Long-term sustainability of spatial data infrastructures: a metadata framework and principles of geo-archiving

Arif Shaon
Science and Technology Facilities Council, UK
arif.shaon@stfc.ac.uk

Carsten Rönsdorf
Ordnance Survey, UK
Carsten.Roensdorf@ordnancesurvey.co.uk

Urs Gerber
The SWISS Federal Archive, Switzerland
Urs.Gerber@lt.admin.ch

Kai Naumann
Landesarchiv Baden-Württemberg - Staatsarchiv Ludwigsburg, Germany
kai.naumann@la-bw.de

Paul Mason
Ordnance Survey, UK
Paul.Mason@ordnancesurvey.co.uk

Andrew Woolf
The Bureau of Meteorology, Australia
A.Woolf@bom.gov.au

Michael Kirstein
Generaldirektion der Staatlichen Archive Bayerns, Germany
Michael.Kirstein@gda.bayern.de

Marguérite Bos
The SWISS Federal Archive, Switzerland
Marguerite.Bos@bar.admin.ch

Göran Samuelsson
Mid Sweden University, Sweden
goran.samuelsson@miun.se

ABSTRACT

With growing concerns about environmental problems, and an exponential increase in computing capabilities over the last decade, the geospatial community has been producing increasingly voluminous and diverse geographical datasets. Long-term preservation of these geographical data exposed through uniform and interoperable Spatial Data Infrastructures (SDIs) is not typically addressed, but highly important for meeting legislative requirements, the short and long term exploitation of archived data as well as efficiency savings in managing superseded datasets. In this paper, we attempt to set out the path and describe what needs to be done now to future-proof the investment government agencies around the world have made in digital geographic data. We take the INSPIRE SDI as an exemplar to investigate the requirements for ensuring sustained access to geographical data from the perspective of a preservation-aware and INSPIRE-conformant SDI. We also outline a number of principles for the long term retention and preservation of European digital geographic information defined by the EuroSDR Geographic Data Archiving working group. In addition, we present a preservation profile of the ISO 19115 metadata standard to enable recording and exposing important preservation related information about geographical data through large-scale SDIs like INSPIRE.

Keywords

preservation, archive, metadata, INSPIRE, ISO 19115, geographical data.

1. INTRODUCTION

Geo-information systems (GIS) have become an indispensable means of storing and analysing geographical data for government, business, and research. In Europe, the National Mapping Agencies (NMAs) and other geographic institutions today experience rising demand for historical geographical data that describe how land, cities and countries have developed over time. Government agencies around the world have invested heavily in this type of geographical data. Unfortunately, high storage costs and difficulties in finding, accessing and delivering older datasets and raster data[1] are making the task of satisfying this demand extremely challenging. Unlike paper maps, digital geographical data without efficient curation and preservation could become unusable within about one decade due to software, hardware or data model obsolescence. Safeguarding today's fundamental geographical data for future generations in order to understand history as well as historic trends needs to be a core objective of the National Mapping Agencies and other data providers.

The European Union INSPIRE Directive[2] aims to address the need for interoperability across the geographical datasets held by its different member states. To facilitate such a high level of interoperability, the directive mandates the adoption of common Implementing Rules (IR) for metadata, data specifications, network services, and data sharing through a pan-European Spatial Data Infrastructure (SDI). While this is an effective way of ensuring interoperability across disparate datasets, it does not guarantee sustainability of those datasets over an indefinite period of time. For instance, INSPIRE does not address ensuring compatibility with future technology or ensuring continued access even after a provider has ceased to exist. To further illustrate, we can consider the specific requirement of INSPIRE for data providers to use the OGC[3] standardised Web Map Services[4] to expose GIS Maps. Currently, there is no standardised way of defining precisely which data tables, attributes, geometries or raster images are contained within such a service. But each of those components has different properties that will need to be migrated into newer systems or formats at some point in time to ensure continued accessibility and usability.

Geographic information is already at the heart of environmental analysis that informs policy as well as practical implementation. For example, adding preserved digital snapshots of detailed land ownership and use, river and transport networks together with historical environmental measurements such as pollution or

[1] Raster graphical data - http://en.wikipedia.org/wiki/Raster_graphics

[2] INSPIRE Directive - http://inspire.jrc.ec.europa.eu/

[3] Open Geospatial Consortium - http://www.opengeospatial.org

[4] http://www.opengeospatial.org/standards/wms

water quality over 10 or 20 years, coupled with new analysis techniques not available today will identify correlations and trends that allow better scenario models for the future and also inform environmental policy. As this shows, properly historicised geographic information provides tremendous value for government, economy as well as for individuals. We need historic data to meet economic and legal requirements for government and business, but also for citizens as a means of gaining deeper understanding of their lineage, for example, by tracing back their individual or family history. A large-scale SDI like INSPIRE has a crucial role to play in facilitating the availability of this type of geographical data over the long-term.

In this paper, we investigate the requirements for developing a preservation-aware SDI based on the OAIS reference model [5], an important ISO standard for digital preservation. We also outline a number of principles for the long term retention and preservation of digital geographical information with a view to introduce fundamental concepts of digital geographical data archiving for the public sector information providers in Europe. These principles have been proposed by the EuroSDR Geographic Data Archiving working group[5] – a group of 11 National Mapping Agencies, Archives and Research institutions across Europe collaborating to address the issues of preserving geographical data in Europe. In addition, we present a preservation profile of the ISO 19115 metadata standard[6] that is designed to enable an archive to record preservation-related information about geographical data and make it available to the users through the associated SDI.

2. THE MAIN CHALLENGES OF PRESERVING GEOGRAPHICAL INFORMATION

In general, geographical data inherit the preservation challenges inherent to all digital information [3]. These challenges are further complicated by some of the characteristics of geographical datasets, such as diverse and highly structured data formats, and the need for special domain knowledge for accurate interpretation. Moreover, in the context of SDIs, such as INSPIRE, state-of-the-art service-oriented infrastructures adopt exchange formats (i.e. application schemas) that reflect domain-specific conceptual data models ('feature types') rather than directly reflecting underlying database storage schemas. These application schemas and their relationships (e.g. mapping) with the corresponding datasets would need to be preserved to ensure appropriate accessibility and re-use of those datasets in the future.

