Useful Step Towards Giant, New Generation Radio Telescopes for Long Wavelength Radio Astronomy.” In *Radio Astronomy at Long Wavelength*, GM 119:321–28. AGU.

[3] Lamy, L, P Zarka, B Cecconi, K-L Klein, S Masson, L Denis, A Coffre, and C Dumez-Viou (2017). “1977– 2017: 40 Years of Decametric Observations of Jupiter and the Sun with the Nançay Decameter Array.” *Planetary, Solar and Heliospheric Radio Emissions (PRE VIII)*.

[4] Boischot, A. 1974. “Radioastronomy on Decameter Wavelengths at Meudon and Nançay Observatories.” *Sol. Phys.* 36 (2): 517–22. doi:10.1007/BF00151218.

[5] Boischot, A, J de la Noë, M du Chaffaut, and C Rosolen. 1970. “Radiospectrographie Solaire À Haute Sensibilité À Arecibo (Porto-Rico).” *C. R. Acad. Sci. Paris* 272 (B): 166–69.

[6] Leblanc, Y. 1975. “Expérience STEREO-MJS 77” Internal report of Station de Radioastronomie de Nançay.

[7] ImageMagick. http://www.imagemagick.org/script/index.php

Designing DAFNI : a national facilityfor modelling infrastructure

Brian Matthews, Sam Chorlton, Peter Oliver, Ron Fowler, and Erica Yang

STFC Scientific Computing Department
Brian.Matthews@stfc.ac.uk

DAFNI is a major UK national facility to advance infrastructure system research. DAFNI will host national infrastructure datasets and provide a complex hybrid- cloud platform for modelling, simulation and visualisation. We discuss some of the challenges and aims of the DAFNI system, outline its architecture and technical challenges, and summarise its status.

Introduction

The infrastructure systems of a country or region, including energy supplies, water systems, transport routes, digital networks, land use, and the built environment, are subject to environmental (e.g. climate, geology, hydrology), social and economic pressures. Researchers in a variety of disciplines, including environmental sciences, geography, civil engineering, urban planning and economics use computer modelling and analysis of infrastructure to explain and predict the effects of changes, whilst policy makers use the outputs of such models to make planning decisions. Infrastructure systems are becoming ever more complex, and models are becoming more detailed, combining data from different infrastructures and disciplines, and at different scales, from a country or a region down to an individual locality or building. Thus, there is a need for advanced high-performance and high-throughput computing, large-scale data infrastructure to manage and combine data, together with cloud systems for on-demand remote access.

The Data Analytics Facility for National Infrastructure (DAFNI)[1] is a major national facility under development in the UK to provide world-leading capability to advance infrastructure system research. It involves a consortium of 14 universities led by the University of Oxford, and is being developed and hosted by the Science and Technology Facilities Council's (STFC) Scientific Computing Department, as part of the UK Collaboratorium for Research on Infrastructure and Cities (UK-CRIC). It will provide a scalable hybrid cloud platform supporting storage and querying of heterogeneous national infrastructure datasets in a multi-modal architecture, and will support the execution, creation and visualisation of complex modelling applications to enable significant new advances in infrastructure research, and to improve the readiness of research tools and methods for real-world challenges at scale, nationally and internationally. This platform will improve the quality and opportunities for National Infrastructure Systems research whilst reducing the complexity of using data and models for end users.

Motivations and objectives

The Organisation for Economic Co-operation and Development estimates that globally US$53 trillion of infrastructure investment will be needed by 2030 (OECD 2012). In line with this, the UK's National Infrastructure Plan has set aside over £460 billion of investment for the next decade.
However, the impact this investment is hard to predict as projections are underpinned by the quality of the analytics used to inform decision making. Advanced by big data analytics, simulation, modelling and visualisation are now providing the possibility to improve this situation, but there are a number of challenges to be overcome, including:

Distributed teams: analytics are currently undertaken as an isolated activity at disparate institutions with minimal instances of coalescing and collaboration of outputs. However, infrastructure networks and their interactions are inherently complex and heterogeneous, with interactions with people and the environment. Handling this complexity is beyond the capacity of any one team.

[1] https://www.dafni.ac.uk/

Data Heterogeneity: The variety and variability in data has become a limiter for the modelling community and presents a constant hurdle with respect to model collaboration, interoperability and accessibility with extensive subject matter expertise required to exploit each model. Existing data arrangements such as EDINA[2] provide shared academic access to a number of the data sources used in modelling activities, but these represent a subset of the total data landscape.

Maintaining traceability: there is a need to ensure results are reliable and repeatable. It will therefore be essential to store versioned copies of the datasets to support this. This will cause an exponential growth in scale as the range of supported datasets increases and the platform ages.

*Data security***:** licensing of data and models in this field is complex. The difficultly in ensuring that security is maintained presents a barrier to data sharing, and safeguarding the integrity of the data for researchers and data providers represents a key challenge the development of a common platform.

Model granularity: the increase in data availability and resolution has enabled new modelling applications with increasing granularity of modelling applications, with a corresponding increased demand for computational resources. The resource availability limits the ability for modelling activities to understand impacts of simulations at a national scale whilst maintaining fine grain resolution.

Consequently, the shared DAFNI platform is being developed to provide a dedicated compute resource specifically for the National Infrastructure modelling community. It will improve the quality and opportunities for research; and reduce the complexity of all aspects related to conducting the research including data access and processing, model execution, security and visualisation. It will provide specific optimisations to support tasks such as sensitivity analysis and parameter optimisation required by the community. The combination of these facets with a functional platform that addresses the data, licencing and scalability challenges delivers a platform enabling research in areas such as:

- modelling of energy, transport, digital and water networks, for system planning and optimisation;
- real-time data assimilation from sensors and the Internet of Things (IoT) to enable more efficient and reliable operation of infrastructure networks;
- modelling of changing patterns of demand for infrastructure services to enable investment planning;
- modelling of extreme events and their impact upon infrastructure networks to target vulnerabilities and enhance network resilience.

DAFNI Architecture and Capabilities

Although the problem space that DAFNI is addressing is broad and complex, an initial requirement gathering exercise identified core capabilities, as illustrated in Figure 1, and briefly described below.

National Infrastructure Database (NID): a centrally managed access point to national infrastructure and other datasets required to support infrastructure research. This includes: a centrally managed datastore; an Extract Transform and Load (ETL) framework to maintain data currency and interoperability; a data catalogue; and a data access and publication service.

[2] The University of Edinburgh Centre for Digital Expertise (https://edina.ac.uk/)

National Infrastructure Modelling Service (NIMS): support for users to improve performance of existing models, reduce the complexity of creating models and facilitate the creation of complex system-of-systems models. This will include: workflow framework and creator; a workflow engine; a model catalogue and a data transformation library.

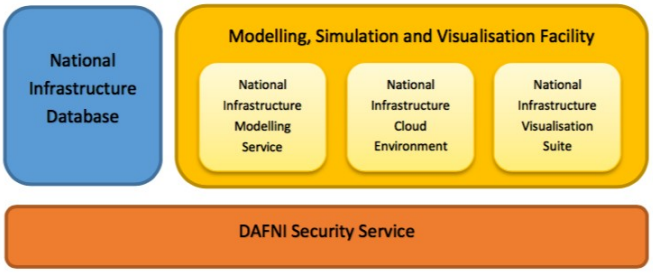


Figure. 1. DAFNI Core Capabilities. Cloud, modelling and visualisation services are grouped into a Modelling, Simulation and Visualisation facility

National Infrastructure Cloud Environment (NICE): a high performance, flexible and scalable hybrid cloud environment with: a cloud services user interface and command line interface; a number of PaaS offerings (e.g. data science notebooks and databases); and a centralised resource pool in support of PaaS and workflow manager services.

National Infrastructure Visualisation Suite (NIVS): a set of high-resolution visualisation tools to facilitate understanding of data, models, outputs and translation of findings to decision makers. This will include: traditional visualisation as a service (e.g. graph and tabular representations); and advanced visualisation as a service (e.g. virtual/augmented reality)

DAFNI Security Service (DSS): tools and processes to manage security aspects of the platform, which allows users to seamlessly access and use services they have rights to without being hindered by the platform, while at the same time maintaining security and integrity of data. Services will include authentication, authorisation, monitoring, and accounts management.

These components will be implemented in a micro-services architecture. This allows the capabilities within DAFNI to be independent with an extensible and flexible delivery of the platform in line with the evolving nature of the National Infrastructure modelling landscape. Following a structured design approach, a hierarchical overview of the platform has been derived leveraging the capabilities and functions outlined as part of the core capabilities analysis. We highlight three key features within the design that present a research challenge for the platform. They are the Central Datastore (CD), Query Manager (QM) and the Model Workflow Engine (MWE).

Central Datastore: The CD is the focal point for the storage of National Infrastructure datasets. The architecture aims to use the multi-modal approach utilising relational, graph and non-relational database technologies to facilitate the data heterogeneity. This is complemented by an object store for access to datasets and a MetadataDB to maintain a record and audit trail of the diverse datasets. This combination of database technologies allows the data to be stored in the manner most suitable and often therefore most performant. This reduces the complexity of ingesting the datasets, and allows models interacting with existing databases to remain predominately the same. Although there are a number of general purpose database technologies[3], these often sacrifice performance or detail in favour or interoperability.

Query Manager: The QM provides a means through which to extract data from the CD and to generate new insights. The QM is attempting to solve the extremely complex task of enabling querying of data

across databases and across database technologies. DAFNI aims to create a new cross database query

[3]An example of this is MongoDB which although a document database is designed as a general database with support for documents, graphs and in part can enable use of relational data (https://www.mongodb.com/what-is-mongodb). manager specifically to address the distinct requirements of its modelling community. This will likely represent one of the biggest implementation challenge for the project.

Model Workflow Engine: System-of-systems modelling is hard because of the required model interoperability. Each model has a unique set of dependencies making coupling extremely time consuming and often an impossible task. To address this, DAFNI is developing the MWE that will utilise containerisation to encapsulate functionality and dependencies whilst maintaining the workflow structure. Each workflow will consist of a series of chained containers characterising each operation with a centralised job manager to handle data collection and data exchange between the containers. This flexibility can allow for more dynamic allocation of resources within DAFNI and allows for use of any operation that can be containerised to be used within the workflows (e.g. data transformation and visualisation).

Current Status

The DAFNI construction programme (2017–2021) is currently in progress, with an early detailed requirements and design phase, and detailed architecture. At the time of writing, it is undergoing its initial implementation and deployment.

As the DAFNI platform is evolving, a series of pilots is validating the functionality and refine the platform requirements. The first pilot focused on the NISMOD-1 *System of Systems* modelling application (Ives, Pant & Robson 2018) developed as part of ITRC-MISTRAL project (Mendes 2016) and hosted at Newcastle University. NISMOD-1 is a collection of codes that currently run on a single machine supporting five models of UK infrastructure: Energy Supply; Water Supply; Solid Waste; Transport; and Waste Water. The models explore the needs of these infrastructure components based on estimates of trends in areas such as population growth, economic growth, and climate change. A key need for NISMOD-I is sensitivity analysis: determining whether the uncertainty of a given input parameter change the "preferred" solution to an infrastructure problem. Without proper understanding of sensitivity, predictions are of limited use. With a large number of input parameters to each of the NISMOD models, a full sensitivity analysis requires running very many simulations while varying each input in turn. This is highly compute intensive and is impractical to run through the existing NISMOD GUI. The first pilot ported the NISMOD-1 system onto the DAFNI cluster and provided a batch processing system to submit multiple sensitivity analysis jobs using HTCondor. As a result, the NISMOD-1 team have successfully run a number of sensitivity analyses on the Water Supply models and achieved a speed up of 10 times over that of the original service.

The NISMOD-1 sensitivity analysis is an example of the benefits that can be derived by moving existing, proven infrastructure models onto a high throughput cluster. The analysis can easily be extended using private and public cloud systems. Moving the data as well as the software to the DAFNI system is key to obtaining scalable performance.

As DAFNI is developed and deployed, it will address the increasing the demand of research capabilities, in order for the UK's national infrastructure research effort to remain at the cutting edge. Further, as DAFNI begins to develop into maturity it could act also as a focus for government and industry, and it is working towards ensuring the expected availability and platform robustness to achieve this goal.

References

Organisation for Economic Co-operation and Development. 2012. Strategic Transport Infrastructure Needs to 2030, OECD Publishing,

Ives, M., Pant, R., Robson, C. 2018 NISMOD: http://www.itrc.org.uk/nismod/#.WrT2ypPFJ24.

Mendes, M. 2016. The UK Infrastructure Transitions Research Consortium: Mistral Programme: 2016

NASA's Earth Observing Data and Information System – Near-Term Challenges

Jeanne Behnke[1], Andrew Mitchell[1] and Hampapuram Ramapriyan[1, 2]
[1]NASA Goddard Space Flight Center
[2]Science Systems and Applications, Inc.
Jeanne.Behnke@nasa.gov

NASA's Earth Observing System Data and Information System (EOSDIS) has been a central component of the NASA Earth observation program since the 1990's. EOSDIS manages data covering a wide range of Earth science disciplines including cryosphere, land cover change, polar processes, field campaigns, ocean surface, digital elevation, atmosphere dynamics and composition, and inter-disciplinary research, and many others. One of the key components of EOSDIS is a set of twelve discipline-based Distributed Active Archive Centers (DAACs) distributed across the United States. Managed by NASA's Earth Science Data and Information System (ESDIS) Project at Goddard Space Flight Center, these DAACs serve over 3 million users globally. The ESDIS Project provides the infrastructure support for EOSDIS, which includes other components such as the Science Investigator-led Processing systems (SIPS), common metadata and metrics management systems, specialized network systems, standards management, and centralized support for use of commercial cloud capabilities. Given the long-term requirements, and the rapid pace of information technology and changing expectations of the user community, EOSDIS has evolved continually over the past three decades. However, many challenges remain. Challenges in three key areas are addressed in this paper: managing volume and variety, enabling data discovery and access, and incorporating user feedback and concerns.

