Towards a Long-term Preservation Infrastructure for Earth Science Data

Brian Matthews

Scientific Information Group
Scientific Computing Department
STFC Rutherford Appleton Laboratory

Arif Shaon, David Giaretta[1], Esther Conway, Shirley Crompton: *STFC* (*[1]*APA)
Jinsongdi Yu: *Jacobs University;* **Fulvio Marelli,.***ESA*;
Ugo Di Giammatteo, Advanced Computer Systems
Yannis Marketakis, Yannis Tzitzikas: *FORTH,* **Raffaele Guarino**, *Capgemini*
Holger Brocks, *InConTec GmbH*, **Felix Engel**: FTK

scidip-es

Science & Technology Facilities Council

Contents

- Introduction
- The Problem
- Example: Meris
- Preservation Infrastructure
- Conclusion

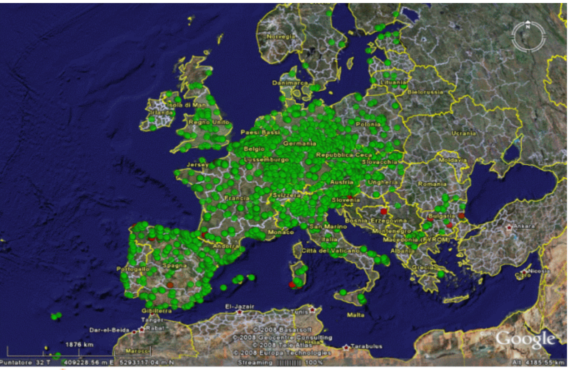
Contents

- Introduction
- The Problem
- Example: Meris
- Preservation Infrastructure
- Conclusion

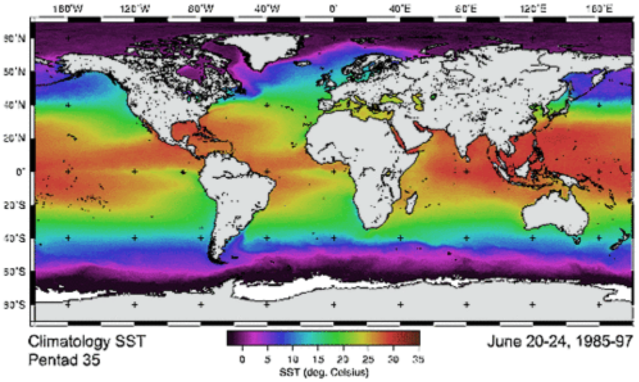
Earth Science Data

Data on:
- Oceans
- Atmosphere
- Land use
- Biosphere
- Geology
- Seismology
- Cryosphere

Used for:
- Disaster management
- Health
- Energy
- Climate Change
- Water
- Ecosystems
- Agriculture