On the positive side, it should be possible in principle, to apply existing widely adopted preservation mechanisms and standards, such as the OAIS reference model (Section 4) to the long-term preservation of geospatial data. In fact, a number of European archives [10] are currently adopting or are looking to adopt the OAIS model and other related specifications for the long-term preservation of their geospatial datasets. These organisations would, therefore, significantly benefit from a best-practice implementation profile of the OAIS model for geospatial datasets and an INSPIRE-compliant metadata model for describing and sharing the relevant preservation aspects (Section 5) of such datasets through the INSPIRE SDI – neither of which exist at present.

3. EXISTING ENDEAVOURS

Aside from a handful of initiatives, such as the NGDA[7] project funded by the NDIIPP[8] initiative of the US Library of Congress, the GER[9] project and some exploratory work by the Digital Preservation Coalition (DPC) [3], there have not been many noteworthy endeavours for long-term preservation of geospatial information. Amongst the existing initiatives, the GER project has introduced a new metadata model for describing geospatial information, which is essentially an amalgamation of FGDC[10] (the current US Federal Metadata standard), the ISO 19115 metadata model and a few preservation metadata specifications including the PREMIS Data Dictionary [9]. In general, the GER model is a comprehensive metadata model designed to enable capturing and managing a wide variety of preservation-related information (e.g. accessibility, provenance, distribution etc.) about a geospatial dataset during its entire life-cycle. The metadata-related notions defined in the GER are represented as relational database tables and their corresponding fields, with a view to facilitate the development of new archives for preserving geospatial data as well as improving the capabilities of existing archives [6]. As a result, the GER metadata model is not a true 'profile' of any of the existing metadata standards on which it is based; e.g. it does not follow the rules of profiling specified in Annex C of ISO 19115. From that perspective, it would not be fit for capturing and sharing metadata about geospatial datasets through large-scale SDIs, such as INSPIRE which requires the adoption of ISO 19115-conformant metadata models for describing geospatial data.

The NGDA approach, on the other hand, is specifically intended to address the preservation requirements of the US-based geospatial datasets at archive or repository levels. In particular, this approach includes a comparative assessment of a number of existing metadata standards, including the aforementioned GER and FGDC metadata model with a view to address the metadata capturing and management requirements of a long-term archive of geographical information [1]. However, such archive-specific technical solutions may not directly benefit large-scale SDIs more generally (including INSPIRE), where the main focus is on the provision of uniform accessibility of geospatial datasets, not specific techniques for preserving such datasets. Further, an SDI typically consist of many different data providers with different organisational remits and constraints – so, it would be impractical for an SDI to impose the adoption of a 'one-size-fits-all' preservation approach on all the data providers involved. Nevertheless, the NGDA approach could serve as

---

[5] EuroSDR Geographic Data Archiving working group - http://bono.hostireland.com/~eurosdr/start/index.php?option=com_content&task=view&id=60&Itemid=88

[6] ISO 19115:2003 Geographic information – Metadata

[7] National Geospatial Digital Archive (NGDA) Project - http://www.digitalpreservation.gov/partners/ngda/ngda.html

[8] National Digital Information Infrastructure and Preservation Program (NDIIPP) - http://www.digitalpreservation.gov/library/

[9] Geospatial Electronic Records (GER) project - http://www.ciesin.columbia.edu/ger/

[10] Federal Geographic Data Committee (FGDC) Metadata Format - http://www.fgdc.gov/metadata

guidelines for implementing geospatial preservation archives in Europe, mainly for the exploratory work done on various general aspects (e.g. data format, metadata mapping etc.) of long-term preservation of geospatial data.

Aside from the aforementioned endeavours, the European Space Agency (ESA) has recently established a major preservation initiative, the ESA Long-Term Digital Preservation (LTDP)[11] programme, with a view to formulate a coordinated and coherent approach to the long-term preservation of the EO space data archives across its member states. Although this ESA LTDP initiative primarily focuses on the preservation of Earth Observation (EO) space data, the end result of this initiative should also be applicable to other types of geographical data, and to INSPIRE. The work presented in this paper should be of considerable relevance to this ESA initiative, since ESA adopts ISO 19115 for collection-level discovery[12].

4. THE OAIS REFERENCE MODEL

The Reference Model for an Open Archival Information System (OAIS) is a very important ISO standard (ISO 14721:2003) for addressing the issues associated with the long-term preservation of digitally encoded information [5]. The OAIS describes a number of conceptual models in order to aid formulation of a suitable preservation strategy for digital objects. Of particular importance, among the OAIS models, is the Information Model that broadly describes the metadata requirements associated with retaining a digital object over the long-term (Figure 1). We consider the different components of the OAIS information model from the perspective of long-term preservation of geospatial datasets.

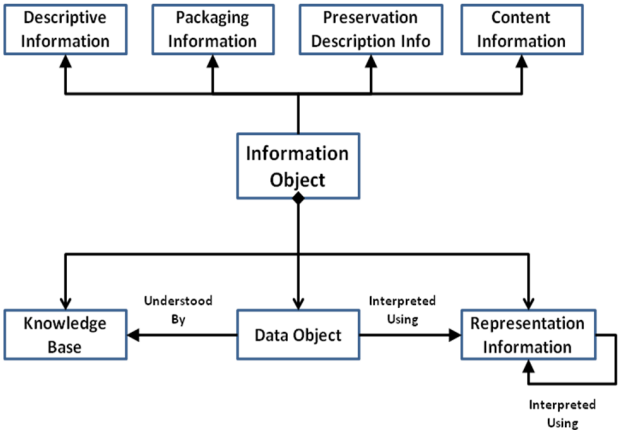

Figure 1: A Partial View of the OAIS Information Model [5]

4.1 Content Information

This is the set of information that needs to be preserved over the long-term. In the case of spatial datasets, it should be the 'original' version of a dataset rather than a domain specific representation of that dataset. For example, in the INSPIRE SDI, where geospatial datasets are mapped on to 'application schema' to represent particular facets of phenomena on the earth as 'geographic features' (e.g. a pan-European road transport network), the source dataset rather than its 'mapped view' should form the 'Content Information'.