Keywords: Data systems, Earth science, Remote sensing, Big data, Data discovery, Data access, Data preservation, Metadata

Introduction

NASA's Earth Observing System (EOS) Data and Information System (EOSDIS) has been a central component of the NASA Earth observation program since the 1990's. The data collected by NASA represent a significant public investment in research. Consequently, NASA developed a free, open and non-discriminatory policy consistent with existing international policies to maximize access to data. EOSDIS manages data covering a wide range of Earth science disciplines. The data managed by EOSDIS include observations from instruments on board satellites and aircraft, and field campaigns, as well as derived products. The EOSDIS is comprised of partnerships among NASA Centers, other US agencies and academia that process and disseminate remote sensing and in situ Earth science data. One of the key components of EOSDIS is a set of twelve discipline-based Distributed Active Archive Centers (DAACs). Because of their active role in NASA mission science and with the science community, they perform many tasks beyond basic data stewardship, representing a distinct departure from typical data archives. They are collocated with scientific expertise in their respective Earth science disciplines.

Managed by NASA's Earth Science Data and Information System (ESDIS) Project at Goddard Space Flight Center and distributed across the United States, these DAACs serve over 3 million users globally. The ESDIS Project provides the infrastructure for EOSDIS including other components such as the Science Investigator-led Processing systems (SIPS), common metadata and metrics management systems, specialized network and security systems, standards management, and centralized support for use of commercial cloud capabilities. Given the long-term requirements, and the rapid pace of information technology and changing expectations of the user community, the ESDIS Project has had to evolve EOSDIS continually over the past three decades to address many challenges. The purpose of this paper is to describe some of the key challenges and the approaches being taken to address them. In the era of big data, it is important to consider emergent issues in light

of the ongoing challenges that science archives like EOSDIS have addressed and are continuing to address.

Challenges

As a long-lived system that manages data from many diverse sources and serves a multi-disciplinary user community, EOSDIS faces many challenges. These challenges can be grouped into three main categories: 1. Managing volume and variety; 2. Enabling data discovery and access; and 3. Incorporating user feedback and concerns. These challenges are discussed in the three subsections below.

Managing volume and variety

Back in the 1990s, when EOSDIS was conceived it was understood that it would always be a growing collection of Earth science datasets. It started with very small collections that NASA had funded at various locations, which became the DAACs of the EOSDIS. The NASA EOS program was planned to consist of several multi-instrument platforms that would collect data continuously. From the management and funding perspective, it makes sense to have a single system that manages multi-mission operations, as opposed to the old model of creating a new processing/archiving system with each mission.

Since its inception, EOSDIS has added new missions to the Earth science collection expanding the variety and volume every year. With each orbit, instruments continue to acquire data adding to the collection. However, the data also grows as scientists improve the measurements from the instruments deriving new parameters and products. Data formats that were chosen at launch must adopt to meet new standards and feed new software applications reliant on improved metadata. In the 1990s, staff at the ESDIS Project had a difficult time convincing scientist data providers of the value of metadata – assuring them that the most metadata they would have to insert into the complex Hierarchical Data Format – HDF format would be no more than 18 individual fields. Today, metadata is a ubiquitous word – everyone understands the value of it and the EOSDIS metadata model has grown to cover not only data, but services, identifiers, humanizers and so on.

At this time, EOSDIS has over 400 million granules identified in its repository from over 7,000 data collections. The number of providers who are allowed to load data into and delete data from the repository is controlled. The EOSDIS Common Metadata Repository software to manage this is carefully maintained, but open source versions of the software are available along with programming interfaces that allow anyone to access the repository. The ESDIS Project provides an infrastructural software system, Earthdata Search, as a user interface to the repository. Because of the diversification by discipline, the workload in maintaining the EOSDIS collection is shared by the DAACs. Vigilance is still needed as inconsistencies across the collected information become more readily apparent in Earthdata Search and other user interface applications. One way ESDIS manages inconsistencies is to establish an independent review committee composed of metadata professionals. These professionals have scrubbed through the DAAC collections, focusing on metadata, to create targeted reports of errors, misspellings, inconsistencies, etc. This two-year task is expected to improve the user experience in searching through the EOSDIS data collection.

As has been the case over the history of EOSDIS, the diversity of science data producers contributing to the variety and volume of the collection continues to present challenges. Physical storage and hardware challenges are always expected as the collection grows. However, the challenge presented by the diversity of data producers is inescapable. Although we have required standards for data and metadata, like the HDF and ISO19115 (ESDIS 2017) , we do not precisely control the way the standard is implemented by a particular science instrument team or SIPS. This challenge means that with each organization-wide system change, whether for new versions or transferring to new technology, each dataset must be handled individually. An additional challenge now is that several of the original EOS instruments are at end of their life. In the old days, data would be written to tape and racked and users could request the data but would have to figure out how to read the tapes. Since all

data are now online and available, the data are easier to maintain and re-version, but we may no longer have final versions of data. These heritage datasets will always need to be maintained at the DAACs and the ESDIS Project plans ahead to keep them updated to the latest data and metadata standards. In the case of ICESat (2003-2010) data, the DAAC archived the final version of the data several years ago but continues to update the metadata to make it discoverable by users. However, researchers who have better calibration data have reprocessed new versions of the entire dataset and it is difficult for the archive staff to know what to do about these newer datasets. Procedures need to be established for deciding whether they should replace the older versions and distributed by the archive to users.

EOSDIS, like many other space data archives, is looking at the use of the commercial cloud as the next avenue for data storage and services. Instead of managing in-house hardware systems at the DAACs, the use of cloud systems is very appealing. Several prototyping tasks have been undertaken to gauge the effort of managing data in commercial cloud structures (McInerney 2017). The chief effort is to build an infrastructure that allows all of the EOSDIS components to work in a controlled fashion on the various cloud platforms. One challenge is the effort to make certain that we understand the security aspects associated with the use of a commercial system. We are working on a specific security plan that would identify all risks and contingencies associated with the use of a commercial entity. Another issue is the use of various networks to access the cloud systems. Improper use of the networks could increase the cost of cloud use significantly.
Several test efforts are ongoing to document various traffic patterns and usage. In addition, we have prototyped the development of a system, based on existing processes, for ingest and archive of data. This system is undergoing functional and performance tests to work out the many issues that have been encountered. However, one of the greatest challenges will be managing the overall cost of using the cloud by the various components of EOSDIS. Developing the processes for such management are ongoing and proving to be problematic but not insurmountable. We expect that by using the commercial cloud as a platform, the advantages to the user community will be myriad. Researchers will be able to gain new insights into the data and users will enable new applications, which ultimately is the end goal for the big data era.

Enabling Data Discovery and Access

A continuing challenge is to provide users with just the data they need. Typically users search for data using keywords as well as spatial and temporal constraints. In EOSDIS, with thousands of datasets, typical queries from users may result in hundreds of hits meeting their criteria. Ensuring that the most relevant of the datasets appear first in the results list is crucial to users. The obvious steps one can take towards increasing search relevance are ranking the datasets based on spatial and temporal relevance. Also, ranking newer versions higher, and applying information about community usage of datasets (e.g., through automated analyses of scientific literature) for ranking are useful steps. Observation of the real usage of the Earthdata search capability in EOSDIS and characterizing the search and access will also help in continuous improvements to data discovery.

In the case of data that can be represented as images, it is beneficial for users to be able to visualize them and select the data that they want to download and analyze. Enabling this for large volume datasets is a challenge that we have successfully addressed through its Global Imagery Browse System (GIBS) and the WorldView client software (Murphy et al. 2015 a). The GIBS consists of a database of images stored in a hierarchical manner to enable rapid access to data at multiple resolutions.
The WorldView client takes advantage of this data structure and enables users converge within a few seconds on their area of interest at the highest resolution offered by the dataset.

The access to data by users has changed significantly over the last two decades. In the 2000's, EOSDIS data were stored in robotic tape silos. Users would discover what they needed and place on-line orders for data from the respective DAACs. The DAACs would copy data to media and mail them to the users or stage the data on disk and email users so that they could download the data. With the move starting in 2006 to online storage, users now select data granules (files) that meet their search criteria

and are provided with the URLs, which they can use to download the granules. The on-line storage also has enabled the users to request services such as subsetting, reformatting and reprojection conveniently prior to downloading the data. However, as the volume of data is expected to increase significantly in the near future, new challenges arise.

Providing to users data that are ready for ingest into algorithms and for analysis saves them considerable amount of traditional preparatory work, such as downloading large amounts of data, subsetting, reprojection, mosaicking, etc. This idea of "analysis-ready data" is becoming more popular recently, especially with respect to Landsat data (USGS 2018). Of course, to extend this idea to all the Earth science disciplines is a challenge due to the differences in the way different science disciplines deal with data. Defining analysis-ready data for different disciplines and preparing the data to meet their diverse needs would take significant effort, especially in a well-established system such as EOSDIS with hundreds of millions of data files requiring reorganization. The next step is to carefully evaluate typical use cases in different disciplines and prioritize implementation efforts. Also, the large and increasing volumes of data make it impractical for users to download them into their own systems for analysis. Near-archive analysis capabilities, as in the case of archiving data in a commercial cloud environment, will alleviate this problem significantly. The challenges of security and managing costs in the cloud environment are real, and are being addressed as described earlier.

Another challenge in this area is ensuring access to data decades into the future. The data and derived products from NASA's missions are a valuable asset resulting in many important scientific discoveries and influential findings. Therefore, they need to be preserved so that future users are able to discover, ac0cess, read, understand and reuse them. Future users should be able to verify, reproduce or question the science as necessary without having access to the science teams that produced the products. The contents needed to be preserved with the data can be referred to as associated knowledge. ESDIS developed a "Preservation Content Specification" that identifies the classes of content that need to be preserved (NASA, 2011). Similar efforts have been documented by the European Space Agency and the Committee on Earth Observing Satellites (CEOS) Working Group on Information systems and Services (WGISS) (CEOS, 2015) Since then having educated our components, DAACs and SIPS, on preservation of content, especially with regard to instruments at end-of-life. It is important that the broader community also consider the serious issues of long-term data archive and accessibility, so we have worked to encourage the universal comment on ways to preserve data along with adoption of these practices.

Incorporating User Feedback and Concerns

As a system that serves a diverse global community of over 3 million users, EOSDIS receives feedback from them in several different ways. Responding to the diversity of the feedback is a challenge. Users can provide direct feedback including suggestions, problem reports or questions on the webpage http://earthdata.nasa.gov. Each of the DAACs has a user services team responsible for analyzing applicable user feedback and responding to requests for help. Also, each of the DAACs has a user working group (UWG) consisting of science and applications users representing the DAAC's specific discipline(s). The UWGs meet periodically with the ESDIS and DAAC staff members to review the data holdings, tools and services offered by the DAACs and provide advice on priorities and future plans. The ESDIS Project employs an independent organization to conduct an annual survey of users to derive the "American Customer Satisfaction Index (ACSI)". While the ACSI is a number indicating how satisfied the users are, the survey also includes several questions for which users provide free-form answers. The ESDIS Project and DAACs analyze these answers for suggestions for system improvement. In addition, focused efforts have been made within NASA's Earth Science Data System Working Groups (ESDSWG) for user needs assessment.

A related challenge is a concern by users regarding privacy while the system requires them to be registered in order to obtain most of the data and services. As a system that manages data from NASA as well as other non-U.S. partners, EOSDIS must comply with different rules regarding access restrictions and privacy policies regarding collection of information about data users. NASA has had

a free and open data policy for Earth science data since the beginning of the EOS Program in 1990. However, working under agreements for archiving and distributing data from international partners, NASA complied with more restrictive policies regarding charging for data and requiring users to be registered and authorized to obtain the partners' data. Until 2012 NASA did not have a registration system for users to access the data from NASA missions. NASA's "earthdata login" is now used for registering users, but with minimal information needed for registration so that more accurate metrics are collected about numbers and organizations of users, and users can be contacted about new datasets and features offered by EOSDIS. The need for better metrics and services to users is balanced relative to privacy rules.

Conclusions

As a long-lived data system, EOSDIS has faced a number of technological, organizational sociological challenges over the past two decades. It continues to evolve in response to such challenges, but the challenges are not unique to EOSDIS. By sharing our issues and solutions, we look forward to discussions of state-of-the-art solutions and novel data services used in other scientific data archives. It is clear that we are all stepping into the big data era.

Acknowledgements

Jeanne Behnke and Andrew Mitchell contributed to this paper as a part of their duties as employees of NASA. Hampapuram Ramapriyan was supported by NASA contract NNG15HQ01C with Science Systems and Applications, Inc.