Collected by:
- Samples
- Monitoring sites
- Traverses
- Satellite Observation



Climatology SST
Pentad 35

June 20-24, 1985-97

scidip-es

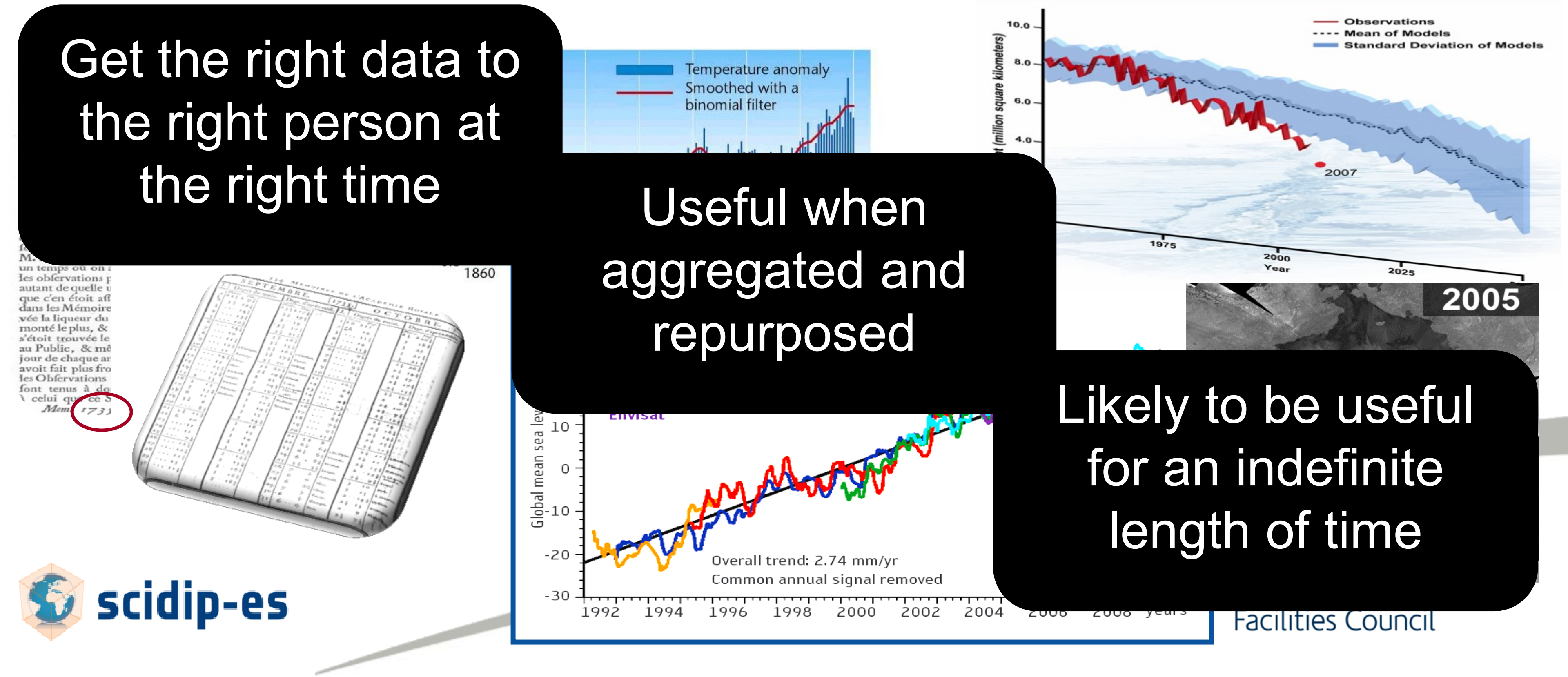
Science & Technology
Facilities Council

The Need

"A fundamental characteristic of our age is the raising tide of data – global, diverse, valuable and complex . In the realm of science, this is both an opportunity and a challenge."

Report of the High-Level Group on Scientific Data, October 2010

"Riding the Wave: how Europe can gain from the raising tide of scientific data"



Get the right data to the right person at the right time

Useful when aggregated and repurposed

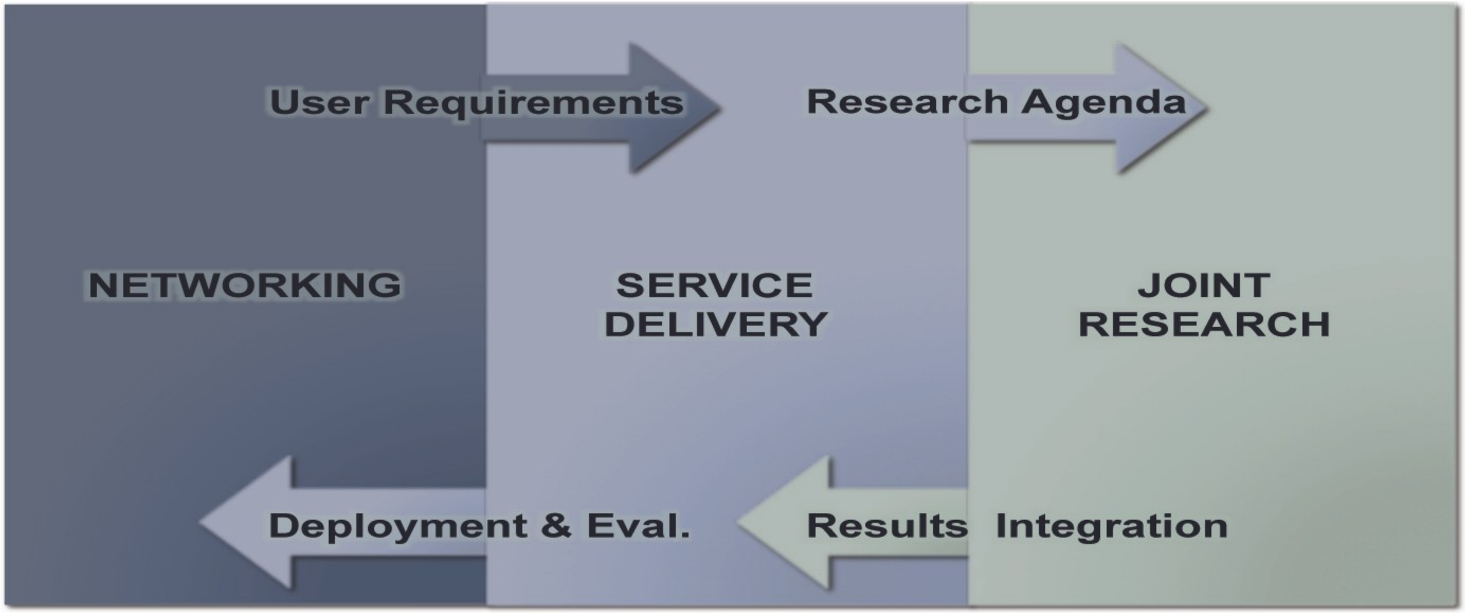
Likely to be useful for an indefinite length of time