4.2 Preservation Description Information (PDI)

This type of information is needed to efficiently manage and preserve a digital object over an indefinite period of time. This includes various information about the life-cycle of a dataset, such as its provenance and versioning history, as well as reference and annotation-related information.

4.3 Representation Information (RI)

This is a component of the Content Information that is required to accurately render a preserved digital object on a future technological platform. This encompasses all levels of abstraction and refers to both the structural and semantic composition, such as recreating the original appearance of the digital object, or analysing it for a concordance [5]. The use of RI can be recursive, especially in cases where meaningful interpretation of one RI element requires further RI (Figure 1). The RI for a dataset may include information about its technical dependencies, such as software required to access the dataset, compatible operating platform and so on.

With respect to an SDI, RI refers to the ability to continue to be able to interpret the semantics of a digital dataset, i.e. how the digital objects relate to a conceptual model of some universe of discourse (ISO 19101:2002 - Geographic information -- Reference model). For instance, a transport network dataset stored in a geo-database or a Shapefile[13] will be meaningless unless the tables or digital objects can be interpreted as 'road features' defined in a relevant conceptual model.

4.4 Packaging Information

This type of information is used to bind a data object and its associated metadata (such as PDI and Descriptive Information) into an identifiable unit or package for preservation. For example, if a data object is compressed before being ingested into an archive, the packaging information for that dataset would include information about the underlying structure of its compressed form.

4.5 Descriptive Information

The information needed to facilitate efficient discovery and accessibility of a preserved data object, typically through search and retrieval facility provided by the long-term preservation archive. Descriptive information about a data object may be derived from its PDI and other metadata. For a spatial dataset that is exposed as a 'feature type' through for example, an OGC standardised Web Feature Service (WFS)[14], the descriptive information could include the information (e.g. keywords, abstract) about that 'feature type' provided in the 'GetCapabilities' document of the WFS.

[11] European Space Science (ESA) Long-Term Digital Preservation (LTDP) Programme - http://earth.esa.int/gscb/ltdp/. (See also: http://www.digitalpreservationeurope.eu/publications/briefs/dp_for_longterm_environmental_monitoring.pdf)

[12] ESA HMA Standards - http://earth.esa.int/gscb/HMAstandards.html

[13] Shapefile - http://en.wikipedia.org/wiki/Shapefile

[14] OGC Web Feature Service - http://www.opengeospatial.org/standards/wfs

4.6 Designated Community/ Knowledge Base

This encompasses all identified potential consumers (e.g. human, software application etc.) to whom the preserved data object is beneficial in terms of its accurate interpretation and proper utilisation. The level of recursion for a particular element of representation information (RI) about a data object is likely to depend on the level of knowledge that the designated community has about that element. For example, if the designate community has considerable understanding of the OGC Web Feature Service, then the representation information of a dataset that is exposed through WFS as 'feature types' could just include the service name – 'OGC Web Feature Service'. Conversely, if the designated community has no understanding of WFS, the representation information of such dataset would have to include detailed implementation and use specification of the OGC WFS among other related information.

A generic viewpoint assumption in an SDI for long-term preservation would define the user community of the SDI as the OAIS 'designated community', with the semantics of harmonised conceptual models that enable domain-specific representation (e.g. 'feature types') of a spatial dataset within the SDI constituting the OAIS 'knowledge base'.

5. A Preservation-aware Spatial Data Infrastructure

We have analysed the INSPIRE architecture in the context of the OAIS reference model with a view to determining the requirements for a preservation-aware SDI. Functionally, INSPIRE consists of the following components (Figure 2):

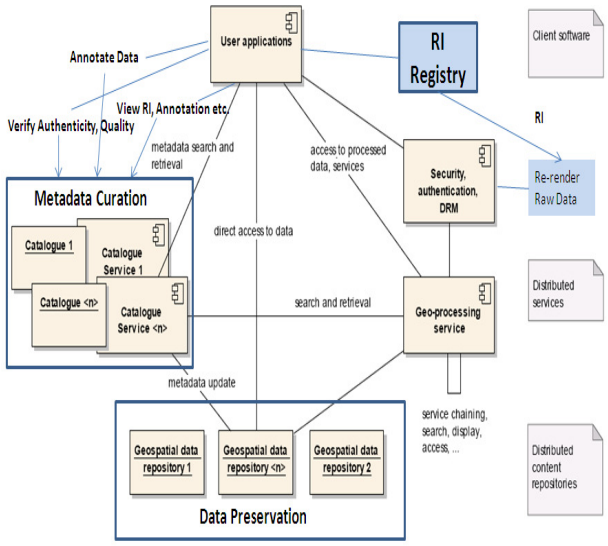

Figure 2: A Preservation-aware SDI

- **Geospatial Data repositories** made available and maintained by different member states and other approved data providers.
- **Metadata catalogues** containing metadata - additional information about the data held in the repositories, typically provided by the data provider(s) - based on the ISO 19115 metadata model to enable efficient discovery of the data exposed through the repositories.
- **Geo-processing Web services** to enable accessing, analysis and processing of the data discovered using the metadata catalogues; includes view and download services.
- **User applications**, i.e. client software to enable users to search the metadata catalogues in order to locate datasets for further processing and/or analysis using the geo-processing services as required.

An analysis of the applicability of the OAIS reference model to the INSPIRE SDI identifies the following three core requirements for ensuring sustained accessibility and usability of the data exposed through such SDIs.