References

CEOS 2015 Earth Observation Preserved Data Set Content (PDSC), http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf (last accessed April 26, 2018)

ESDIS 2017 Standards, Requirements and References, https://earthdata.nasa.gov/about/system-performance, (last accessed April 26, 2018)

McInerney, M 2017 EOSDIS Cloud Evolution, https://earthdata.nasa.gov/about/eosdis-cloud-evolution, (last accessed April 26, 2018)

Murphy, K. J. et al. 2015 LANCE, NASA's Land, Atmosphere Near Real-Time Capability for EOS,

Time-Sensitive Remote Sensing, Springer, New York, NY, 2015. DOI: https://doi.org/10.1007/978-1-4939-2602-2_8

NASA 2011 NASA ES Data Preservation Content Spec (423-SPEC-001, Nov 2011), http://earthdata.nasa.gov/sites/default/files/field/document/NASA_ESD_Preservation_Spec.pdf (last accessed April 26, 2018)

USGS 2018 U.S. Landsat Analysis Ready Data, https://landsat.usgs.gov/ard (last accessed April 26, 2018)

Building the Data Management Plan of Observatoire de Paris

Aurélie Kasprzak[1a], Baptiste Cecconi[1b], Hélène Veillard[1a], Karine Thomas[1a], Pierre Le Sidaner[1c], Stéphane Aicardi[1c], Stéphane Erard[1b], Catherine Boisson[1d], Véronique Stoll[1e], Virginie Barbet[1e], Frédéric Saconnet[1e], Carlo-Maria Zwölf[1f], Catherine Muset[1c], Florence Henry[1b], Agnès Fave[1b], Jean Aboudarham[1b], Florent Deleflie[1g]

[1] Observatoire de Paris, CNRS, PSL, France : [a] SRCV (Paris), [b] LESIA (Meudon), [c] DIO (Paris),

[d] LUTh (Meudon), [e] Bibliothèque (Paris), [f] LERMA (Meudon), [g] IMCCE (Paris). Corresponding author: Aurélie Kasprzak (aurelie.kasprzak@observatoiredeparis.psl.eu)

During the last decade, the production of science data increased in parallel with the decreasing cost of digital storage and the increase of data processing and computation capabilities. Science institute have to find a way to manage and preserve this data inflow. Most of the calls of funding agencies now require to provide a description of how the data produced in the project will be managed (archiving, curation, distribution...) and published. This usually takes the form of a Data Management Plan. Funding agency also required more and more to select open source licenses for any production of the project, for instance by enforcing FAIR (Findable, Accessible, Interoperable, Reusable) principles.
Some departments of Observatoire de Paris were identified with needs for setting up data management policies: Informatics Department of the Observatory (DIO), which is hosting scientific computing servers and data storage facilities for the sciences teams of the observatory; Paris Astronomical Data Centre (PADC), which provides interoperable access on data collections produced within the observatory; the library of the observatory. Several science teams (linked with projects funded by EU or space agencies) showed interest as well.

Several actions have now been started: Identification of the various sources of data and data collection in each department of the Observatory; Identification of the needs in term of citation (data collections, artefacts, documents, software...) and licenses; study of possible authoritative delegations (e.g., on DOI attribution, long term preservation...) and to whom; proposing a Data Management Plan template to support science teams when applying for funding. Those actions are all aiming at building a generic Data Management Plan for the Observatory, that would propose rules and practices for preserving, distribution and sharing science products.

The PADC team is deeply involved in data-related international data alliances. This is ensuring that: (a) this study is conducted with up-to-date technologies and concepts, and that (b) the results of the study will be discussed and advertised in those international contexts.

Keywords: Open Data; Data Management Plan; Heliophysics; Planetary Sciences; Astrophysics; Library

Introduction

The Open Science movement leading to the concept of open research requires academic and research institutions such as the Observatoire de Paris (hereafter referred to as OP) to think out a new strategy to deal with these requirements. OP is a leading research and teaching institution, which plays a major role worldwide in all fields of astronomy. Founded in 1667, OP is the largest national research centre for astronomy in France. 30% of all French astronomers work in its five laboratories and its institute. It hosts over 800 researchers and academic staff, and 200 students. Due to its specific fields of research in astronomy, astrophysics and time and frequency, OP has been involved over time in many European and international projects. The European Union through its Horizon2020 (H2020) research and innovation program has been a leader of the open science movement in Europe, prompting OP to build its first data management plan (DMP) in 2014.

Since then, several H2020 projects have requested to work on a DMP. Moreover, there has been a specific need due to the amount of data collected, to build a more generic DMP for OP in general. Laboratories, researchers, the library, the virtual observatory Paris Astronomical Data Centre (PADC), the IT department (DIO) as well as the contracts and technology transfer office (SRCV) all work together on the management of data at OP and more specifically on building its DMP in accordance with the regulations and the FAIR

principles (Findable, Accessible, Interoperable, Reusable). More than a DMP type of document, the working group is confronting ideas and looking for solutions to manage data produced by the laboratories and the administrative departments and to preserve and distribute the data efficiently.

Context and Regulations

The European Commission (EC) has been following the international movement of opening scientific data through its H2020 research and innovation program by promoting open science over Europe. The European Commission promotes open access to research results and introduces open access to data produced by projects founded through the H2020 framework program. The aim is to improve science and innovation by making project results and data accessible to everyone, from the researchers to the public and the civil society.

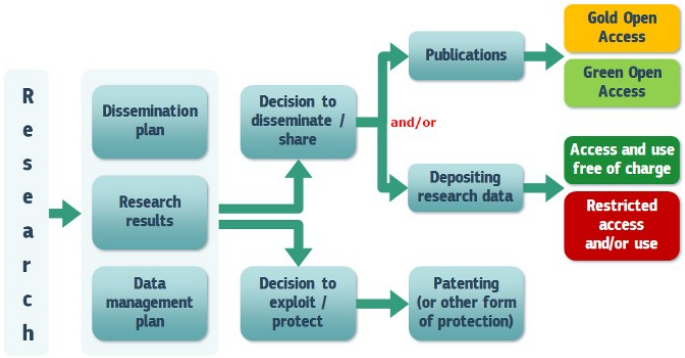


Figure 1. Open access in context: dissemination & exploitation of research results [1]

In addition to providing open access to peer-reviewed publications, EC has enabled access to, and reuse of research data generated by H2020 projects through the Open Research Data Pilot (ORD Pilot) [2]. OP, one of the main actors of data management in the Europlanet-RI-2020 project funded under H2020 research and innovation program, has been working on the project's DMP since 2014. The ORD Pilot project aims to enhance and maximise access to research and science data generated by H2020 projects and to reuse the data, considering the following aspects: the need to find a balance between openness and protection of scientific information; commercialization and Intellectual Property rights; concerns regarding the protection of personal data and privacy; security; preservation and data management issues. EC is also promoting the use of the FAIR principles through the process of writing a DMP. These principles act as guidance to help build strategies for the management of data, its preservation and dissemination.

In France, open access was introduced by the French law on higher education and research of July 22nd, 2013, which modifies article L112-1 of the Code of research. It namely states as one of the objectives of public research "organizing open access to science data" [3].

Research in astronomy produces a lot of data and scientists working in the field have been dealing with the preservation and management of data for a very long time. The international movement towards open access and the EC's will to manage research data in H2020 projects has led various actors of OP to work together on a broader DMP.

The members of this working group are staff from scientific and administrative departments, as listed in the author list. They also participate to the data-related international alliance in the field, such as the International Virtual Observatory Alliance (IVOA), the International Planetary Data Alliance (IPDA), and the Research Data Alliance (RDA). The PADC team is one of the world's most active in the field of astronomical interoperability infrastructures. It collaborates with all other relevant teams world-wide. They have participated to the development of most of the IVOA standards. This context ensures that the recommendations by these world-wide entities are taken into account and the outcome of the study is shared.

Needs for OP

The management of data is not only handled by researchers. When discussing with colleagues from various departments, we realized that it would be more effective to all gather within a working group and to share

and add up our ideas and documents in a joint effort towards a common goal.

The first need for a DMP was expressed by laboratories and researchers involved in a European H2020 project. The contracts and technology transfer office (SRCV) then worked together with scientists to build its first DMP on a European project. Writing a DMP requires complementary skills from IT, archives handling, librarian expertise, science, law, to technology transfer and engineering. This exercise led to further discussions and the will to write a global DMP for the institution.

We first decided to prepare an inventory of the needs of the different actors of the DMP. For instance, the IT department (DIO) needs to deal with the scientific data produced and hosted on its servers. The amount of data created and collected increases in parallel with the decrease of storage costs. This cannot be overlooked because of the financial implications (e.g., for the maintenance and renewal of hardware). The preservation and curation of those large quantities of data needs to be taken into account by OP in its global management plan.

Regarding PADC, the DMP must clarify rules and create internal guidelines for the management of data and metadata produced by the projects of PADC, as well as for their distribution interfaces. Licensing should also be taken into consideration, knowing that EC specifically states its preference for the creative commons licenses. PADC needs to set up a scheme for persistent identifier attribution such as Digital Object Identifiers (DOI) [4]

Considering the needs of the SRCV, the main objective is to build a DMP that is capable of adapting to any project, keeping in mind that its first use is for European funded projects but that the model may be used by any laboratory hosted by OP and for any kind of project. SRCV is also updating researchers on legal developments related to open science and open access.

Current work and Solutions

Elaborating a template of a DMP for EU funded projects is the first step taken by OP. A DMP should describe the life cycle of the data generated or collected within a Horizon 2020 project. The European Commission also provides advice on which type of information should be included in the DMP, such as: the handling of research data during and after the end of the project; what data will be collected, processed and/or generated; which methodology and standards will be applied; whether data will be shared/made open access; and how data will be curated and preserved (including after the end of the project). [5]

In order to do so, the European Commission insists on using the FAIR principles as the core of the DMP. For research data to be FAIR, it must be findable, accessible, interoperable and reusable.

- Making a datum findable means that the data produced or used in the project must be identifiable and locatable. This can be attained from metadata or with digital object identifiers;
- For a data to be openly accessible, the DMP will state where the data and associated metadata, documentation and code are deposited. This can be achieved through a deposition in a repository for example;
- For a datum to be interoperable, the standards or methodologies of data and associated metadata to be implemented by the projects must be described;
- The reusability criterion focuses on the type of license under which the data is protected. A period of embargo may be defined, as it is common for astronomical results but then, the DMP will have to specify the period of embargo and why it is needed.

The European Commission also provides a Horizon 2020 FAIR DMP template. This document assesses the issues to be addressed in the DMP by defining six separate sections. The first section is a data summary. This data summary will be completed by the FAIR principles with its four subsections. Then the question of allocated resources and of data security will have to be assessed. The IT department's expertise is essential for writing the plan for these two components. The fifth section is composed of ethical aspects. Considering the scope of activities at OP, this specific section is usually not applicable. The final section to be addressed is the "other" section in which the participants may refer to other specific procedures for data management, such as national or funding procedures.

The DMP is a tool for managing all kinds of data created in the course of a European project. It should be "as open as possible, as closed as necessary" [5]. It is also a means for discussion amongst researchers, IT

and contracts and TTO departments of the types of results expected during the project and to work out the use of the results and eventually a plan to protect proprietary intellectual property when needed. Projects can lead to results that need to be protected by intellectual property rights. EC accepts that projects opt out of the submission of a DMP in the proposal but encourages participants nevertheless to submit a DMP on a voluntary basis. Several tools such as DMP-Online [6], DMP-Tool [7] or OPIDOR [8] are available to help partners build their own DMP. PADC proposes to use IVOA standards for data formatting, tagging (metadata dictionaries and nomenclatures) and sharing (protocols, registry). The IVOA protocols can be used in a variety of contexts, see, e.g., the Europlanet-RI-H2020/VESPA (Virtual European Solar and Planetary Access) [9] or CTA (Cherenkov Telescope Array) [10] developments on data discovery and provenance technologies.

Discussion

Although there is a lot of documentation and guidelines available for building data management plans and policies, many questions remain open in the case of OP. We list below the current discussion topics of the OP team.

- *Citation and Persistent Identifiers*. Conforming to FAIR principles implies that data is findable and citeable. Libraries have been using ISSN (series) and ISBN (books) as publication identifiers for a long time. DOIs are now used for scientific paper citation. Transition to data collection citation with DOIs started a few years ago. The citation of data through persistent identifiers is being studied by all data centres in the world. The current option is to offer citation of data collections instead of individual data products. OP science teams produce “routine” observation data collections (i.e., the data collection is regularly updated with new data, without modification of former observations). The citation of such collections with static DOIs is under study.
- *Attribution of Persistent Identifiers*. The attribution of DOIs requires a financial participation to DataCite, as well as long term DOI landing pages’ maintenance. OP will have to conduct a trade-off analysis between the cost of hosting and maintaining a local DOI attribution and landing page system (better visibility and independence of OP), and relying on an external DOI provider (e.g., EU-funded or world-wide scale). Such a decision is related to the institutional policy.
- *Personal data*. Recent EU regulations reinforces the personal data protection (GDPR) [11]. Although most of OP activities do not contain personal data, some concerns were raised about the scientist’s personal data in data files, headers and descriptions, e.g., the name of an observer, the location and date of the observation. Such data are part of the observation metadata, and are required for data analysis.
- *Metadata and nomenclatures*. Interoperability means speaking the same language. Data centres and tools need to agree on common terms and descriptors. This applies to all aspects of data description: observatories, instruments, physical quantities, observational parameters and conditions, observed target. The current focus is on the nomenclature of observation facilities (ground based and space borne) in an international effort between IVOA and IPDA, led by PADC.
- *Data Centre Certification*. Data centre certification aims to offer trustworthy contents to scientists. A list of requirements has to be fulfilled [12]. PADC is willing to apply for certification to ensure international recognition, as well as, long term sustainability.

The OP DMP working group aims to deliver a first series of plans and guidelines by the end of 2018.

Acknowledgments

The Europlanet H2020 Research Infrastructure project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654208.