SCIDIP-ES Overview

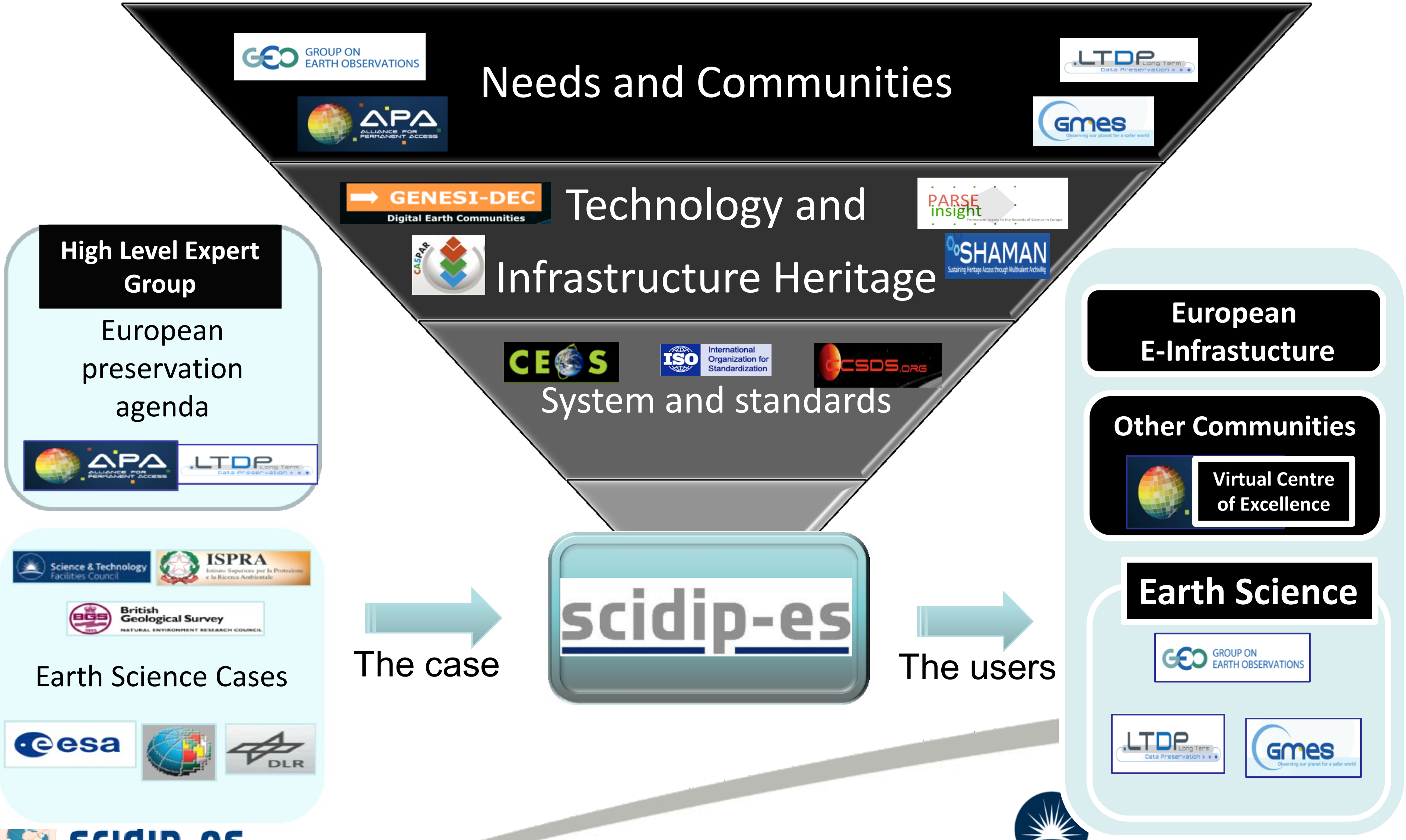
- An EC FP7 funded project; under the Grant Agreement 283401.
- 13 Work packages to be completed by 17 beneficiaries
  - Project Coordinator : European Space Agency (ESRIN, Italy)
- Project time frame: **October 2011 – September 2014**
- Antecedent projects: CASPAR, SHAMAN

SCIDIP-ES Objectives

- **Deliver generic sustained services** for long-term preservation and usability as part of the data infrastructure for e-Science.
- **Harmonize data preservation** policies, approaches and tools in the Earth Science Domain.



The Context


GROUP ON EARTH OBSERVATIONS
APA ALLIANCE FOR PERMANENT ACCESS
LTDP Long Term Data Preservation
GMES
GENESI-DEC Digital Earth Communities
PARSE insight
CASPAR
SHAMAN
CEOS
ISO International Organization for Standardization
CCSDS.ORG
Science & Technology Facilities Council
ISPRA
British Geological Survey
esa
DLR
scidip-es

Needs and Communities

Technology and Infrastructure Heritage

System and standards

High Level Expert Group

European preservation agenda

Earth Science Cases

The case

The users

European E-Infrastructure

Other Communities

Virtual Centre of Excellence

Earth Science

Contents

- Introduction
- The Problem
- Example: Meris
- Preservation Infrastructure
- Conclusion