5.1 Long-term preservation of geospatial data repositories

An effective and coherent approach is required to preserve the individual data repositories made available through the SDI over the long-term (Figure 2 – "Data Preservation" box). This needs to address various complex issues, such as compatibility of data with future repository technology and ensuring its continued access even after its provider has ceased to exist. While this aspect is provider-specific, and dependent on the adoption of suitable preservation policies and strategies, it should be possible for the repository owners to identify, define and adopt a set of common fundamental concepts or principles of archiving geographical data over the long-term. INSPIRE can play an important role in defining and promoting such preservation concepts and principles, or at the very least, creating an awareness of the importance of long-term preservation of geographical data among the data providers.

5.2 Preservation-aware Metadata Model

The ISO 19115 metadata model adopted in the INSPIRE SDI is comprehensive enough for capturing enough of the context surrounding the data (for example, data quality, maintenance, use/processing) to enable its effective discovery. However, the metadata elements defined in ISO 19115 do not capture other important preservation-related metadata specified in the OAIS Reference model, such as PDI and RI (Section 4). For example, the ISO 19115 model does not address the mappings between a source geospatial data set and its canonical representation, which typically describes particular facets of phenomena on the earth as 'geographic features'. Such 'feature-based' representation of a geospatial dataset is usually described by an appropriate 'application schema' and exposed by the INSPIRE SDI. This type of information is a significant aspect of a geospatial dataset's RI, without which accurate interpretation and re-use of the dataset on a future technological platform may not be possible.

Therefore, a preservation-aware SDI would require a preservation-focused metadata model that would help capture accurate and sufficient description of all aspects (including the aforementioned preservation-related aspects) of a geospatial dataset as well as well as being flexible for addition of future requirements. However, as RI of a dataset could be highly complex and detailed (depending on the requirement of the designated community), it may be sufficient for a preservation metadata model for a SDI to include only an overview of the RI associated with a dataset. Access to the complete set of RI could be provided through a RI repository or registry (Figure 2), if supported by the data provider. There are other benefits in adopting such an approach that are discussed in Section 7.1.

5.3 Long-term curation of metadata catalogues

The metadata catalogues (Figure 2 – "Metadata Curation" box) are instrumental in facilitating discovery of the datasets held in the repositories by enabling searching of the metadata that describe those datasets. However, without curation - proper management, quality assurance and preservation - the metadata, too, may become unusable over time (Figure 2 – "Metadata Curation" box). For example, it may become out of step with the data that it describes. Therefore, it is also crucial to apply effective long-term curation measures to the metadata catalogues within an SDI [8].

6. PRINCIPLES OF ARCHIVING GEOGRAPHICAL DATA

As mentioned before, the data providers of large-scale SDIs, such as INSPIRE should benefit from a set of common and practical principles applicable to the task preserving geographical data over the long term.

In recognition of the importance of long-term archiving of geographical data in Europe, a number of National Mapping Agencies, archives and research councils across Europe formed the EuroSDR Geographic Data Archiving working group in 2010. Since its inception, the group has been working together to identify, articulate and address the challenges faced by European data providers for preserving their geographical data. As an outcome of this exercise, the group has recently defined and agreed upon a set of common and practical fundamental concepts and principles of archiving geographical data.

Here, we outline a selected few of these principles as agreed by some of the important European National Mapping Agencies and archives who expose their geographical data through SDIs like INSPIRE.

The order of the principles follows the lifecycle of data from creation to maintenance, archival, preservation to accessing archived data. Notably, more generic and comprehensive conceptualisations of the lifecycle of an archive already exist. For example, the Draft DCC Curation Lifecyle Model has been designed to facilitate a lifecycle approach to the management of digital materials in an archive, and to enable their successful curation and preservation from initial selection for reuse and long-term preservation [11]. The principles presented here are the outcomes of a preliminary exploration of the applicability of these existing models to Geographical archives.

Suggested action points in the principles are indicated by this symbol: ►.

Principle 1: *Archiving of digital geographic information begins at the point of data creation, rather than at the point of withdrawal from active systems.*

Today archiving is often seen as an afterthought, though the long term value of a dataset can often be appraised at the outset. If this is done, archival requirements are clear from the start and can be acted upon.

► Define whether long term preservation is desired or necessary, determine and document the retention period. This can be changed at a later date if requirements change but will clarify archival needs from the outset. It should also be done for all existing datasets.

Principle 2: *Establishment and agreement of a common preservation planning process and a set of common preservation objectives between data producers and archives is the backbone for any archiving business case.*

► An archive should look across borders and beyond its domain, and consult other experts to formulate an efficient preservation strategy. Using a common vocabulary and reference model (such as the OAIS model) will improve clarity and understanding. One of the key goals of a long term archiving/preservation strategy is risk mitigation against loss and corruption.

► The preservation objectives of an archive should be defined and articulated in its archival policy. The policy should cater for the requirements of both data providers and future users (the so-called designated community).

► A good governance regime is needed to be established to ensure that the policy is implemented in the foreseeable future.

Principle 3: *Be selective and decide what to archive and what to lose.*

Archiving is an economic issue, as well as a technical challenge. Long term benefits are likely to be intangible, so it is advisable to concentrate on short and medium term benefits. Long term archiving may prove to be less challenging if the medium term actions are considered, prepared and undertaken well. The survival rate for data might be better if less material is archived well, than a vast amount of material being archived poorly.

► An archive should define for each dataset, product or feature group, the required retention period. It should also preserve the documentation that explains what it has chosen to lose and why. This means that it needs to be explained why which aspects of a dataset are important in the shorter and longer term (collection policy).

Figure 3: Geo-archiving Lifecycle

Principle 4: *Consider archiving timeframes of 1, 10, 100 years*

1 year, operational archives focus on short term needs, proprietary formats and specialist solutions may be appropriate.