References

[1] EUROPEAN COMMISSION - Directorate-General for Research & Innovation, *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020.* Version 3.2, 21 mars 2017. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020- hi-oa-pilot-guide_en.pdf

[2] European IPR Helpdesk, Making the Most of Your H2020 Project – Boosting the impact of your project through effective communication, dissemination and exploitation. Version EU March, 2018. https://www.iprhelpdesk.eu/sites/default/files/EU-IPR-Brochure-Boosting-Impact-C-D-E_0.pdf

[3] Loi n°2013-660 du 22 juillet 2013 relative à l'enseignement supérieur et à la recherche, NOR: ESRJ1304228L. https://www.legifrance.gouv.fr/eli/loi/2013/7/22/ESRJ1304228L/jo/texte - https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071190&idArticle=LEGIARTI000006524135&dateTexte=&categorieLien=cid

[4] https://www.doi.org/index.html

[5] EUROPEAN COMMISSION - Directorate-General for Research & Innovation, *H2020 Programme – Guidelines on FAIR Data Management in Horizon 2020,* version 3.0, 26 July 2016. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[6] https://dmponline.dcc.ac.uk (UK)

[7] https://dmptool.org (Univ. California)

[8] https://dmp.opidor.fr (CNRS)

[9] Erard, S., Baptiste Cecconi, Pierre Le Sidaner, Angelo Pio Rossi, M T Capria, Bernard Schmitt, V Génot, et al. 2017. "VESPA: a Community-Driven Virtual Observatory in Planetary Science" *Planet. Space Sci.* doi:10.1016/j.pss.2017.05.013

[10] Servillat, Mathieu, Catherine Boisson, Julien Lefaucheur, Johan Bregeon, Michèle Sanguillon, and Jose-Luis Contreras. 2018. "Structuring Metadata for the Cherenkov Telescope Array." *arXiv eprint.* arXiv:1706.06512

[11] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679
Data Archiving Network Service. CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS. https://doi.org/10.17026/dans-22n-gk35

Space Data and Associated Information Long Term Preservation, Discovery and Access

Razvan Cosac[1], Sergio Folco[1], Rosemarie Leone[2] and Mirko Albani[2]

[1] Engineering and Management Support Services for the ESA Heritage Data Programme (LTDP+), Rhea Systems S.A., Frascati, IT
[2] Heritage Data Programme (LTDP+), European Space Research Institute (ESRIN), European Space Agency (ESA), Frascati, IT

Knowledge management practices ensure the identification, capture, organisation, preservation and sharing of core knowledge and information in order to continuously improve an organisation's effectiveness and efficiency in pursuing its mission. The European Space Agency recognizes that knowledge represents the most valuable resource of the organisation, and therefore, that knowledge management represents a crucial aspect when considering the successful completion of the Agency's goals. This paper will give an overview of the European Space Agency's LTDP+ Earth Observation Data Preservation System and will discuss in more detail the ESA LTDP+ Knowledge Management System (KMS), which is composed of the OMNES Platform and the Preserved Data Set Content Management System. These two environments focus on ensuring the long term preservation and discovery of ESA and TPM Earth Observation knowledge and information.

Keywords: associated information; knowledge management; preservation; discovery; access; PDSC; documentation; information; Knowledge Management System; OMNES

1. Introduction

In order to understand the present and to be able to shape the future, we need to know the past. Historical information and knowledge is key to making informed decisions. The European Space Agency has the mandate to assure the long term preservation, sharing and exploitation of space data and its associated knowledge. ESA aims to achieve these goals through the Heritage Data Programme (LTDP+), which is built on four main pillars:

- Preservation – preserve and manage ESA's Space Mission Data and Information
- Discovery – inventory and assure discoverability of all ESA Space Mission Data and Information
- Access – share ESA Space Mission Data and Information
- Value Adding – enhance the value of ESA's Space Mission Data and Information

In order to achieve these objectives an Earth Observation Data Preservation System (EO-DPS) was set up. This system is primarily composed of a Master Archive, a Cold Back-up data archive, and a Knowledge Management System (KMS). The main aim of the EO-DPS is to preserve the EO Mission/Sensor Data Set, which is comprised of the Data Records and the Associated Knowledge, acquired or procured by ESA EO and Third Party Missions (TPM).

Data Records include raw data and/or Level-0 data, higher-level products, browse images, auxiliary and ancillary data, calibration and validation data sets, and descriptive metadata.

Associated Knowledge includes all the Tools used in the Data Records generation, quality control, visualization and value adding, and all the Information needed to make the Data Records understandable and usable by the designated community (e.g. mission architecture, products specifications, instruments characteristics, algorithms description, calibration and validation procedures, mission/instruments performances reports, quality related information). It includes all Data Records representation information, packaging information and preservation descriptive information.

One of the main components of the ESA LTDP+ EO Data Preservation System represents the Knowledge Management System, which is composed of the OMNES Platform and the Preserved Data Set Content (PDSC) Management System. Together, these two environments aim to ensure the long term preservation and discoverability of all ESA and TPM missions' Earth observation space data knowledge and information.

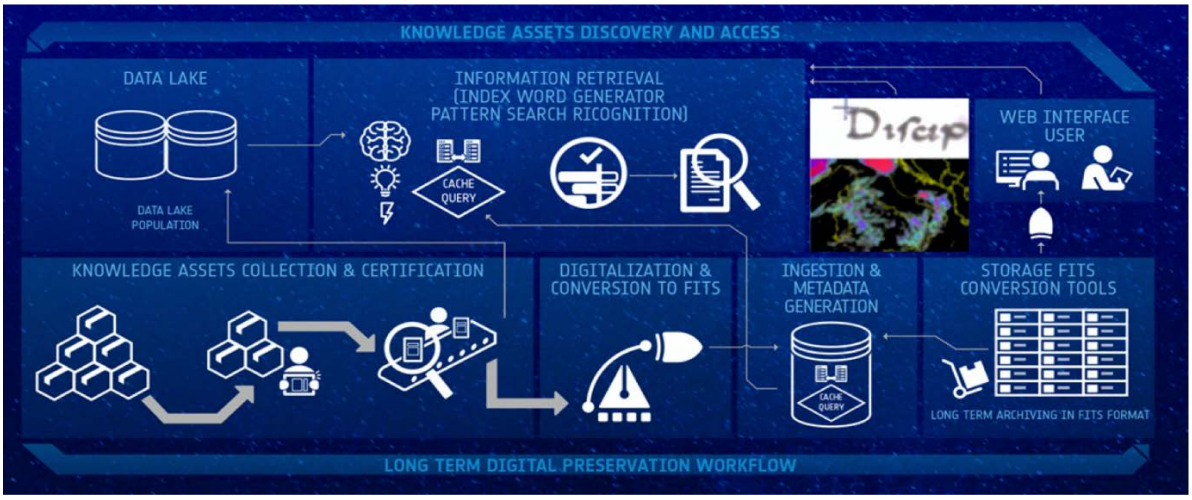


Figure 1 – Knowledge assets long term digital preservation, discovery and access workflow

2. OMNES Platform

The OMNES system is a complex software application used for the preservation, traceability, discovery, and access of digital resources captured in various digital formats. It is designed to adopt an easy, fast and flexible ingestion process for archiving and information retrieval of digital content. The objective of the OMNES platform is to ingest digital information, such as documentation or images, and to preserve it in a digital repository, with an appropriate long term archive format.

The OMNES Platform currently supports the following file formats as input:

- Native PDF (digital text and images)
- PDF with images (e.g. scanned document)
- Images in TIFF, JPG or PNG format
- MP3 audio file (currently only supported in the OMNES ISE system)
- MP4 video file (currently only supported in the OMNES ISE system)

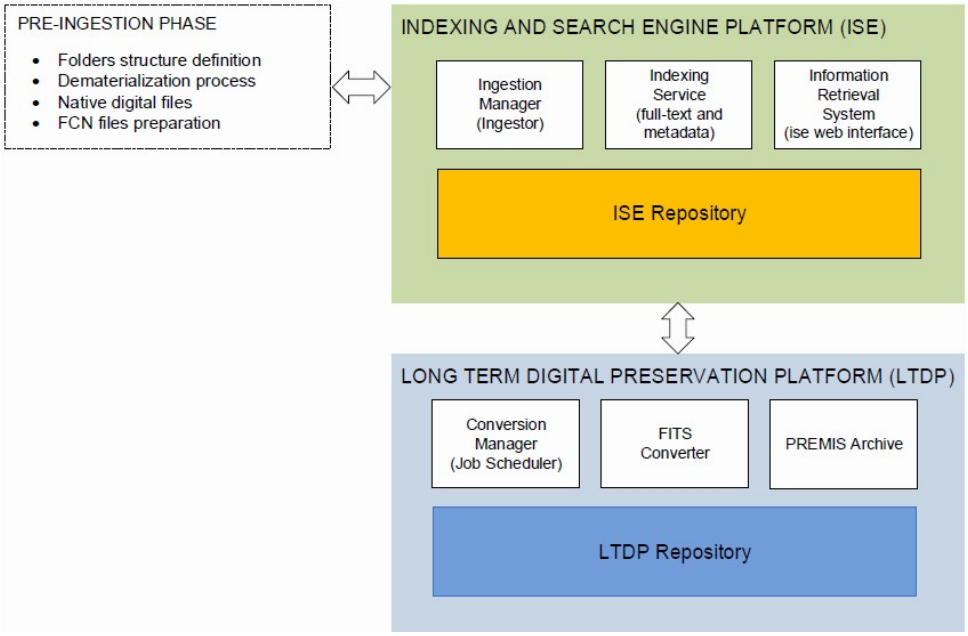


Figure 2 – OMNES Platform high-level system architecture

The OMNES platform is composed of two sub-systems: OMNES ISE (Indexing and Search

Engine) and OMNES LTDP (Long Term Digital Preservation).

OMNES ISE has the following characteristics:

- Conversion of digital content during ingestion process (e.g. searchable PDF/A creation, page previews, etc.)
- Full-text and metadata indexing using Optical Character Recognition (OCR) and Text Extraction processes
- Use of Dublin Core metadata for discoverability
- Web front-end for information retrieval, discovery and access (e.g. search metadata/resource and download)

OMNES LTDP sub-system focuses on:

- Long Term Digital Preservation using FITS (Flexible Image Transport System) file format
- Conversion manager to/from FITS archiving format
- PREservation Metadata Implementation Strategy (PREMIS) metadata model for long term preservation of digitalized documents and images

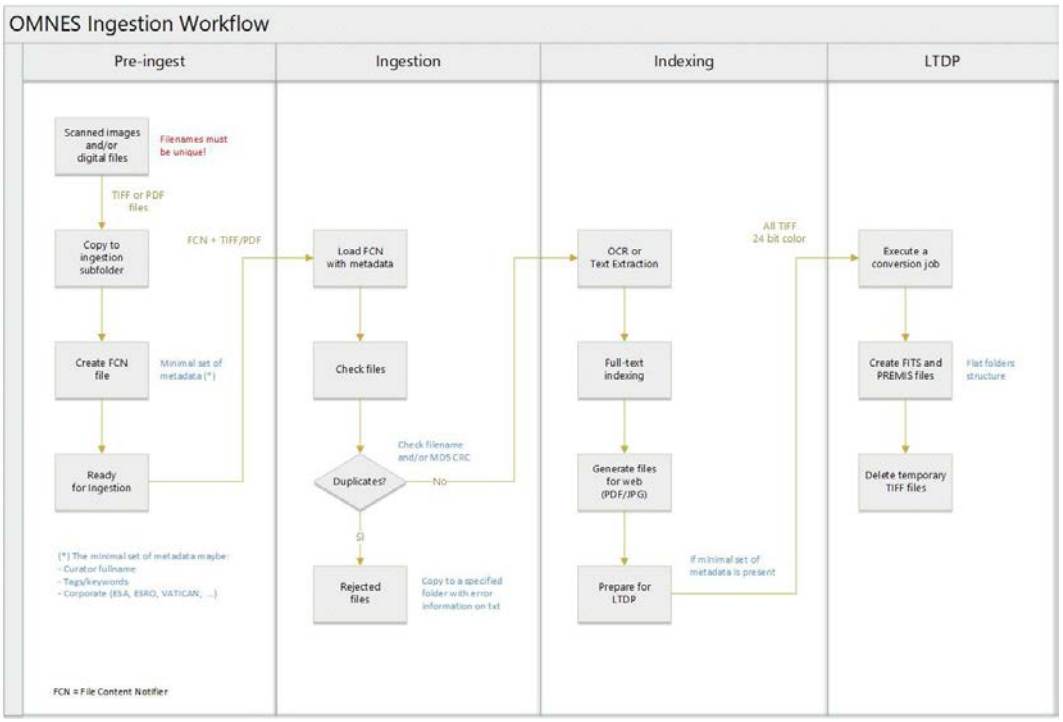


Figure 3 – OMNES Ingestion

Workflow OMNES Ingestion Workflow comprises the following

four stages:

- Pre-Ingestion
- Ingestion
- Indexing
- Long Term Digital Preservation

As part of the pre-ingestion phase, documentation in digital format (e.g. PDF) or scanned images (e.g. TIFF) that are ready for ingestion are copied to the ingestion subfolder by an operator. Here, a File Content Notifier, containing a minimal set of metadata, is created for each digital resource. At this stage, additional Dublin Core metadata can be added by the operator to the FCN file. In the future it is envisaged to have templates used for the generation of documentation, which will facilitate automatic generation of Dublin Core metadata. It will be the responsibility of each document author/owner to fill in the correct corresponding information.

During the ingestion stage, the files are loaded into the system, together with their corresponding FCN files. The system then checks if the digital resource has already been ingested

and either rejects the file, reporting the reason for rejection, or ingests it and indexes the content. As part of the indexing phase, Optical Character Recognition (OCR) and Text Extraction processes occur, and the full text is indexed, allowing a user to search for the resource or its content, through the ISE platform. The OMNES system then prepares the resource for Long Term Digital Preservation. As part of this stage, the digital resources are converted to FITS files and PREMIS metadata is generated. The FITS files together with the PREMIS metadata represent what is preserved for the long term. If, for a certain reason, a digital resource (e.g. PDF document) has to be regenerated, this can be performed starting from the FITS files and PREMIS metadata.

The OMNES platform is aimed at ensuring the preservation, discoverability and accessibility for all ESA EO Space Data Associated Knowledge, in line with LTDP+ objectives. It is initially intended for ESA internal use and will comprise the information generated during each mission phase, from all ESA Earth Observation Programmes.