Ensuring Intelligibility and Re-Usability of Data

- Need format information
  - And tools to render
  - But not enough
- What do the numbers mean?
  - Field meanings
  - Units
  - Accuracy
  - Context
  - Interpretive software
- Add Representation Information
  - Forms a dependency graph
  - Maybe alternatives
  - Can calculate costs and risks
- Preservation Network Model (PNM)

```
Slough (SL051) : 1975-01 to 1977-01
SLOUGH SL051 0 51.5359.4 Manual Edited
1975 1 31 2 744 24 24 24 24 24 24 24 24 24 24 24 24
24 24 24 24 24
24 24 24 24 24 24 24 24 24 24 24 24 24 24 foF2
M3000F2 0.1 MHz 0.01
0003 00000 10000 20000 30000 40000 50000 60000
70000 80000
900001000001100001200001300001400001500001600
00170000180000190000
200000210000220000230000 00000 10000 20000
30000 40000 50000 60000 70000 80000
000010000011000012000013000014000015000016000
0170000180000190000200000210000220000230000
00000 10000 20000
30000 40000 50000 60000 70000 80000
90000100000110000
120000130000140000150000160000170000180000190
000200000210000220000230
000 00000 10000 20000 30000 40000 50000 60000
```

Ionosonde Data Retreival results (CEDA)
IIWG format
URSI Codes

Further Barriers and Challenges of ES Data Preservation

- Designing a cost effective preservation solution
  - Maintaining reusability in preservation complex
  - Need methods and tools to make this manageable
- Reacting to changes in preservation requirements
  - Things change
  - Monitor change and propose suitable actions
  - Sharing knowledge
- Maintaining Authenticity
  - Gathering evidence of the extent that authenticity can be maintained.
  - Provenance, fixity, context, access rights, reference

- Supporting Practical Business Models for Data Preservation
  - Minimising costs
  - A community approach – pooling expertise, services, rep info

Challenge Use Cases

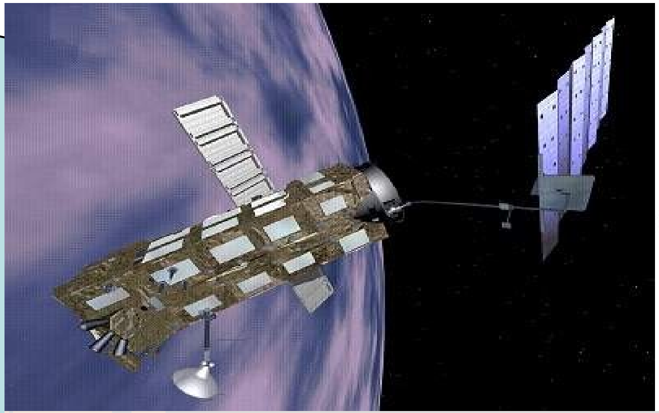
- **Preservation Archive Creation:**
  - what information should be preserved for future use, by an identified Designated Community (DC)?
- **Archived Data Access**:
  - what kind of enhanced information could be provided to current and future consumers to add value to the preserved data, ?
- **Archive Change/Evolution**:
  - how to preserve data against changes in related technology and in the designated community ?

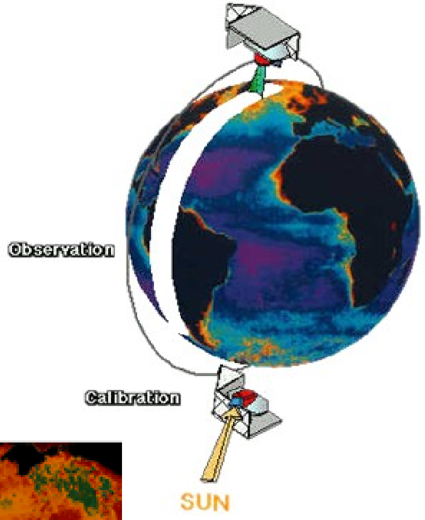
Contents

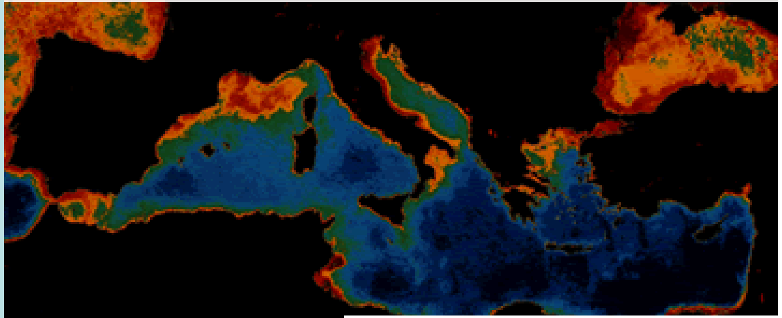
- Introduction
- The Problem
- Example: Meris
- Preservation Infrastructure
- Conclusion