10 years, a strategic, internal business archive, the focus should be on reusability and access of data. This builds a bridge between shorter term data provider's needs and archivists' needs.

100 or even 1000 years, long-term archive aimed at preservation. Focus on robustness against data loss and corruption, ability to curate and migrate. Data preferably held in flat files, open format.

► Planning should be made to shift data between these archives which may be based on different technical solutions. Access to the 100 year archive can be through a replicated data in a 10 year archive.

Principle 5: *The output of the planning process should also be preserved over the long-term to accommodate future preservation requirements.*

► The documents describing the archival planning process and policy need to be linked to the geographic data in order to provide the context for decision made at the time at or before ingestion of data into an archive.

Principle 6: *Archiving is not backup.*

► It is necessary to backup an archive on at least two uncorrelated storage systems. One backup system should be at a remote and secure site.

Principle 7: *Geographical data should be preserved in a way that non geo-specialists can handle it.*

The likelihood that data survives and can be accessed will be higher if data is structured in a way that archivist are familiar with from other, non-geospatial mainstream content.

► Document migrations, format, and structure so it can be understood by archivists and curators.

► Document the motivation behind applying certain preservation action (e.g. migration) to the data. This type of information forms the preservation history of a dataset and may assist future archivists in understanding and determining the updated preservation requirements for that dataset.

► Also archive data specifications, definitions of coordinate systems and anecdotal material that will help to interpret and understand the data at a later point in time.

Principle 8: *Ensure effective management and quality assurance of the metadata associated with your data.*

► Define the types of metadata needed to enable efficient discovery, accurate rendering, understanding and re-use (e.g. significant properties), and effective preservation of your data over the long-term

► Use appropriate, widely-adopted metadata standards and formats (e.g. ISO 19115, Dublin Core[15], ISO 23081[16])

►Metadata stored in the archive should be both syntactically and semantically valid. For example, an XML-based metadata record can be validated the corresponding XML schema to ensure structure validity. Semantic validation is more complex, and may involve the use of controlled vocabulary defined by the archive, preferably through collaboration with the user community.

► Apply appropriate and efficient versioning mechanism to manage changes made to the metadata in the archive over time.

[15]Dublin Core Metadata Elements Set - http://dublincore.org/documents/dces/

[16] ISO 23081: Records Management Processes - Metadata for Records

► Consider enabling the users to annotate the metadata in the archive to facilitate adding value to the metadata.

► Define a set of broad and high-level principles that form the guiding framework within which the metadata curation (management) can operate. The metadata curation policy would normally be a subsidiary policy of the archival data preservation policy statements and should have reference to the rules concerning legal and other related issues regarding the use and preservation of data and metadata, as governed by the data policy statements.

7. A PRESERVATION PROFILE OF THE ISO 19115 METADATA MODEL

As identified in the analysis of the INSPIRE SDI above (Section 5), the ISO 19115 metadata model is not sufficient for capturing and providing the users with the information needed to enable accurate interpretation of geospatial data in the future. To address this issue, we have developed a preservation profile of ISO 19115 based on the metadata requirements specified in the OAIS reference model and the PREMIS data dictionary[17]. The rationale of this profile is to enable recording preservation-related information about a geospatial dataset, while retaining the ability of the core ISO 19115 model to capture descriptive and contextual metadata about that dataset. The preservation profile incorporates the following key preservation concepts into the core ISO 19115 model as shown in Figure 4 below.

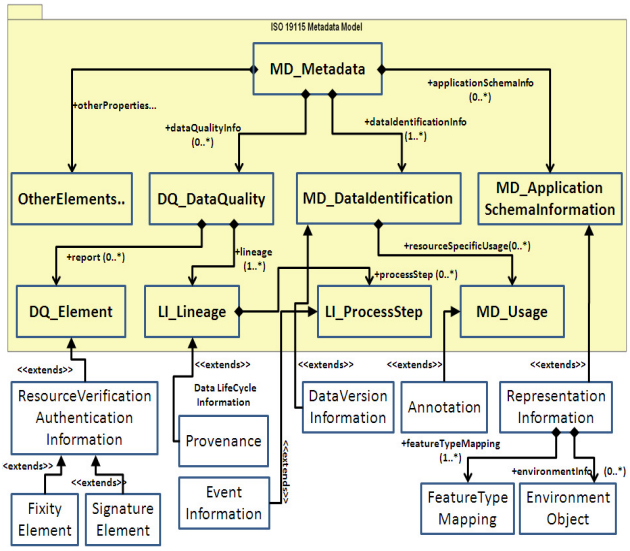

Figure 4: A preservation profile of ISO 19115 Metadata Model

7.1 Representation Information

The OAIS reference model defines the Representation Information (RI) about a digital object as the information required to enable access to preserved digital objects in a meaningful way [5]. In ISO 19115, the only notable RI related information defined is the information about the application

[17] A framework for defining and describing a set of core preservation metadata (based on the OAIS reference model) that would be required to facilitate a long-term data preservation process in a digital archive [9].

schema(s) (i.e. the ***MD_ApplicationSchemaInformation*** class – Figure 3) used to create a particular feature view of a source geospatial dataset. The preservation profile extends this concept to incorporate information about the mappings between the source data and application schema along with the applications/software/services required to effectively apply the mappings (Figure 4).