3. Preserved Data Set Content Management System

The OMNES platform does a good job at managing digital information, such as documentation. However, as part of the Heritage Data Programme, ESA is also interested in managing and preserving the all the knowledge around the data. This is where the Preserved Data Set Content (PDSC) Management System steps in. The PDSC Management System deals with the management of the Earth Observation Preserved Data Set Content. The platform aims to link ESA Earth Observation data records with related documentation and software, in order to be able to manage the Preserved Data Set Content and trace the provenance of the data. The PDSC System has to objective of ensuring the long term preservation, discoverability and stewardship of ESA EO associated knowledge, in line with the LTDP+ Programme objectives.

The PDSC Management System is composed of an open source web enterprise environment, CMDBuild, and an Associated Knowledge repository, Alfresco. CMDBuild allows the management of database of assets so that the organisation may be able to keep under control the situation of data and associated digital resources, knowing at all times the composition, dislocation, functional relations, rules for updating over time, and managing the complete EO mission associated information life-cycle. The metadata used by the PDSC Management System are related to the Dublin Core metadata, while taking into consideration the particular needs for ESA Earth observation information. The digital resource (e.g. document) are ingested together with the corresponding metadata into the PDSC System and they are saved in the Alfresco repository, for discovery and access.

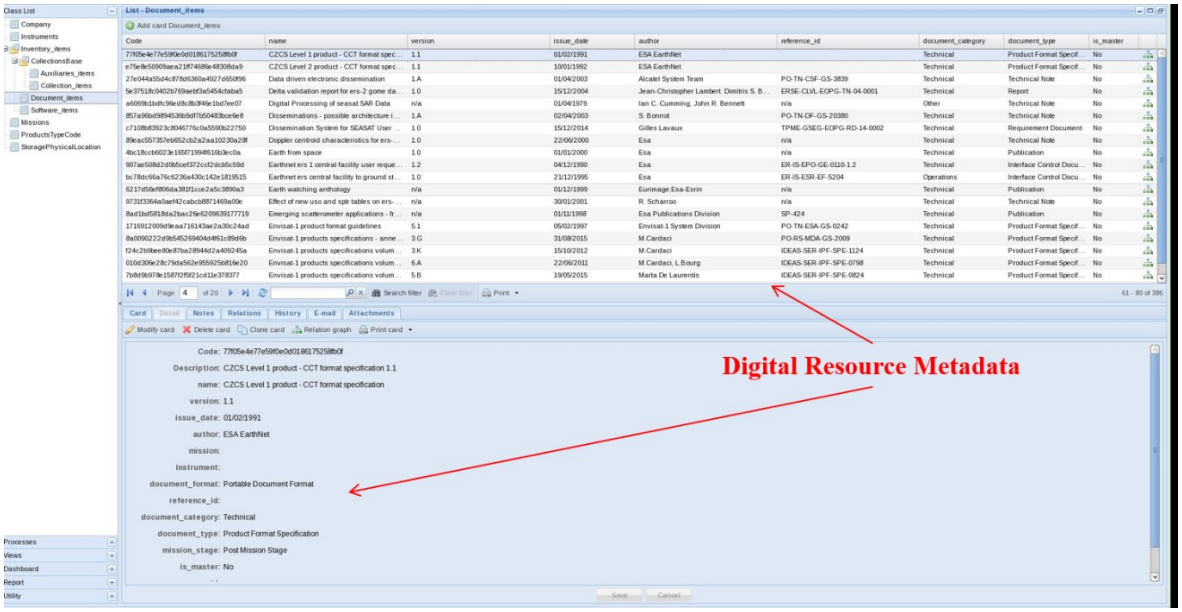


Figure 4 – PDSC Management System knowledge resource metadata representation

The ingested digital resources are then linked to the related data records and software tools, based on the associated metadata. The criteria for ingesting content into the PDSC Management System is given by the tailoring of the Earth Observation Preserved Data Set Content for each Earth Observation mission/sensor/phase. Once the associated knowledge and software tools are linked to the corresponding data records (including their physical location), the data set is defined as master.

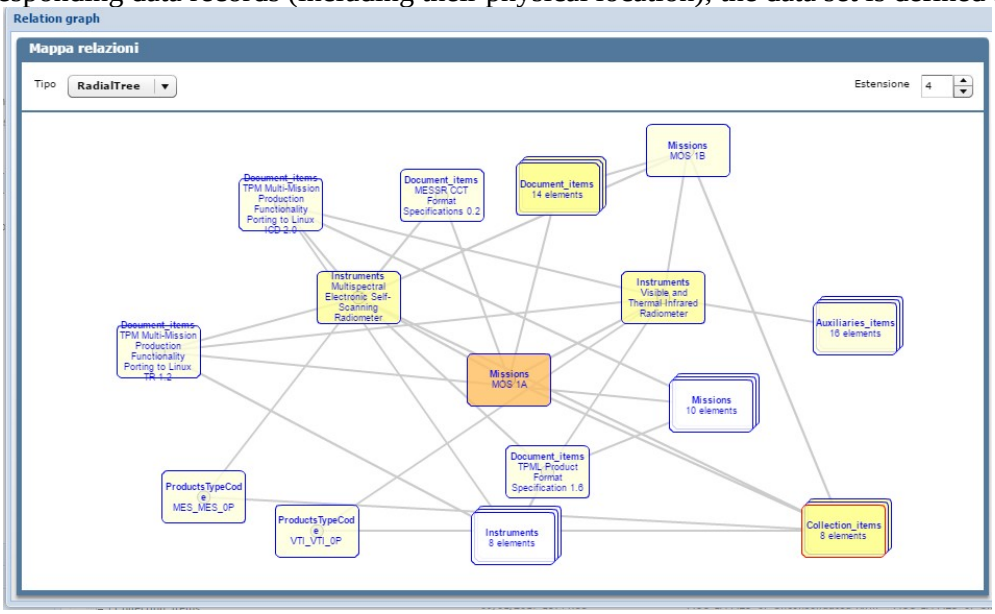


Figure 5 – PDSC Management System diagram showing relationships between data records, knowledge resources and software tools

It is important not to confuse the PDSC Management System with a catalogue. It is a system that enables EO heritage data and information discoverability, insight, exploitation and curation activities. The PDSC Management System facilitates the valorisation of Historical EO Missions Data Sets for long time data series and innovative applications, for multi-disciplinary use, favouring "cross-fertilization" with current and future missions, thus adding value to heritage data.

4. Conclusion

Knowledge and information are the most valuable resources of the European Space Agency, and therefore, knowledge management represents a crucial aspect when considering the successful completion of the Agency's goals. The ESA LTDP+ Programme, contributes to reaching these objectives through the implementation of the Knowledge Management System, ensuring the long term preservation, discovery and access of ESA Space Data and Associated Information.

References

Albani, M., Leone, R., Maggio, I. and Cosac, R. 2015. *Long Term Preservation of Earth Observation Space Data – Earth Observation Preserved Data Set Content.* CEOS-WGISS Data Stewardship Interest Group. [http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf]

ESA Data Preservation Team. 2014. *EO Data Preservation System Scope and Assets.*

ESA Director General's Office. 2017. *ESA Knowledge Management Policy.*

[http://intramedia.sso.esa.int/public/corporate/ESA_KM_Policy_admin-ipol-know-2017-001e.pdf]

Maggio, I. 2017. *Associated Knowledge Preservation Best Practices.* CEOS Data Stewardship Interest Group. [http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS%20Best%20Practices/C EOS%20Associated%20Knowledge%20Preservation%20Best%20Practices_v1.0.pdf]

Embedding Research Data Management Support in the Scholarly Publishing Workflow

Iain Hrynaszkiewicz and Rebecca Grant

Correspondence to iain.hrynaszkiewicz@nature.com

Springer Nature, The Campus, Trematon Walk, Wharfdale Road, London N1 9FN

In 2016 the publisher Springer Nature introduced four standard research data policies for its journals, enabling more journals to adopt a data policy appropriate for their discipline and community. These standard policies have been adopted by more than 1,500 journals, and similar initiatives to standardise journal research data policies have since been introduced by other large publishers. To support researchers and editors Springer Nature launched a Research Data Helpdesk, which, to October 2017 had received more than 300 enquiries. A large survey of researchers with more than 7000 respondents in 2017 revealed that many researchers need support with data management and curation tasks. In 2017, Springer Nature introduced a pilot service to provide additional support to researchers who wish to make their data available alongside their published articles. This Research Data Support service provides hands-on assistance to researchers in uploading their data to a repository, selecting an appropriate licence, enhancing metadata, and cross-referencing the data and its associated publication. The data curation standards were subject to blinded testing by professional editors and curated datasets scored much higher for metadata quality and completeness on average. We describe the implementation of – and lessons learned from providing – a third party data deposition and curation service at a large scholarly publisher, which has been used by authors publishing in journals including *Nature*, and *BMC Ecology*. We conclude with current and future developments, which extended the Research Data Support service to any published researcher, and to research institutions and conferences, providing opportunities to embed research data management support earlier in the scholarly publishing workflow.

Research data policies

In 2016 the publisher Springer Nature introduced four standard research data policies for its journals (Iain Hrynaszkiewicz, 2016), enabling more journals to adopt a data policy – one which is appropriate for their discipline and community. This was in response to evidence (Naughton & Kernohan, 2016) that the research data policies of journals and publishers were found to be confusing for researchers and support staff – and were in need of standardisation. The four policies aim to support researchers in following good practice in the sharing and archiving of research data in accordance with community expectations; facilitate compliance with institutional and research funder requirements to share data, and simplify the journal policy landscape for researchers and support staff. As of April 2018, these standard policies have been adopted by more than 1,500 journals at Springer Nature (the second largest scholarly publisher), and similar initiatives to standardise journal research data policies have since been introduced by other large publishers: Wiley, Elsevier and Taylor & Francis in 2017-18.

Several major publishers are now participating in an international effort to standardise journal and publisher data policies via the global Research Data Alliance (RDA), chaired by representatives from Jisc in the UK, the Australian National Data Service, Springer Nature, and Wiley (Iain Hrynaszkiewicz, Simons, Goudie, & Hussain, n.d.). This interest group in 2018 released for public comment a draft master policy framework, which is ultimately intended to be adopted by all journals and publishers. Policy is known to have an impact on researchers' willingness to share data (Schmidt, Gemeinholzer, & Treloar, 2016) but there is a need to provide ensure policy is implemented effectively (Vines et al., 2013).

Providing support with a Research Data helpdesk

To support the introduction of its standard policies, Springer Nature introduced in July 2016 a Research Data Helpdesk. The helpdesk provides free email-based advice to Editors and authors (researchers), and support staff to help them implement and comply with journal research data policies. The helpdesk is accessible at any point in the peer-review and publication process, and outside of it, although the helpdesk

is not intended to become part of the peer-review process. To October 2017, the helpdesk had received more than 300 enquiries. More than half (53%) of people contacting the helpdesk were researchers (authors); 43% were editors, both professional and academic editors, and the remaining 4% were other stakeholders such as librarians and repository managers. The most common types of queries related to policy implementation (122 queries); data repositories (53 queries); policy compliance (46 queries) and on writing data availability statements (31 queries). A survey of users of the helpdesk found a high level of satisfaction, with 84% of respondents being satisfied or very satisfied with the advice they received (Astell, Hrynaszkiewicz, Grant, Smith, & Salter, n.d.).

Understanding costs

As well as promoting the benefits of data sharing, it is important to understand the costs if research data management support is to be available to a high proportion of researchers. Funding agencies have begun to more explicitly acknowledge and provide funding for research data management in grant awards and guidelines ("Frequently Asked Questions (FAQs) for Public Access (nsf18041) | NSF - National Science Foundation," 2018). There can be increased costs in the publishing process when research data policies are enhanced. Understanding these costs is important for journals and publishers who routinely monitor manuscript processing time and costs, and typically strive to peer review papers rigorously and rapidly. Some types of data policy require additional checks on submitted or accepted manuscripts, which can have associated costs – while also providing benefits (Piwowar & Vision, 2013) – and are important to quantify, particularly for large publishing operations.

At the Nature journals, in 2016, an analysis of the impact of adding data availability statements (DASs) to every article, as part of enhancements to their policies, found that on average increased manuscript processing time by professional editors by 10 minutes (Grant & Hrynaszkiewicz, 2018). The publisher deemed the increased time (cost) reasonable given the importance and benefits of including a DAS, and this cost did not limit implementation of mandatory DASs at all Nature journals.

Where researchers need more support

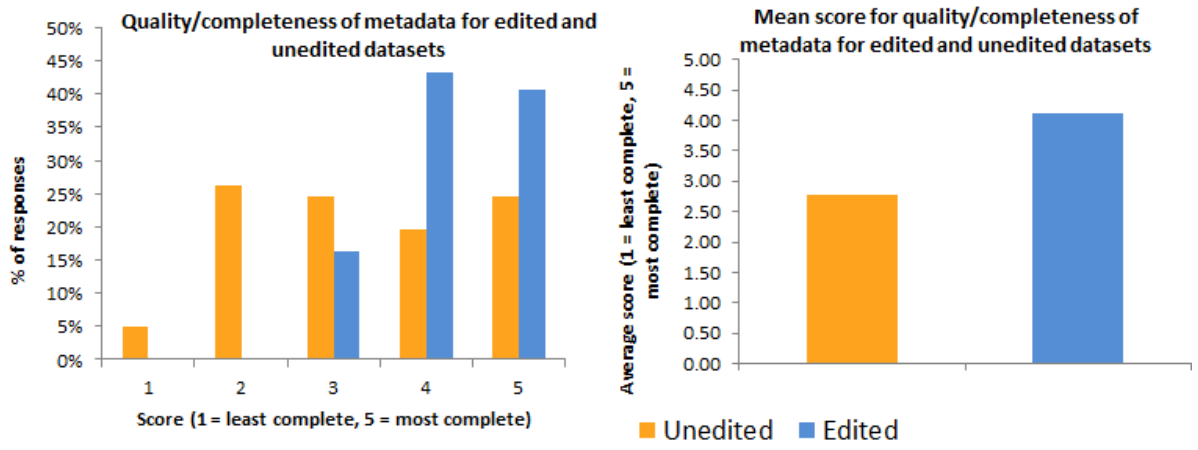
Seeking to understand more about the practical challenges researchers face in sharing research data, in 2018 Springer Nature published a survey of researchers that had received more than 7,700 responses (Stuart et al., 2018). The main barrier to data sharing was identified by respondents as 'Organising data in a presentable and useful way' (46%). Other common challenges were 'Unsure about copyright and licensing' (37%); 'Not knowing which repository to use' (33%); 'Lack of time to deposit data' (26%) and 'Costs of sharing data' (19%). The most commonly reported barrier suggests there may be a lack of skills, support or resources available to researchers curate, share, manage and archive data. Three quarters of researchers also expressed how important discoverability (find-ability) of their research data was to them, rating it highly for importance. Interviews with funding agencies (n = 13) and higher education institutions (n = 15) have also revealed that metadata skills and expertise are a commonly perceived barrier to data sharing.