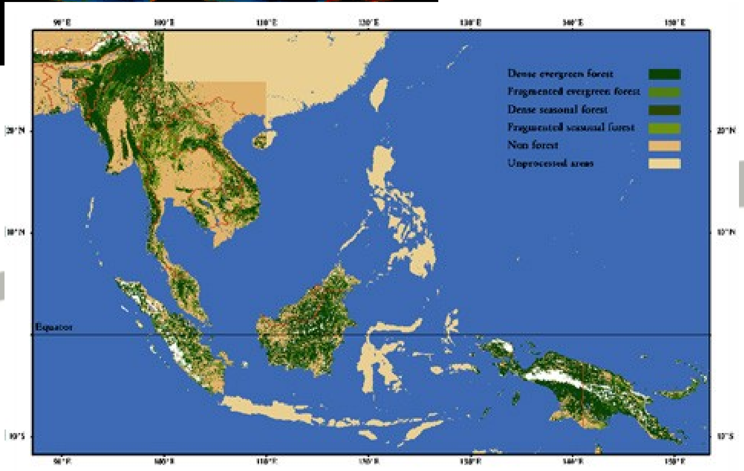
Example: data from the MERIS Instrument

- Medium Resolution Imaging Spectrometer (MERIS)
  - an instrument on the ESA ENVISAT EO satellite
- Primarily: sea colour measurement
  - Chlorophyll
  - Suspended sediment
  - Atmospheric aerosol over water
- Also land vegetation
- Understand the carbon cycle
  - How this changes under climate change
  - Also agriculture and fisheries







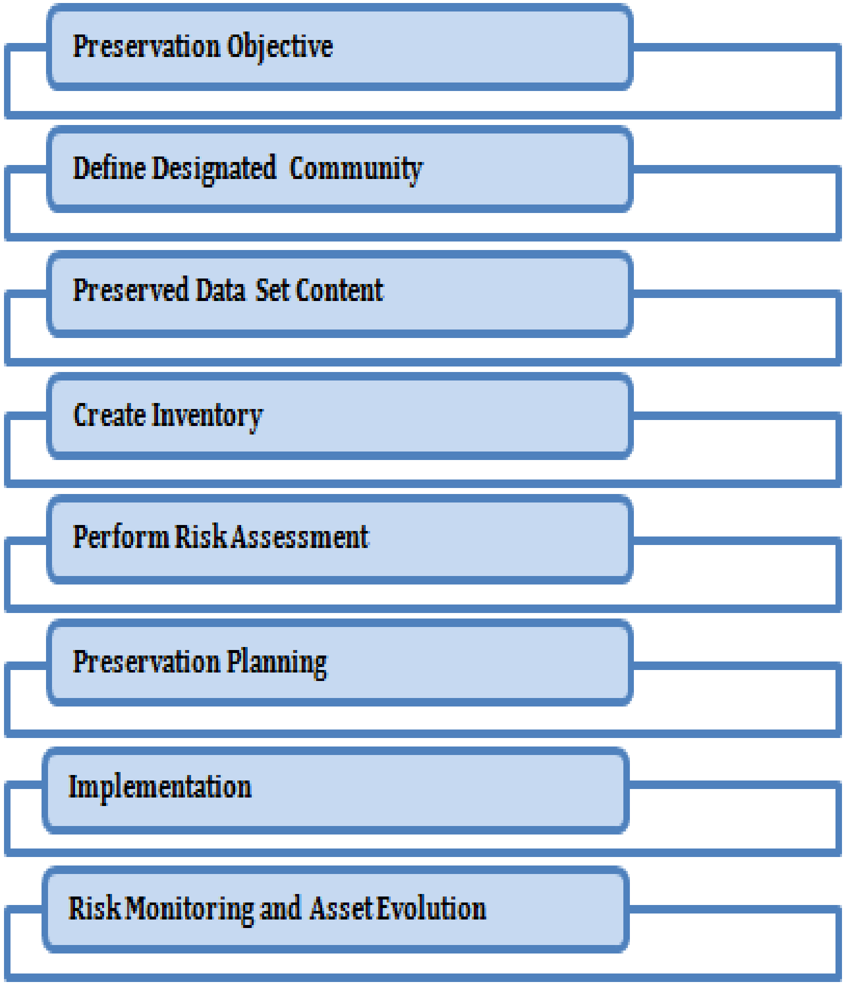


Preservation Archive Creation for MERIS

- Undertake a preservation analysis of MERIS data

- Preservation Objective
  - Preserve ESA MERIS data to maintain its time series usable for 50 years.
  - storage/archiving of the **ESA MERIS N1 File Level 0** (L0) and
  - storage/archiving of the **ESA MERIS N1 File Level 1 (L1).**

- Designated Community
  - ESA staff – with full specific knowledge of ENVISAT datasets.
  - Principal Investigator (PI) - They know the MERIS data's scientific value but don't have the skills
  - University Students - they are learning MERIS data and need to fully understand it.