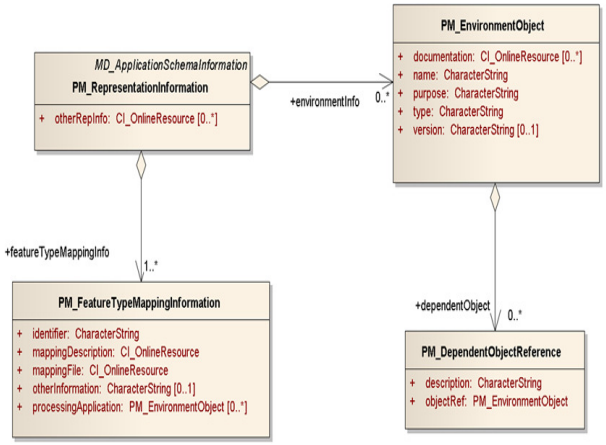

Figure 5: Representation Information elements of the preservation profile of ISO 19115 Metadata Model

In particular, as illustrated in Figure 5, the preservation profile defines the ***PM_FeatureTypeMappingInfo*** class to record information about the mapping(s) between a source dataset and its canonical 'feature-based' representation. The preservation profile also defines additional elements (***otherRepInfo*** and ***environmentInfo*** properties of ***PM_RepresentationInformation*** class – Figure 4) to enable capturing other data specific RI (e.g. data formats, storage media), in the form of web-accessible resources (through HTTP URLs). It is envisaged that detailed RI about a geospatial dataset may not directly benefit its typical users, as they are likely to rely on the current data provider or preservation body to make the data available to them, generally through web services, which apply the aforementioned mappings.

Nevertheless, this approach provides the users with the option to access the RI (made available on the web through e.g. a RI registry by the data provider/preservation body) about a dataset, which, if necessary, could be used to reconstruct and re-use that dataset on a future technological platform (Figure 2). From an archivist's perspective, it is an important mechanism for providing access to the data in a consistent manner into the future. As well, it provides flexibility in terms of the metadata model/format used to capture data-specific RI without being constrained by the ISO 19115 model.

7.2 Data life cycle information

Detailed information about changes (e.g. change of ownership or archive) and events occurring during the life-cycle of a dataset is essential for verifying the provenance of a dataset as well as the reliability of its preservation in the future. In addition, this type of information could contain a detailed history of every preservation measure (e.g. migration) applied to a dataset during its lifecycle, in order to assist its future curators in understanding and determining the updated preservation requirements for that dataset. For instance, a provider may choose to migrate an existing road transport dataset into a new database schema more closely reflecting an INSPIRE application schema (a process sometimes known as 'Extraction-Transformation-Load', or ETL); it is important to document this schema transformation for preservation purposes. Similarly for quality assurance purposes it is important to be able to verify the history of ownership of a dataset.

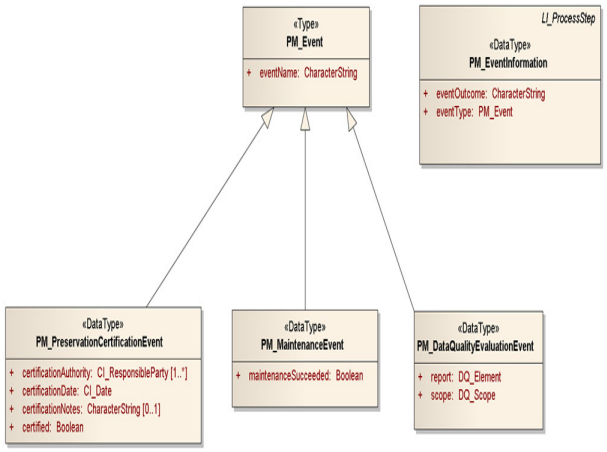

Figure 6: Dataset Event information elements of the ISO 19115 Preservation Profile

With this in mind, the preservation profile extends the ***LI_Lineage*** and ***LI_ProcessStep*** elements (Figure 4) defined in the ISO 19115 model to capture detailed information about the lifecycle of a dataset. The dataset lifecycle information in the preservation profile is divided into two main categories: **Dataset Provenance Information**, (i.e. change of ownership and/or preservation body) and **Dataset Event Information** (i.e. all major events, including preservation-related ones, such as major platform change and preservation certification process that have affected the data during its life cycle - useful for audit trailing and quality checking purposes).

Important among these elements is the ***PM_PreservationCertificationEvent*** (a specialised ***PM_Event*** class shown in Figure 6) defined to provide information about any certification examination(s) conducted, to ensure adequacy of the preservation measure(s) applied to a dataset. This should provide the users with some level of confidence in the preservation method(s) applied to, and consequently, in the longevity of the data of their interest. In the OAIS, this type of information is referred to as 'Preservation Descriptive Information' (See Section 4.2).

7.3 Data Authenticity Verification Information

The ISO 19115 model adopts a number of data quality related concepts (e.g. ***DQ_Elements*** – Figure 4) from the ISO 19113[18] and 19114[19] standards (for representing the quality principles and evaluation procedures associated with geographic information) in order to provide detailed description of the quality assurance measures applied to a dataset. The

[18] ISO 19113:2002 - Geographic information -- Quality principles

[19] ISO 19114:2003 - Geographic information -- Quality evaluation procedures

preservation profile adds to this the ability to verify unauthorised modifications to a dataset by recording its fixity information, such as a checksum and digital signature. This may be important, for instance, where major asset management or security programmes depend on the accuracy of information in a dataset, and it is important to be sure that data has not been altered.

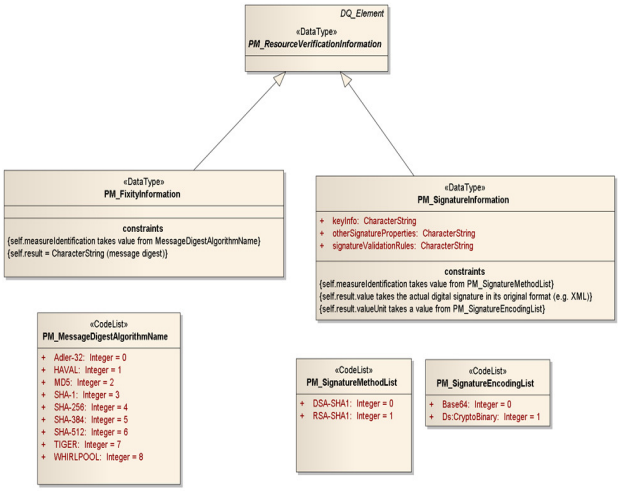

Figure 7: Resource Authenticity Verification information elements of the ISO 19115 Preservation Profile

As illustrated in Figure 7, the preservation profile defines the ***PM_ResourceVerificationInformation*** class as a specialised ***DQ_Element*** class (of ISO 19115:2003 core). It is intended to record fixity information (***PM_FixityInformation*** class), such as a checksum and digital signature (***PM_SignatureInformation*** class) about a dataset to enable verification of unauthorised alterations made to that dataset.

