Computer Standards & Interfaces 36 (2014) 909–917

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

Point of view: Long-Term access to Earth Archives across Multiple Disciplines

Jinsongdi Yu [a,*], Peter Baumann [a], Dimitar Misev [a], Piero Campalani [a], Mirko Albani [b], Fulvio Marelli [b], David Giaretta [c], Shirley Crompton [c]

[a] Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany
[b] EuropeanSpaceResearchInstitute/EuropeanSpace Agency, Via Galileo Galilei, CasellaPostale 64, Frascati, Rome, Italy
[c] Science and Technology Facilities Council, Rutherford Appleton Laboratory, Didcot, UK

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form 16 December 2013
Accepted 1 April 2014
Available online 13 April 2014

Keywords:
Interoperability
virtualization
coverage
multi-disciplinary
data preservation

ABSTRACT

Without an approach accepted by the communities at large, domain disagreements will continue to thwart current global efforts to harmonize information models. The research presented here reviewed current standardization activities. A number of observations and possible solutions are proposed to address the topic of standardizing long term access to multi-discipline Earth System archives by considering the application of the knowledge base concept to facilitate data interpretation. Finally, we present a case study as an initial entry point for the further discussion about standardization.

1. Background

Multi-disciplinary Earth System Science research often involves the use of unfamiliar geo archives from different domains. Raw data, derived products or representation data are delivered either by offline ordering systems or online and web service based delivery systems. However, the heterogeneity of archiving models employed in these systems tends to limit the interoperability of the data and hence their usefulness in today's highly multidisciplinary Earth system science research. Due to continuous technology change and research development, data access technologies come and go. Although the archived bits may remain the same, the information or knowledge encapsulated by these bits and bytes may be lost. Goodchild et al. [1] argue that scientific grounded information about the planet's future should be fully understood and absorbed. To ensure these archives are both sustainable and sustained for the long term, the Research Libraries Group (RLG) and Commission on Preservation and Access (CPA) formed a Task Force on Archiving of Digital Information to investigate the means of ensuring "continued access indefinitely into the future of records stored in digital electronic form" [2]. The publication of the final report marked an important point for the digital preservation community and has proved to be a fundamental document identifying core challenges of digital preservation [3].

To standardize digital preservation practice and provide a reference model for repositories, the Consultative Committee for Space Data Systems (CCSDS) developed the Open Archival Information System (OAIS) Reference Model to provide a framework for the standardization of long-term preservation which is applicable in any domain or context beyond its initial space science community [4]. Instead of defining concrete metadata standard, the model provides an abstract framework for designing archival systems or repositories, including all technical aspects of a digital object's life cycle, from ingestion to distribution. In terms of interpreting the information encoded within an object's bitstream, representation information maps a Data Object into more meaningful concepts [4]. A specific example provided by Mbaye [5] is using the Data Request Broker (DRB) [6] API to handle the ENVISAT product data [60], regardless of physical formats, for the extraction and interpretation of relevant product information.

Building on this example, long term access to Earth Science data needs to involve a multi-disciplinary interoperability approach, as the community evolution will always result in a change of the disciplinary knowledge base. Standardization work by combining domain best practices and making them generally accepted within communities at large has the potential of achieving long-term global sharing of geospatial information in the heterogeneous world of Earth archives. In this paper,

* Corresponding author.
E-mail addresses: j.yu@jacobs-university.de (J. Yu), p.baumann@jacobs-university.de (P. Baumann), d.misev@jacobs-university.de (D. Misev), p.campalani@jacobs-university.de (P. Campalani), Mirko.Albani@esa.int (M. Albani), fulvio.marelli@esa.int (F. Marelli), david.giaretta@stfc.ac.uk (D. Giaretta), shirley.crompton@stfc.ac.uk (S. Crompton).

http://dx.doi.org/10.1016/j.csi.2014.04.001
0920-5489/

possible approaches to achieve long-term interoperability of existing archives are investigated, on both the horizontal and vertical domain, as well as beyond the Earth science disciplines.