In 2017, Springer Nature introduced a pilot service (Iain Hrynaszkiewicz, 2017) to provide additional support to researchers who wish to make their data available. This Research Data Support service (https://www.springernature.com/gp/authors/research-data-policy) provides hands-on assistance to researchers in uploading their data to a repository, selecting an appropriate licence, enhancing metadata, and cross-referencing the data and its associated publication. Data are hosted and preserved on Springer Nature's figshare repository. The service was initially available to authors submitting manuscripts to select BMC journals, before being expanded to nearly 100 journals and, from January 2018, being made available to any published researcher and to institutions.

The editorial standards and data curation criteria were developed by, and in consultation with, data curation and data publishing experts and journal editors and were subjected to blinded tested by professional journal editors - many of whom are former researchers. 10 professional editors representing life, social and physical sciences were randomly assigned four datasets each to assess, half (20) of which had been curated according to the standards of the Research Data Support service and half (20) which had not. Responses

were collected via an online survey and the curated datasets scored much higher for metadata quality and completeness on average (rated 4.1 out of 5.0 for overall quality, compared to 2.8 for unedited; Figure 1).

***Figure 1**: Metadata quality and completeness scores for edited (curated) compared to unedited datasets*



The service focuses on curation of metadata rather than directly editing the contents of files, but also intends to guard against the publication of sensitive data and inappropriate file formats, and the use of inappropriate license terms.

Users of the service have included authors publishing in *Nature* (Giles, Xu, Near, & Friedman, 2017a, 2017b), and *BMC Ecology* (J. Rakotoniaina et al., 2017; J. H. Rakotoniaina et al., 2017). Publication and sharing of sensitive data remains a challenge, and is the most common reason for submissions to the Research Data Support service being rejected. Authors' submissions are checked against established guidelines on data anonymisation (I. Hrynaszkiewicz, Norton, Vickers, & Altman, 2010), and authors are asked for details of participant consent for data publication where appropriate, although anonymisation of the data themselves ultimately remains the authors' responsibility. Authors are advised of alternatives to publicly sharing non-anonymised data, such as repositories that provide managed or safeguarded access - such as the UK Data Archive.

Integration of research data support services into the traditional scholarly publishing workflow and systems provides opportunities to increase data sharing, and visibility to researchers of the available support, but also presents challenges. Research Data Support is currently available as an integrated service to selected journals using the Editorial Manager submission system. Testing and development of the integration is ongoing, as with other manuscript submission systems, and has revealed that visibility and engagement of users with the service via this route could be improved. Supporting authors whose associated manuscript is rejected from their initial target journal was also an important workflow consideration. The service provides three options where authors' associated manuscript has been rejected (and will often be sent to another journal for consideration): The curated data can be transferred to a personal figshare account; the data can be deleted; or the data can be held privately for fixed period. Key to provision of the level of curation offered (details at https://www.springernature.com/gp/authors/research-data-policy/help-faqs/15369356 'What is involved in data curation?') is access to the associated manuscript or documentation, to provide a record for the dataset that links to the published article but which is also understandable without the article.

Current and future developments

While the Research Data Support service ended its pilot phase in early 2018, and introduced fees to ensure sustainability of the service, different use-cases and workflows continue to be tested and developed. These include the provision of research data support to conferences, such as the International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS;

https://springernature.figshare.com/tacas). This collaboration led to a higher level of uptake by authors compared to current journal based workflows.

Research data support provision by scholarly publishers is a new development and traditionally researchers may seek support via dedicated teams within their institutions, seeking advice from their funders, or from external expert agencies and services such as Jisc or the Digital Curation Centre or by working directly with specialist data repositories – where and if they exist for a given a community. Not all researchers have access to these services, however, and researcher support for data sharing will likely remain a mixed economy for the foreseeable future. The best way to support a researcher on their specific situation varies from case to case and may need to consider applicable funder and institution research data policies. Research Data Support from publishers can potentially complement institutional services, for example where institutional services are not able to scale or provide sufficient flexibility for the volume of research data outputs of their researchers. Working directly with institutions and funding agencies and with researchers earlier in the research life cycle than the point of publication in a peer-reviewed journal also provide opportunities to further embed research data management support in the scholarly workflow.

Data availability

Data underlying the analysis of helpdesk queries are not publicly available due to the inability to suitably anonymise these customer data, and as consent was not obtained to publish the queries. The data are available on reasonable request from the authors and via researchdata@springernature.com. The data supporting the plots in figure 1 will be made available publicly in figshare with the identifier https://doi.org/10.6084/m9.figshare.6200357 (Smith, Grant, & Hrynaszkiewicz, n.d.) upon publication and are available to editors and reviewers before publication.

Acknowledgement

The authors thank Mr Graham Smith, Springer Nature, for support with production of figure 1 and preparation of the data supporting the figure.

Conflict of interest statement

Both authors are employees of Springer Nature.

References

Astell, M., Hrynaszkiewicz, I., Grant, R., Smith, G., & Salter, J. (n.d.). Have questions about research data? Ask the Springer Nature Helpdesk. https://doi.org/https://doi.org/10.6084/m9.figshare.5890432.v2

Frequently Asked Questions (FAQs) for Public Access (nsf18041) | NSF - National Science Foundation. (2018). Retrieved April 30, 2018, from https://www.nsf.gov/pubs/2018/nsf18041/nsf18041.jsp#q46

Giles, S., Xu, G.-H., Near, T., & Friedman, M. (2017a). Fukangichthys: CT scan data and surface files from middle Triassic fossil scanilepiform fish. figshare. https://doi.org/https://doi.org/10.6084/m9.figshare.c.3814360

Giles, S., Xu, G.-H., Near, T. J., & Friedman, M. (2017b). Early members of "living fossil" lineage imply later origin of modern ray-finned fishes. *Nature*, *549*(7671), 265–268. https://doi.org/10.1038/nature23654

Grant, R., & Hrynaszkiewicz, I. (2018). The impact on authors and editors of introducing Data Availability Statements at Nature journals. *bioRxiv*, 264929. https://doi.org/10.1101/264929

Hrynaszkiewicz, I. (2016). Promoting research data sharing at Springer Nature. Retrieved from http://blogs.nature.com/ofschemesandmemes/2016/07/05/promoting-research-data-sharing-at-springer-nature

Hrynaszkiewicz, I. (2017). New services to support open research. Retrieved April 30, 2018, from https://blogs.biomedcentral.com/bmcblog/2017/04/27/new-services-to-support-open-research/

Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., & Altman, D. G. (2010). Preparing raw clinical data

for publication: guidance for journal editors, authors, and peer reviewers. *BMJ*, *340*(jan28 1), c181–c181. https://doi.org/10.1136/bmj.c181

Hrynaszkiewicz, I., Simons, N., Goudie, S., & Hussain, A. (n.d.). Research Data Alliance Interest Group: Data policy standardisation and implementation. Retrieved April 30, 2018, from https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation

Naughton, L., & Kernohan, D. (2016). Making sense of journal research data policies. *Insights the UKSG Journal*, *29*(1), 84–89. https://doi.org/10.1629/uksg.284

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Rakotoniaina, J. H., Kappeler, P. M., Kaesler, E., Hämäläinen, A. M., Kirschbaum, C., & Kraus, C. (2017). Hair cortisol concentrations correlate negatively with survival in a wild primate population. *BMC Ecology*, *17*(1), 30. https://doi.org/10.1186/s12898-017-0140-1

Rakotoniaina, J., Kappeler, P., Kaesler, E., Hämäläinen, A., Kirschbaum, C., & Kraus, C. (2017). Capture-mark-recapture data modelling survival rates of Microcebus murinus in relation to glucocorticoid level, parasite infection and body condition. figshare. https://doi.org/https://doi.org/10.6084/m9.figshare.5259415

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. *PloS One*, *11*(1), e0146695. https://doi.org/10.1371/journal.pone.0146695

Smith, G., Grant, R., & Hrynaszkiewicz, I. (n.d.). Quality and completeness scores for curated and non-curated datasets. figshare. https://doi.org/https://doi.org/10.6084/m9.figshare.6200357

Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., & Astell, M. (2018). *Whitepaper: Practical challenges for researchers in data sharing*. https://doi.org/https://doi.org/10.6084/m9.figshare.5975011.v1

Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., … Yeaman, S. (2013). Mandated data archiving greatly improves access to research data. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, fj.12-218164-. https://doi.org/10.1096/fj.12-218164

Posters

1. EUFAR Flight Finder & CEDA Satellite Data Finder

Wendy Garland, Ag Stephens and Richard Smith

Searching for data collected on moving platforms such as research aircraft can be difficult – unless you were involved in the flight planning you are not likely to know where or when a flight took place, or what instruments were operated. The EUFAR Flight Finder tool (EFF) is a geospatial/ temporal/keyword search tool developed during the EUFAR2 (The EUropean Facility for Airborne Research in Environmental and Geo-sciences 2014-2018) FP7 project to maximize discovery, access and re-use of data collected by the many European environmental research aircraft – both atmospheric in-situ and remote-sensing measurements - during EUFAR funded flights whose data are stored in the EUFAR data archive at CEDA.

The capability of this flight-finding tool has been widened to include the full catalogue of flights from the FAAM Bae-146 aircraft and the NERC-Airborne Research Facility (formerly ARSF), also stored in the CEDA archive, and now covers 120TB of data and nearly 2000 flights.

Following the success of the EFF, the concept was adapted and reconfigured to facilitate the discovery of satellite data in the CEDA archives in the CEDA Satellite Finder. This enables geographic and temporal search capability for 2 million Sentinel 1, 2, 3 and Landsat 5, 7, 8 scenes.

2. Supporting large scale, iterative metadata enhancement and delivery with a persistent, distributed, event streaming platform

David Fischman, Evan McQuinn and Nancy Ritchey

Unsurprisingly, the rate at which data flow into NOAA's National Centers for Environmental Information (NCEI) is ever increasing, and so too is the volume of metadata to support stewardship, discovery and access of those data. At the same time, the logic used to produce high quality metadata is increasingly varied across data sets and is also always evolving along with both the technology and the standards. The multifaceted challenge that NCEI faces is applying the unique processing steps to the data in order to obtain the prescribed metadata, and storing that information in an agnostic format to support the different standards and downstream uses. Having rich and accurate metadata is required for populating downstream systems supporting customer search, discovery and access. In addition, the metadata supports internal analytics, stewardship, lineage, and data management. By adopting a few modern tools developed by organizations who already attempted to solve this issue, and embracing an architecture based on distributed event sourcing and processing, NCEI aims to build a scalable platform to handle the flood of metadata required for downstream processing and data management and to provide accurate and consistent representation for viewing internally and externally. This poster will present the intended framework to support NCEI's metadata needs while enhancing our customers' experience.

3. Facilitating Accessibility and Exploitation of Historic AVHRR Products

Gina Campuzano, Matthias Hofmann, Torsten Heinen and Katrin Molch

The German Satellite Data Archive (D-SDA) provides data archiving and access services in the context of space-borne Earth observation mission and ensures long-term data preservation (LTDP) of the payload data and derived spatial information products. The archived data are accessible to a variety of users and communities via established, largely automated ordering processes including online data delivery. To keep pace with technological progress, D-SDA also develops and applies new technologies and tools for facilitating data access and use, with options for directly analyzing Earth observation data without the need to download products or to deploy specialized software.

The EOC Geoservice is the spatial data infrastructure of the DLR Earth Observation Center (EOC). The EOC Geoservice provides interoperable spatial data services for efficient data discovery, visualization, and direct download based on the standards set by the Open Geospatial Consortium (OGC). A selection of

historic and current geospatial data hosted by the D-SDA have been integrated into the EOC Geoservice with more data being added continuously. These state-of-the-art services facilitate data interoperability and exploitation.

D-SDA has been archiving the data of the Advanced Very High Resolution Radiometer (AVHRR) on board of the National Oceanic and Atmospheric Administration's (NOAA) fleet of Polar Operational Environmental Satellites (POES) since 1983. For three decades DLR has been generating and archiving regional coverages of daily AVHRR level 3 products – vegetation index, sea surface temperature and land surface temperature. D-SDA is now making a subset of its AVHRR level 3 data archive available via geospatial web services provided by the EOC Geoservice. This will demonstrate the value of online data services to facilitate exploitation of valuable time series of data otherwise 'buried' deep in a preservation infrastructure. Additionally, integrating these data into the EOC Geoservice infrastructure serves as a prototype for the DLR 'TIMELINE' project. As a 'mission to archive' TIMELINE will consolidate and re-process the entire D-SDA AVHRR data holdings into a range of AVHRR level 3 products. In the future these products will also be made available via the EOC Geoservice geospatial web services.

Integrating more and more of its historic data into spatial data services, D-SDA is building a bridge between preservation and value adding. Combined with open data policies, it simplifies exploitation and trend analysis and makes valuable Earth observations accessible to researchers and non-expert users alike.