In the context of the OAIS information model, this type of information is categorised as the 'Preservation Descriptive Information' associated with a dataset.

7.4 Annotation

Annotation in the digital world has long been recognised as an effective means of adding value to digital information. It can, in effect, help establish collaborative links between data providers, data users and a preservation body. Thus, annotation has the potential to facilitate enhanced efficiency of a preservation process, and thereby improve the quality of both data and metadata. However, annotation without the intended context may become meaningless. For example, an annotation may be used to label particular map features with descriptive text, which may contain values of some attributes associated those features [7]. These attribute values alone, i.e. without the correct association with the corresponding map features (the annotation context) would be meaningless. For more complex and dynamic geographical datasets, it may be useful for users to be able to annotate specific features or attributes for collaborative analysis or interpretation, for instance in an emergency response scenario. While not directly related to preservation, it is not difficult to appreciate the long-term value of such information, e.g. during post-disaster audit of response capability.

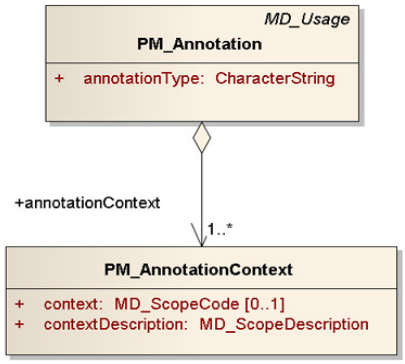

Figure 8: Annotation elements of the ISO 19115 Preservation Profile

Therefore, the preservation profile defines as extensions to the ***MD_Usage*** elements of the core ISO 19115 (Figure 8) a number of suitably structured elements to capture detailed annotation related information (***PM_Annotation*** class) with traceability to the data context (***PM_AnnotationContext*** class) to which the annotation refers.

7.5 A Test Case

We tested the preservation profile of the ISO 19115 by recording preservation metadata about some weather observation datasets exposed by an OGC-compliant Web Feature Service (WFS). This WFS is built on the 'Complex Datastore' version of GeoServer[20], which enables representation of data from a relational database in a GML[21]-based application schema (e.g. Climate Science Modelling Language, CSML[22]) defined independently of the underlying database structure. This special edition of GeoServer was a research endeavour by SeeGrid[23] with contribution from the GeoServer community.

Considering the aforementioned special capability of the WFS, the dataset exposed by it provided ideal examples of 'feature-based' representations of source spatial datasets. Therefore, we used the preservation profile of ISO 19115 to record a number of useful Representation Information (RI) about some of the datasets served up by the WFS. This RI captured included the mappings used to generate a "feature-based" canonical representation of a dataset as well as other metadata. The following XML snippet provides an example of such an RI:

```
<geop:PM_RepresentationInformation>
        <geop:featureTypeMappingInfo>
                <geop:PM_FeatureTypeMappingInformation>
                        <gco:identifier>
```

[20] GeoServer, an open source Java-based web server that provides a suitable means of promoting and publishing Geospatial information on the web using various OGC standards - http://geoserver.org/display/GEOS/Welcome [Accessed 1 February 2011]

[21] Geography Markup Language is an XML grammar written in XML Schema for the description of application schemas as well as the transport and storage of geographic information - http://www.opengeospatial.org/standards/gml [Accessed 1 February 2011]

[22] http://ndg.nerc.ac.uk/csml/ [Accessed 1 February 2011]

[23] https://www.seegrid.csiro.au [Accessed 1 February 2011]

```
                                        <gco:CharacterString>
                                        mapping1
                                        </gco:CharacterString>
                                </gco:identifier>
                        <geop:mappingDescription>
                                        <gmd:CI_OnlineResource>
                                                <gmd:linkage>
                <gmd:URL>http://www.stfc.ac.uk/geopres/mappings/dataset1/descr
iption.html</gmd:URL></gmd:linkage>
                        </gmd:CI_OnlineResource>
                                </geop:mappingDescription>
                                <geop:mappingFile>
                                        <gmd:CI_OnlineResource>
                                <gmd:linkage>
                <gmd:URL>http://www.stfc.ac.uk/geopres/mappings/dataset1/
mapping.xml </gmd:URL></gmd:linkage>
                        </gmd:CI_OnlineResource>
</geop:mappingFile>
<geop:processingApplication>
                <geop:PM_EnvironmentObject>
                        <geop:documentation>
                                <gmd:CI_OnlineResource>
                                        <gmd:linkage>
                <gmd:URL>http://www.stfc.ac.uk/geopres/mappings/dataset1/
application.html </gmd:URL>
                </gmd:linkage></gmd:CI_OnlineResource></geop:documentation
>
                <geop:name><gco:CharacterString>GeoServer WFS
                </gco:CharacterString></geop:name>
                <geop:purpose>
                <gco:CharacterString>produces representation of STFC sample
weather observation datasets in  Climate Science Modelling Language - a
GML-based application schema</gco:CharacterString></geop:purpose>
<geop:type><gco:CharacterString>Software</gco:CharacterString></geop:type
>
<geop:version> <gco:CharacterString>Complex
Datastore</gco:CharacterString></geop:version>
</geop:PM_EnvironmentObject>
</geop:processingApplication>
</geop:PM_FeatureTypeMappingInformation>
</geop:featureTypeMappingInfo>
</geop:PM_RepresentationInformation>
```

Listing 1: an example of Representation Information recorded using the ISO 19115 Preservation Profile

Of particular note in the above XML snippet is the 'CI_OnlineResource' related metadata elements, such as 'mappingFile' and 'processingApplication'. These elements are defined to record references to web-based resources providing more comprehensive (and possibly complex) information about the aspects of the data that they represent. In the above XML snippet, the 'processingApplication' element points to a web-based document providing detailed information about the GeoServer WFS, such as the input parameters and computer platform required to apply the mappings (described by the 'mappingDescription' and 'mappingFile' elements) to the corresponding dataset. These web-based resources could be encoded in any format chosen by the preservation body concerned. Thus, the preservation profile of ISO 19115 provides flexibility in terms of the metadata model/format used to capture data-specific RI without being constrained by the ISO 19115 model while ensuring the accessibility of such information in a uniform and coherent manner.