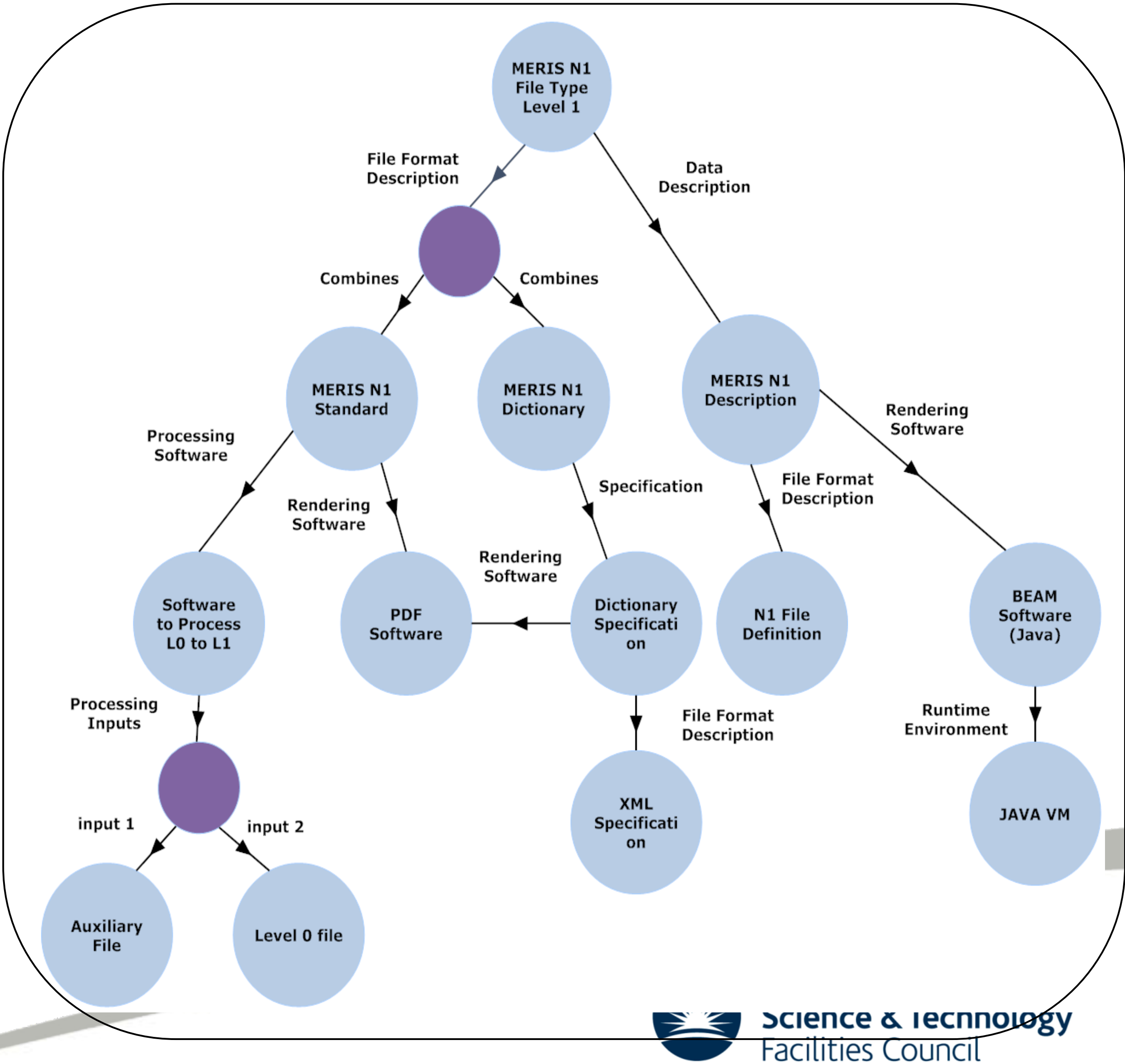


scidip-es

Science & Technology Facilities Council

MERIS Preservation planning and risk analysis

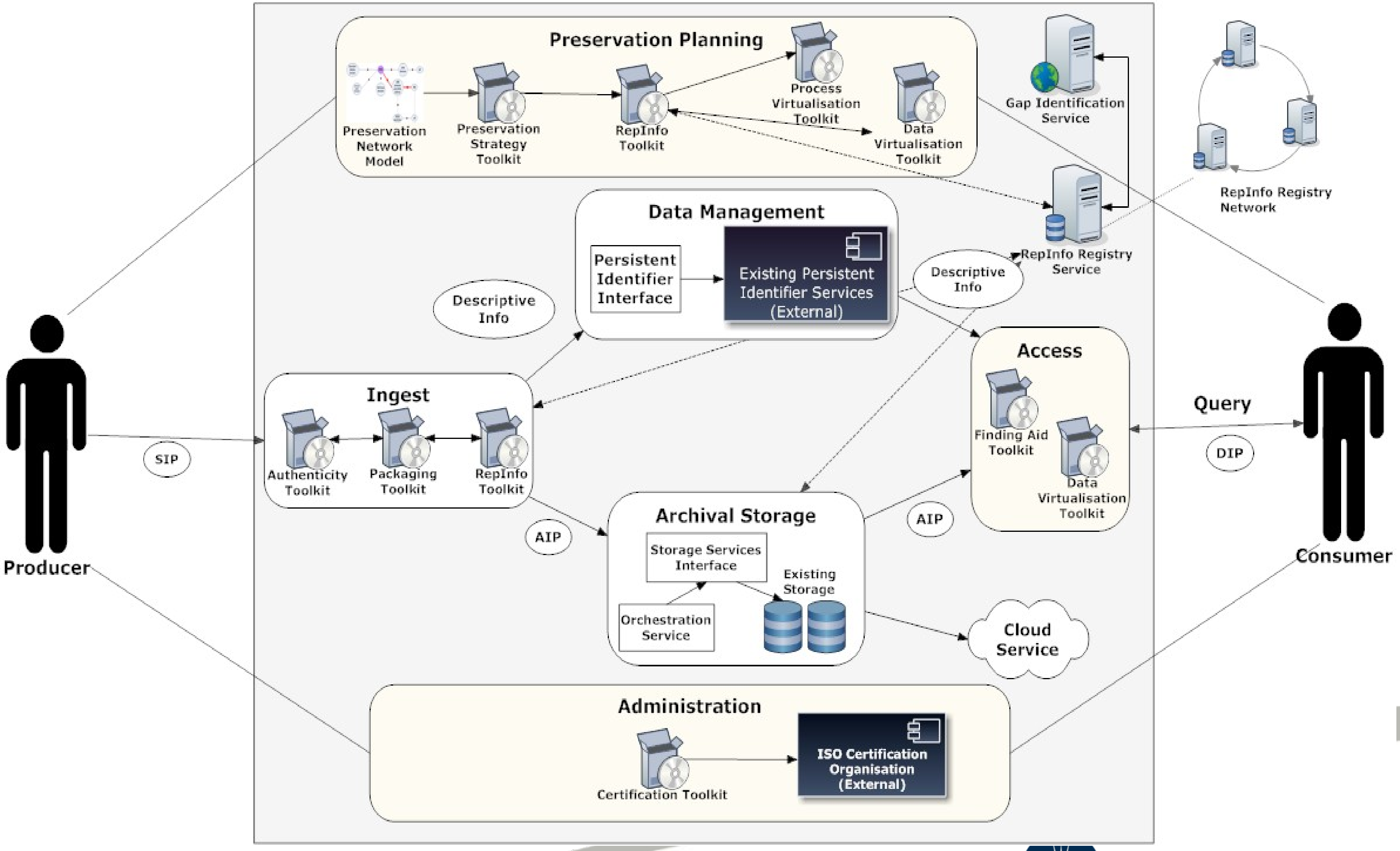
- Create PNM to capture dependencies between Rep Info
- Each item has a risk associated with it
  - Can make a risk assessment of possible strategies and make a choice
- Use the PNM
  - To guide the design of the AIPs
  - To monitor the changes in the environment and DC
  - To plan subsequent preservation actions.



Contents

- Introduction
- The Problem
- Example: Meris
- Preservation Infrastructure
- Conclusion

SCIDIP-ES Preservation Infrastructure



scidip-es

Science & Technology Facilities Council

Tools to Support Planning

- Preservation Strategy Toolkit
  - Allow PNMs to be designed and evaluated to generate a plan
  - Helps design suitable AIPs
- RepInfo Toolkit
  - Design and capture Representation Information about data
  - Stores RepInfo in the Registry
- Process Virtualisation Toolkit
  - Plan and enforce re-processing action on data objects
- Certification Toolkit
  - Self audit for compliance with OAIS (ISO 16363)

Tools to Support Ingest, storage and management

- Packaging Toolkit
  - Build suitable AIPs
  - Links to items in the RepInfo Registry and RepInfo Toolkit
- Authenticity Toolkit
  - Captures and checks Authenticity information associated with a AIP.
  - Provenance, fixity, context, reference, access rights
- Storage Services
  - Stores and migrates AIPs – AIP aware
  - Storage platform independent
- Persistent Identifier Service
  - Maintain Persistent Identifiers to objects

Tools to Support Monitoring

- Gap Identification Service
  - Assesses changes and risks to data objects
  - Evaluates changes in the designated community
  - Identify “gaps” in the intelligibility of the data objects
- Orchestration Service
  - Brokerage between current and future data holders
  - Exchange intelligence about events
    - could be used to monitor changes and risks to data objects
  - Trigger corrective actions.