4. Integration of multiple sources on SELENE HDTV archives

Yukio Yamamoto and Rie Honda

Japanese lunar orbiter SELENE known as Kaguya carried the high definition television system, HDTV. We released the movies of the moon obtained by SELENE HDTV through the Internet for outreach purposes.

However, we determined to archive these movies as well as the other scientific instruments because these movies were precious from the viewpoint of science. We had proceeded data preparation for several years. The data archives were developed and opened in September 2016 to the public.

The data format is a hybrid format using astronomical standards: FITS and planetary data standards: PDS. A movie is divided into still images. Each image includes an observation condition such as observation time, geometry information, movie title, etc.

During the creation of data archives, we integrated multiple sources such as command log, housekeeping telemetry, numeric calculation, and manually input information. All the information was registered into the relational database management system, and it was used both for the meta-data creation and the backend for web services.

We introduce the details of the integration process and utilization of the information.

5. Analysis Ready Data to support the EVER-EST Virtual Research

Iolanda Maggio, Rosemarie Leone, Mirko Albani, Simone Mantovani, Federica Foglini and Francesco De Leo

EVER-EST provides the means to overcome existing barriers to sharing of Earth Science data and information allowing research teams to discover, access, share and process heterogeneous data, algorithms, results and experiences within and across their communities, including those domains beyond Earth Science.

The main components of the EVER-EST Virtual Research Environment are: Presentation Layer, the element that provides the user interfaces and all the technologies that shall guarantee the availability of those services and functions (VRE portal, ROHub, Collaboration Sphere).

Service Layer that provides both generic VRE services and Earth Science specific services. These components represent the reasoning engine of the e- infrastructure and actually orchestrate and manage the services available to the VRE final users.

Central to the EVEREST approach is the concept of the Research Object (RO), which provides a semantically rich mechanism to aggregate related resources about a scientific investigation so that they can be shared together using a single unique identifier. Although several e-laboratories are incorporating the research object concept in their infrastructure, the EVER-EST VRE is the first infrastructure to leverage the concept of Research Objects and their application in observational rather than experimental disciplines.

Data Layer that references the data holdings made available to the VRCs: data is linked and proper means are provided, where feasible, to access it from the VRE.

As a default setting, data will not be copied or duplicated, but will continue to reside on the provider's local servers unless it is directly retrieved by the user. The Data Layer relies on interoperable OGC standard services (i.e. OpenSearch, Web Coverage Service) and permits the integration with Big Data services: the integration with EO Data Service (https://eodataservice.org) enables the provision of Analysis Ready Data (ARD) and makes quicker and easier to explore a time series of images stored in multidimensional geospatial datasets.

The EVER-EST e-infrastructure is validated by four virtual research communities (VRC) covering different multidisciplinary Earth Science domains including: ocean monitoring, natural hazards, land monitoring and risk management (volcanoes and seismicity).

In the framework of the current work the case how an Analysis Ready Data (ARD) service supports the studies "EVALUATE HOW HUMAN ACTIVITIES CAN CAUSE POSIDONIA MEADOWS REGRESSION" and "CROSS-FERTILIZATION BETWEEN JELLYFISH OUTBREAKS & ANOMALIES DETECTION IN THE MEDITERRANEAN SEA" is described.

6. Bit preservation processes in the Centre for Environmental Data Analysis Archive

Sam Pepler

The Centre for Environmental Data Analysis (CEDA) has an archive of over 6PB of Climate science and Earth Observation data used principally by the UK research community. There are a number of systematic processes that the archive performs to enable preservation at the bit level. This poster describes these processes and some of the issues that have been seen as they have been performed in the last 20 years.

Audit of the archive content it is important to make sure that any changes to the data are expected and not the result of IT systems corrupting data files. Very occasionally data files have been seen to change their content with no corresponding update of the modification time. These corruption events are systematically looked for within the CEDA archive so that timely recovery from backups can be made.

Migration to new media is done to minimise wasted space, ensure media are stable, and to take advantage of new technologies. The migration process often highlights problems of the legacy media and can prompt questions about the value on the data undergoing migration.

Currently the master copy of archive data is principally on disk, with secondary copies held in a tape archive system. Pressure on spaces has forced CEDA to move to a model where some data is only held on tape. We are also investigating object store technology, particularly for climate model data. This new mixed media environment is forcing us to change how we manage our archive.

7. Scientific Information Retrieval and Integrated Utilization System

Marina Ohara, Masahiro Ukebe and Yukio Yamamoto

In ISAS (Institute of Space and Astronautical Science)/JAXA (Japan Aerospace Exploration Agency), we use a unique system to preserve and provide telemetry data obtained by the operations of scientific satellites and space probes for a long time.

The system is called SIRIUS (Scientific Information Retrieval and Integrated Utilization System) and has been operating for more than 20 years as an infrastructure of Japanese scientific satellites and planetary explorations. The data generated by instruments onboard a spacecraft is variable length data usually called CCSDS space packet.

The space packets are multiplexed into fixed-length transfer frames when spacecraft transmits telemetry to the ground.

SIRIUS can provide data in both CCSDS space packet and transfer frame data formats in response to needs of users.

Moreover, one of the characteristics of SIRIUS is time calibration of packets. The packet generation timing is embedded in the secondary header of each packet as a counter value called Time Indicator (TI).By converting the counter value to UTC (Coordinated Universal Time) and adding it into the protocol header area, SIRIUS enables users to acquire data both at the earth received time and the packet generation time. Furthermore, one of the essential functions of SIRIUS is the sort function and merge function for packets. Packets made on spacecraft may exist as the multiple redundant packets when processing on the ground. Also, the order may change as compared with the generated order when packets arrive on the ground. These phenomena can occur when simultaneous reception is performed by a plurality of ground stations or when a data recorder is reproduced.

SIRIUS provides merge data to users after combining duplicate packets and sorting the packets in order of creation time. Users do not need to know which ground stations or bands were used during operations, but users are required to remember when their instruments have acquired data.

Currently, SIRIUS is an indispensable system for both spacecraft operation and data analysis.

In this presentation, we report on the details of SIRIUS.

8. Integrated Space and Ground Based FY-4A Satellite Data Service System

Zhe Xu, Di Xian and Yonggang Qi

China has successful launched a new generation of geostationary meteorological satellites called FengYun-4 (FY-4) on 11, Dec, 2016 Following upon the current FY-2 satellite series, FY-4 will carry four new instruments onboard; they are the Advanced Geosynchronous Radiation Imager (AGRI), the Geosynchronous Interferometric Infrared Sounder(GIIRS), the Geostationary Lightning Mapping (GLM), and the Space Environment Package(SEP). The first satellite of the FY-4 series will be experimental with the following satellites being operational.

The main objectives of FY-4 series are to monitor rapidly changing weather systems and to improve warning and forecasting capabilities. The FY-4 measurements aimed at accomplishing this are: high temporal and spatial resolution imaging in over 14 spectral bands; lightning imaging; and high spectral resolution IR sounding observations over China and adjacent regions. Current products from FY-2 will be improved by FY-4, and a number of new products will also be introduced.

This paper introduces the characteristics of FY-4A satellite data service system, which provides several ways for global community to get FY-4A data. Users in FY-4A direct broadcast service area with appropriate receiving equipment can directly receive real time L1 data transmission of FY-4A satellite. The CMACast users can receive L2 or L3 product with DVB-S equipment in near real time. The full FY-4A dataset, both real time and historical data will be available on NSMC satellite data service website in English version(http://satellite.nsmc.org.cn). Users can search and download FY-4A data after registration.

Several new IT technologies are applied in FY-4A satellite data service system. Distributed NAS systerm provide PB level volume real-time and archived satellite data to support internal users. Public Cloud service is used to deliver near real time data to external users. Big Data technology is adopted for system performance monitoring and user's behaviour analysis.

9. Archive Reload Function of the Online Data Management System for Earth Observation Data Exploitation Platforms

Markus Kunze, Stephan Kiemle, Nicolas Weiland and Matthias Hofmann

The amount of exploitable remote sensing data has increased significantly in recent years. As a result, the requirements of applications for fast processing of large coverages and time series in high resolution have also increased, which confronts Earth Observation (EO) data centers with new challenges. In order to cope with these challenges, approaches such as "Bringing users to the data" have emerged. For this purpose, exploitation platforms with large online memory and processing possibilities are being developed, which provide the functionality to process the data on the platform itself instead of transferring them to the user.

The German Aerospace Center DLR is developing an Online Data Management System (ODMS) which provides functionality for managing EO data in exploitation platforms. Due to the enormous amount of data, it is challenging to store all the data on the platform itself despite their large online storage capacity. From time to time, data will have to be evicted based on specific rules. However, in many exploitation use cases large data set series of historic data need to be held online to be processed into consistent time series products. Therefore the ODMS shall provide functionality to efficiently reload evicted data from the backend archives. This introduces additional complexity in the overall concept and results in particular challenges for the development of the system such as performance and optimization for bulk data reloading, handling priorities, taking into account data popularity for eviction and reload while respecting IT security constraints. Additionally, the ODMS strives for decoupling workloads for reload from traditional archiving tasks, and ensuring consistency of data holdings the online platform and the backend archives.

This paper will give an overview of the architecture focusing on the archive reload functionality, describe the current state of development and performance measurements. It will introduce sample implementations in projects such as CODE-DE and DIAS for which DLR provides the backend archives for Sentinel-1, Sentinel-3 OLCI, Sentinel-2 and Sentinel-5 Precursor data.

10. Introduction to the Fengyun satellite data sharing services on the Belt and Road

Di Xian and Xue Li

The Silk Road Economic Belt and the 21st-century Maritime Silk Road is a development initiative and framework proposed by China. As one of the most important member of WMO space-based observation system, Fengyun series satellites, developed and managed by the China Meteorological Administration (CMA), are providing near-real-time data services to countries along the Belt and Road via CMACast and direct ground receiving systems. As the important part of the integration of space and ground system, Fengyun satellite data sharing system via internet is also playing an important role in the earth observation data sharing system. According to the open data policy of CMA, data and products from Fengyun satellites and other related remote sensing satellites are shared via internet to the public since 2005. About 95 datasets derived from 25 satellites can be found and downloaded from the National Satellite Meteorological Center (NSMC) website. There are hundreds users came from about 40 countries along the Belt and Road who have downloaded hundreds Terabytes data from the website. To evaluate the economic influence of Fengyun satellite data sharing service, a web-based survey was held by the end of 2017. The target of this survey was all data users who have applied for and downloaded data from the NSMC website. We have put the notification on the website and send emails to all of our data users. Up to the end of this survey, about 400 respondents have completed the questionnaire online. The results have passed reliability test and validity test. This study analyzed the statistical results from respondents and annual reports of the Fengyun satellite data services, and described the following aspects: user description, data utility, data service opinion, and social analysis. The results confirmed the increasing influence of Fengyun satellite data in

scientific researches, meteorological services and other social activities. People are paying more attention to Fengyun satellites than other satellites, because they are easy to get and its dataset is becoming more and more useful.

11. The ESA CCI Open Data Portal

Fay Done and Kevin Halsall

The purpose of the European Space Agency's Climate Change Initiative (CCI) programme is to ensure that the full potential of long-term global Earth Observation (EO) archives for a number of Essential Climate Variables (ECVs), defined by the United Nations Framework Convention on Climate Change (UNFCCC), are realised. The programme currently covers 14 CCI ECV projects (entailing collaborations between over 85 institutions from all over Europe, the U.S.A and Canada), each generating validated CCI ECV datasets, from harmonised multi-sensor satellite data, which can be used to provide a solid basis for climate science and modelling, specialist application development and, ultimately, European and global policy making.

A single point-of-access (free and open) to these CCI ECV datasets, hosted by the Centre for Environmental Data Archival (CEDA)'s CCI Central Data Archive (CDA), is provided through the CCI Open Data Portal (ODP). The CCI ODP project is being led by Telespazio VEGA UK Ltd, with partners from STFC, CGI, University of Reading and Brockmann Consult. The CCI CDA currently contains more than 150 datasets (incl. older versions, only 111 of which are available to users through the portal) amounting to approximately 99 TB of data. Users can take advantage of a number of useful features provided by the CCI OPD, including the novel CCI Dashboard, a visualisation interface in which users can navigate their way through the datasets (incl. useful metadata and links to documentation) and download data products (depending on the data/data format, download protocols include FTP, WMS, WCS and OpenDAP), in addition to a standard CCI ECV faceted search. The ODP is also complimented by the CCI Toolbox which allows for the user community to read and work (incl. the application of operations) with data (products) from multiple CCI ECV datasets and then display the outputs. The CCI Toolbox is designed to work with an interactive Graphical User Interface (GUI) or as a Command Line Interface (CLI) which allows users to create scripts to for detailed data processing, etc.

The CCI ODP is set to expand, and evolve, even further in the wake of 9 additional CCI ECV projects to be introduced in the next phase of the CCI programme, CCI+ (due to start Q2 2018).

12. Processing surface state vector by temporal regularization of optical, thermal and SAR data

Maxim Chernetskiy, Mathias Disney, Marcel Urban, Alberto Delgado, Maurizio Nagni and Christiane Schmullius

Changes in the Earth’s surface can have very different properties and so can influence very different domains of the electromagnetic spectrum. If we can use these different domains effectively, they can provide input to machine learning methods to help in detection of changes of earth surface. This is particularly useful for trying to detect changes in ecosystem structure and function, a potentially vital application for satellite monitoring of the Earth system.