8. CONCLUSIONS AND FUTURE DIRECTION

Long-term preservation of geographic data exposed through uniform and interoperable SDIs is not currently addressed in the INSPIRE Directive but is highly important for applications that require continued access to both current and historical data e.g. for monitoring climate change. The main drivers for archiving digital geographic information are meeting legislative requirements, the short and long term exploitation of archived data as well as efficiency savings in managing superseded datasets. This paper has attempted to set out the path and describes what needs to be done now to future-proof the investment government agencies around the world have made in digital Geographic Data.

In this paper, we have investigated the requirements for ensuring sustained access to geographical data from the perspective of a preservation-aware and INSPIRE-conformant SDI. We have also outlined a number of principles for the long term retention and preservation of digital geographic information defined by the EuroSDR Geographic Data Archiving working group with a view to introduce fundamental concepts of digital geographic data archiving for the public sector information providers in Europe. In addition, we have presented a preservation profile of the ISO 19115 metadata standard to enable an archive to record preservation-related information about geo-data and make it available to the users through the associated SDI.

Future work in this area would need to focus on the implementation of efficient and interoperable preservation solutions for the data repositories made available through the SDI. To that end, the EuroSDR group aims to define a reference implementation profile of the OAIS reference model for geographical data based on the practical preservation related use-cases extracted from the participating archives and NMAs. A key consideration of this work will be to consider risks and issues for curation and preservation of geographic data throughout the archival phase of its lifecycle [11]. The group will also work towards refining the preservation principles presented in this paper through broader engagement with the NMAs and archives as well as other preservation-related endeavours in Europe.

9. ACKNOWLEDGMENTS

The work presented in this paper was funded in part by the e-Science centre, STFC.

10. REFERENCES

[1] Hoebelheinrich, N. and Banning, J., 2008. An Investigation into Metadata for Long-term Geospatial Formats. *NGDA* Report, (2008) URL= http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp08/docs/session7_hoebelheinrich_paper.doc [Accessed 4 February 2011]

[2] Janée, G., Mathena, J. and Frew, J., 2008. A Data Model and Architecture for Long-term Preservation. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL),*134–144. (2008) DOI:http://dx.doi.org/10.1145/1378889.1378912

[3] McGarva, G., Morris, S. and Janée, G., 2008. Preserving Geospatial Data, *Technology Watch Report, Digital Preservation Coalition (DPC), DPC Technology Watch Series Report* 09-01. (2008) URL= http://www.dpconline.org/technology-watch-reports/download-document/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee.html [Accessed 4 February 2011]

[4] Morris, S. P., 2006. Geospatial Web services and geoarchiving: New opportunities and challenges in geographic information services. *Library Trends*, 55, pp. 285-303. (2006) URL= http://www.lib.ncsu.edu/ncgdap/documents/MorrisLibraryTrendsFall2006.pdf [Accessed 4 February 2011]

[5] CCSDS, 2002. Reference Model for an Open Archival Information System (OAIS). *Recommendation for Space Data Systems Standard, Consultative Committee for Space Data Systems (CCSDS) Blue Book. (2002)* URL= http://public.ccsds.org/publications/archive/650x0b1.pdf [Accessed 4 February 2011]

[6] GER, 2005. Data Model for Managing and Preserving Geospatial Electronic Records Version 1.00 [, *Center for International Earth Science Information Network (CIESIN) Columbia University*. (2005) URL= http://www.ciesin.columbia.edu/ger/DataModelV1_20050620.pdf [Accessed 4 February 2011]

[7] Bose, R and Reitsma, F., 2005. Advancing Geospatial Data Curation, Conference on Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data, *online papers archived by the Institute of Geography, School of Geosciences, University of Edinburgh*. (2005) URL= http://www.era.lib.ed.ac.uk/bitstream/1842/1074/1/freitsma003.pdf [Accessed 4 February 2011]

[8] Shaon, A and Woolf, A. 2008. An OAIS Based Approach to Effective Long-term Digital Metadata Curation, *Computer and Information Science*, 1(2), 2-12. (2008) URL= http://www.ccsenet.org/journal/index.php/cis/article/download/90/79

[9] PREMIS, 2008. PREMIS Data Dictionary for Preservation Metadata, version 2.0 , *PREMIS Editorial Committee*, (2008), URL= http://www.loc.gov/standards/premis/v2/premis-2-0.pdf [Accessed 4 February 2011]

[10] Bos, M, Gollin, H, Gerber, U., Leuthold, J. and Meyer, U. 2010. Archiving of geodata, A joint preliminary study by swisstopo and the Swiss Federal Archive, *SWISS Archive,*(2010),URL= http://www.swisstopo.admin.ch/internet/swisstopo/en/home/topics/geodata/geoarchive.parsysrelated1.59693.downloadList.93958.DownloadFile.tmp/preliminarystudyarchivingofgeodata.pdf [Accessed 11 September 2011]

[11] Higgins, S. 2008: The DCC Curation Lifecycle Model, *The International Journal of Digital Curation.* Issue 1, Volume 2 (June. 2008), URL= http://www.ijdc.net/index.php/ijdc/article/viewFile/69/48 [Accessed 29 September 2011]