Tools to Support Access

- Finding Aid Toolkit
  - Supplements domain specific search facilities
  - Uses common metadata and semantic definitions
  - Provides a user interface onto the AIPs.
- Data Virtualisation
  - A “quick look” at data
  - Uses RepInfo description to inspect and describe contents of data objects
  - Format independent manner

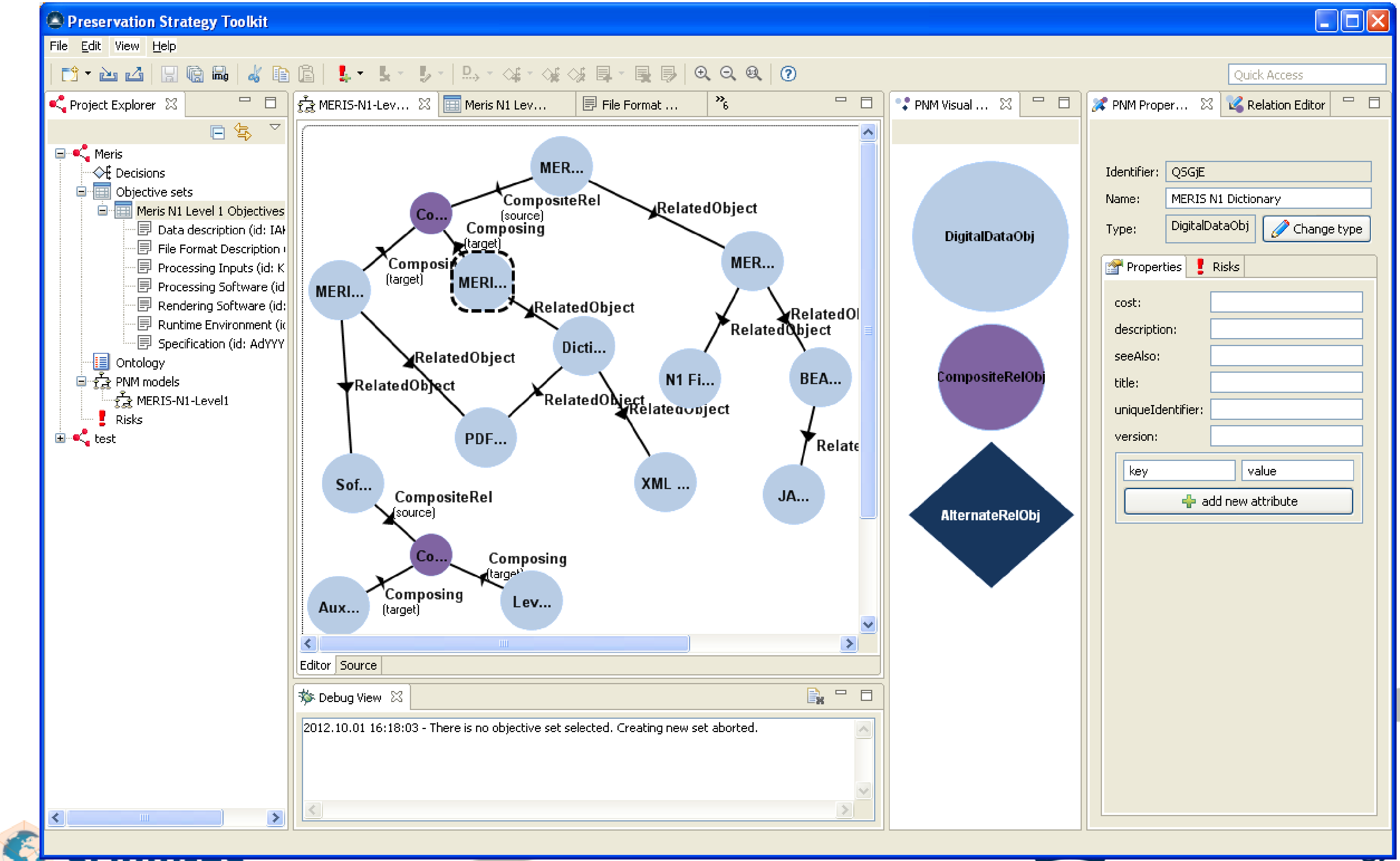
Contents

- Introduction
- The Problem
- Example: Meris
- Preservation Infrastructure
- Conclusion

Current Status

- One year into the project
- Defined use cases with User Partners
  - E.g. MERIS
- Defined a number of outline scenarios
- Initial prototype implementation of services
  - Based on the CASPAR toolkit
  - Identifying shortcomings and improvements
- Specifying and implementing new prototype
  - Need to then test in scenarios with the user cases

Preservation Strategy Toolkit



- Meris
  - Decisions
  - Objective sets
    - Meris N1 Level 1 Objectives
      - Data description (id: IAI
      - File Format Description
      - Processing Inputs (id: K
      - Processing Software (id
      - Rendering Software (id:
      - Runtime Environment (i
      - Specification (id: AdYYY
  - Ontology
  - PNM models
    - MERIS-N1-Level1
  - Risks
- test

Identifier: QSGjE

Name: MERIS N1 Dictionary

Type: DigitalDataObj

cost:

description:

seeAlso:

title:

uniqueIdentifier:

version:

key | value

```
2012.10.01 16:18:03 - There is no objective set selected. Creating new set aborted.
```

Final Word

- Preservation analysis to support a preservation Strategy
  - Earth Science as the “test domain”
  - Applicable to other domains too.
- Preservation tools:
  - Make them simple and robust
  - Deliver as sustainable services
- Make the case to the users
  - Establish the value of preserving
  - Work in a language they understand

Thank You

Questions?

brian.matthews@stfc.ac.uk

www.scidip-es.eu

scidip-es

Science & Technology Facilities Council