One of main complications in combining different Earth Observation (EO) data streams is a requirement of common time and space resolution. We present work from the BACI (Biosphere-Atmosphere Change Index) project, which requires gap free time series of Earth Observation (EO) data with their associated uncertainties, to develop new automated (machine learning) methods of change detection of the biosphere. Here, we present examples of the development and testing of this gap free time series, of EO data across optical (reflectance, albedo), passive microwave (LST) and active microwave (backscatter) domains. In order to normalise reflectance, fill gaps due to cloudiness and calculate uncertainties we use temporal regularisation which is based on Bayesian inference. In this way we are able to combine data across wavelengths, with different spatial and temporal properties, into a common observation framework, which we term the surface state vector (SSV). We use these data to generate optimally smoothed and filtered time

series of reflectance, albedo and backscatter, starting in 2000 and running to the present, as the core SSV output. Crucially, the SSV is provided with consistent uncertainties, which is key for use in downstream quantitative modelling and change detection applications, particularly to help attribute and explain detected change. Inputs to the BACI SSV are MODIS daily reflectance and LST data, Sentinel 1 backscatter and historical microwave (ENVISAT ASAR).

A key innovation of the BACI SSV processing chain is the use of the multitasking facilities of CEMS/JASMIN cluster to process almost 20 years of EO data across domains . A web-interface to the resulting SSV is implemented by means of GeoServer and CEMS. The resulting SSV will be made publically available by the BACI project, as it is likely to be of wider interest for various applications in ecosystem monitoring and change detection.

13. Quality control of CMIP5 data

Ruth Petrie, Martin Juckes, Ag Stephens and Richard Smith

The Climate Projections for the Copernicus Data Store (CP4CDS) is a Copernicus Climate Change Services (C3S) project. One aim of this project is to supply the CDS with a quality controlled subset of CMIP5 (Coupled Model Intercomparison Project - Phase 5) data. CMIP5 data is distributed using the Earth System Grid Federation. The first step in providing the quality controlled data is ensuring that the data supplied to the CDS is the most recent version of the data (for CMIP5 there was no automated way of tracking this), this involves comparing the metadata of published CMIP5 record on ESGF. The next step is to ensure that the data that is passed to the CDS meets a minimum quality control standard. Three quality control checking tools are used. 1) the CF-Checker, this ensures that the file metadata meets the CF-Conventions that were in place at the time. 2) The Centre for Environmental Data Analysis - Compliance Checker (CEDA-CC) this tool checks the file metadata is consistent with the CMIP5 project specified controlled vocabularies and also checks some variable metadata. 3) A time-checking tool, this tool ensures that the temporal metadata is consistent with the modelling calendar used. All three tools are python packages and are freely available on GitHub. Files that do not meet the minimum quality control standards are not included in the subset provided to the CDS. Where possible the file metadata will be corrected and in some cases the data will be modified. Data modifications will be either where a variable is provided in non-standard units or with a sign that is inconsistent with the specified sign convention. All the results of the quality control checks are collated in a database that is available to the user. In the database the user can find information on data availability and the quality control results for each datafile provided to the CDS.

14. STFC Data Analysis as a Service (DAaaS)

Frazer Barnsley

The STFC Data Analysis as a Service (DAaaS) is a system that simplifies the process of scientific data analysis from data produced by the facilities based at the Rutherford Appleton Laboratory (ISIS Neutron and Muon Source, Central Laser Facility and Diamond Light Source). This process is becoming more and more complex due to the advancements of the instruments and the scientific techniques. In some cases the volumes of data have grown so large that it is no longer practical for scientists to transport their data back to their home institution. In other cases, the analysis requires access to high performance computing or GPUs and complex chains of software where these resources, and access to the necessary expertise, may not be available. All of this technical complexity is being exposed to the scientists, who are quite often not computing experts, and is resulting in the analysis process becoming a bottleneck.

To solve this the DAaaS system makes use of virtual machines in an Openstack cloud that provide scientists with customised environments for specific types of scientific analysis. These environments provide the users with all the necessary software, data and compute resources they require to perform their analysis.

Users are also able to share access to these environments with colleagues and instrument scientists for collaboration and support.

15. Online Access to Historical Solar-Geophysical Data: Efforts by UK Solar System Data Centre

Matthew Wild, Yulia Bogdanova and Steve Crothers

The UK Solar System Data Centre (UKSSDC, https://www.ukssdc.ac.uk/) has been working to improve access to its extensive holdings of historical Ionospheric and Solar data recorded on a range of media (paper, film and glass plate photograph).

We have already made available electronic copies of the Royal Observatory Greenwich Photo-Heliographic Reports for the years 1874-1976. These reports contain measurements of the positions and areas of sunspots and faculae taken from photographs of the Sun made routinely at various observatories. In our solar archive, we hold over 10,000 images of the Sun on glass plates and over 20,000 images as prints from the years 1903-1942. Thanks to a Natural Environment Research Council grant we have been able to digitise the glass plate images. We also have scans of the Royal Greenwich Observatory (RGO) Solar images covering the period 1918-1976 and would like to bring these together with the RGO images stored at Cambridge University Library to provide a comprehensive Solar image archive accessible to all. In addition to the historical data archive, we hold data from the SOHO, STEREO, and TRACE missions.

In our archive, ionospheric data from 200 stations worldwide (1930s-present), such as ionograms and scaled ionospheric parameters (e.g., foF2, fmin, h'F2), is held on both digital and physical media. From the 1990s these data sets are available in digital form and can be downloaded via our web-interface. Thanks to a Natural Environment Research Council grant we are in the process of digitising a selection, 2,700 out of ~27,000, of UK ionosonde film data to be made available via the web interface. It is hoped that more funding will be made available to continue this exercise over the next few years. The UKSSDC also provides real-time ionospheric data retrieval from two RAL Space ionosondes, Chilton and Port Stanley, alongside other observatories worldwide.

The UKSSDC is a UK national data archive facility with open data access and can be used by scientists around the globe.

16. Digital Preservation in the Jisc Research Data Shared Service

Matthew Addis, Justin Simpson, Joel Simpson and Peter Van Garderen

The Jisc Research Data Shared Service (RDSS) is piloting a national Shared Service for Research Data Management (RDM) for Higher Education Institutions (HEIs) in the UK. The RDSS supports the deposit, storage, publication and preservation of a wide range of digital research outputs including datasets. Arkivum and Artefactual are delivering an open-source digital preservation solution into the RDSS. Many HE institutions are only just starting to address the issue of digital preservation of research data sets, both to meet funding body requirements on an institutional level and to help their researchers meet the F.A.I.R. principles over the long-term. Challenges include (a) how to integrate digital preservation into the research data lifecycle, including the potential for digital preservation to enable re-use of research data over the long term, (b) how to operate digital preservation at scale with minimal user intervention in order to keep costs manageable and fit with limited staff resources at HEIs, e.g. within the library or at departmental level, and (c) how to best apply digital preservation to the 'long-tail' of data formats found in research environments that are not well supported in existing preservation environments or format registries such as PRONOM. This presentation will present the RDSS preservation solution we have developed and our on-going work in this area. We propose to present three core areas of our work: (i) preservation workflows and how digital preservation can be seamlessly integrated with Institutional Repositories and Reporting Dashboards; (ii) achieving scalable and automated digital preservation through a micro-services architecture, containerised deployment into a cloud environment, and message-based integration with other RDSS components; (iii) the design of a 'Preservation Actions Registry' that allows a community to define and provide a resource

of machine-readable descriptions of what tools can be used for digital preservation of research data formats, including what properties can be extracted or measured from the data, what preservation actions can be taken, and how to execute these actions in practice. We propose to present our work as an oral session. We would also be very willing to present as a poster if the organising committee felt that more appropriate.

17. Rescuing Data to Understand how we Determine our Future

Elizabeth Griffin

Everything we do today, and did yesterday, has a bearing on aspects of the physical world that we shape for tomorrow. What we emit into the atmosphere, what we draw from water reserves, what we remove from the sea, how we redirect natural water-courses, how we shield insolation, or what we toss down the drain – each has a cumulative effect on the passive Planet that we are thus treating. The same is true in reverse: we can determine what the Planet used to be like by analyzing or re-analyzing the many procedures that have combined to shape it into what we find today. Clearly, the more actual evidence that can be brought to bear in such tasks, the more realistic the picture that can be recreated, and the more reliable the models that will then try to simulate the processes involved. But it is also essential to accept that none of those processes has acted in a logically-determinable manner. Chaos enters with brute force into everything, be it a local hill that changed the air-temperature, a local wind that affected the production of tropospheric ozone, a local celebration that influenced food culling and water intake, a mishap that caused flooding or fire, etc. None of those can be predicted by a computer, which is a purely logical brain, yet each has contributed cumulatively to altering the state of Planet Earth. It is therefore vital that we recover as many actual observational data from the past, from whatever reserves, as can be accomplished. All scientific data are part of this critical quest; some are “small” in the sense of being personal or well-understood; some are “big” both in volume and in the modern usage of the term; even ones whose influences are thought of as secondary in nature (such as observations of the cosmos) are necessary to humanity's database for studying changes in the natural world. This talk will expand the above points, and give examples of the effectiveness of applying successfully rescued data to modern research. It will also argue that data rescue, as defined above, is a crucial element of all scientific research.

18. A Space Weather VOEvent service provided by the CDPP in the frame of Europlanet H2020 PSWS

Michel Gangloff, Nicolas André, Vincent Génot, Baptiste Cecconi and Pierre Le Sidaner

The CDPP (Centre de Données de la Physique des Plasmas,(http://cdpp.eu/), the French data center for plasma physics, is engaged for two decades in the archiving and dissemination of plasma data products from space missions and ground observatories.

Under Horizon 2020, the Europlanet Research Infrastructure includes PSWS (Planetary Space Weather Services), a set of new services that extend the concepts of space weather and space situation awareness to other planets of our solar system. One of these services is an Alert service associated with solar wind prediction made using the CDPP Heliopropa service (http://heliopropa.irap.omp.eu), and detection of meteor shower, lunar flash and cometary tail crossing. This Alert service, is based on VOEvent, an international standard proposed by the IVOA and widely used by the astronomy community. The VOEvent standard provides a means of describing transient celestial events in a machine-readable format. VOEvent is associated with VTP, the VOEvent Transfer Protocol that defines the system by which VOEvents may be disseminated to the community. VTP is managed with Comet, a freely available and open source software. Comet is used by PSWS for its Alert service and several partners of PSWS, including the CDPP and Observatoire de Paris.

This presentation will focus on a prototype of the alert system implemented with the current version of the VOEvent standard and the enhancements of the standard necessary to take into account the needs of the Solar System community.

Europlanet 2020 RI has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 654208.

19. Migrating the UMARF Catalogue Database

David Berry

EUMETSAT is the European Organisation for the Exploitation of Meteorological satellites, which provides 24/7 access to weather and climate satellite data. The Unified Meteorological Archive and Retrieval Facility (UMARF) at EUMETSAT is responsible for the long term preservation of satellite data from all current and future missions. The UMARF follows the guidelines of the OAIS (Open Archival Information System) reference model. The UMARF catalogue database constitutes part of the Data Management component, and is responsible for maintaining the catalogue of product metadata, user details and order information. The database is currently being migrated from Oracle to PostgreSQL, a necessary evolution to prepare to the vast number of products that are anticipated in the coming years from new missions, in particular Meteosat Third Generation (MTG) and EUMETSAT Polar System – Second Generation (EPS-SG).

In this poster we present the challenges and lessons learnt during the migration project so far. Firstly we describe the migration strategy, which has been carefully selected to avoid any disruption to operational systems and achieve a smooth transition. We then explore several issues that have been encountered, including; porting issues at database and application code levels, updates to the technologies that are used, how we have improved the maintainability, quality and reliability of the software. Finally we address the lessons learnt from the project.

20. Interactive Visualization and Analysis for Large Time-varying Multivariate Earth Science Data

Jin Wang, Yu Pan, Michael Rilee, Lina Yu, Feiyu Zhu, Kwo-Sen Kuo and Hongfeng Yu

We demonstrate a new interactive visualization and analysis system for large time-varying multivariate Earth Science data. The system consists of multiple views and each view provides specific visual analytics capabilities. The views work in an interactive and linked manner, where user operations conducted in one view can trigger an instant update of visualization and analysis results in other views. This design can facilitate users to simultaneously explore different aspects of data and comprehend phenomena and relationship among multiple variables.

The current datasets stored in the system include (1) an hourly dataset of the NASA Modern Era Retrospective-analysis for Research and Applications (MERRA-2), (2) a reprocessed 5-minute National Mosaic and Multi-sensor QPE (NMQ, where QPE stands for quantitative precipitation estimate), and (3) a swath dataset, from NASA's Tropical Rainfall Measuring Mission (TRMM), deriving vertical hydrometeor profiles using data from Precipitation Radar (PR) and TRMM Microwave Imager (TMI). Our system supports users to conduct interactive query, visualization, and statistical analysis operations on these datasets. First, the main view displays all these datasets, where users can interactively choose any time slide and see the rendering results of the selected datasets. Second, users can input their customized queries in space-time to easily explore features of interest among multiple heterogeneous datasets. Third, several real-time statistical analytics (such as histogram and correlation) can be conducted on data of user queried results or selected regions. Moreover, queried data and analysis results can be easily downloaded as files in conventional formats (e.g., csv files) for post-processing at a user side.

The backend of the system is based on our new data indexing scheme, SpatioTemporal Adaptive-Resolution Encoding (STARE), and a scalable data partitioning and distribution method. These advanced techniques allow us to spatiotemporally co-align arrays of different shapes and models of time-varying multivariate Earth Science data, and devise an optimal placement of multiple datasets in a distributed environment. Our end-to-end system is able to achieve interactive visualization and analysis with scalable performance.

EUMETSAT

esa

DLR Deutsches Zentrum für Luft- und Raumfahrt e.V. in der Helmholtz-Gemeinschaft

ltdp

cnes CENTRE NATIONAL D'ETUDES SPATIALES

DCC

CATAPULT Satellite Applications

PV 2018 CONFERENCE

RAL Space

National Centre for Earth Observation NATURAL ENVIRONMENT RESEARCH COUNCIL

Science & Technology Facilities Council

UK SPACE AGENCY