2. State of the Art

2.1. Standardized geospatial models

In Earth science, GIS data models [7] are designed to capture, manipulate, analyze, manage, and present all types of geographical phenomena, such as roads, land use, elevation, trees, rainfall amount, etc. Traditionally, there are two methods to abstract these geographical features: vector data and raster images. Points, lines, and polygons are the vector data which mathematically describe the location of each vertex in *Coordinate Reference System*; while images or arrays are the raster data whose cells' geographic location is implied by their position in the array matrix. Large multidimensional data are being collected by sensors and by humans [8]. These can be from any number of sources, largely unknown and unlimited, and stored in diverse formats for optimizing either read-write efficiency, security enhancement, or data interchange. "This richness of alternatives is more a curse than a blessing since it has created the confusing and apparently chaotic variety of Geographic Information System (GIS) data structures and formats now confronting GIS users"[9]. A standardized Geographical Feature [9] is adopted to provide foundation models that order the chaos and bridge real-world phenomena and their representation as a collection of Features with Geometry. The abstraction is built on the basic concepts of geometry, reference system, relations, quality, metadata, etc. [10]. Common open model languages, such as Geography Markup Language (GML) [11], JSON Geometry and Feature Description [12], and Keyhole Markup Language [13], which implement or partially implement this abstraction, are widely adopted as *de facto* interchangeable formats for geographic features.

Nevertheless, domain experts have different preferences, e.g., all NASA Earth Observing System (EOS) data products [14] use Hierarchical Data Format (HDF) as the standard data format, while the Standard Archive Format for Europe (SAFE) [15] has been designed to act as a common format for archiving and conveying data within ESA Earth Observation archiving facilities [44]. Further commonly used data formats, such as ESRI shape files, Network Common Data Form (NetCDF) [16] and GeoTIFF formats, are also used to deliver data across the various Earth science disciplines. Actually, there are rich formats, eg. JPG2000 [57] and PNG [58], available beyond these fairly restricted set of data formats. Efforts have been invested into binding existing archives with the standardized models to improve interoperability among various Earth Science domains. For example, Nativi [17] has investigated the mapping model from the Common Data Model of the Unidata [18] to the ISO 19123 coverage [19], which is a special case of Geographical Feature. Based on this research, practical experiments [20] have be taken by GALEON IE [21] to provide interoperable and standard-based solutions [22] for datasets up to 5D and bridge the gap between the atmospheric, oceanographic and GIS communities. The approach provides a model level mapping for generic access interfaces which are independent of how the data are stored physically.

2.2. Standardized packaging models for LTDP

OAIS defines data as any type of knowledge that can be exchanged, and representation information, which maps the data into more meaningful concepts. The Archival Information Package is defined as the data plus representation information plus the additional information needed to support claims of authenticity. The CCSDS/ISO XML Formatted Data Units (XFDU) packaging format is consistent with this approach and is provided as a standard to package data and metadata (including software) into a single package (e.g., file or message) to facilitate information transfer and archiving in the space informatics domain [23]. The standard provides a packaging solution to define the information and its behaviors. The solution leaves the freedom of choosing representation information to domain experts.

For example, by considering access Earth observation archives in long term, SAFE [24] implements the CCSDS/ISO XFDU packaging format, wraps or references Earth Observation (EO) data and associate them with information expressed in EO vocabulary. SAFE is designed to act as a common format for archiving and conveying data within ESA Earth Observation archiving facilities. The integrated representation via XFDU opens the door to the access of bit streams without having to consider the heterogeneous data encodings. The approach is particular important in long-term preservation of Earth Observation data and implies an interoperable framework for packaging a large variety of information for multidisciplinary Earth Science communities.

Furthermore, XFDU allows executable behaviors to be associated with content in the content unit of information package. These behaviors may be represented by abstract definitions or references of concrete modules of executable code that implement and run the behaviors defined abstractly by the interface definition. Domain experts make the decision on which behaviors are to be associated to the information content.

Besides XFDU, OAI-ORE [25], is a recommendation that built on the principles of the web architecture. It defines the description of aggregations of web resources and the digital objects of which they are composed. The specification was extended in the SHAMAN project [26], which focuses on OAIS-based packaging of compound digital objects preserved in a distributed storage environment for future use. The Metadata Encoding and Transmission Standard (METS) [27] is primarily a packaging format for metadata and object references without considering the consistency between the encoding and the information. The BagIt File Packaging [28] is yet another simple layout for exchanging generalized digital content, which is used for document encodings and checksum algorithms associated with the contents of a "bag". However, the semantics behind it is ignored.

OGC members are proposing an open, non-proprietary, platform-independent GeoPackage container for distribution and direct use of all kinds of geospatial data. Obviously, using a packaging standard would help improve access to historical geospatial data be it XFDU's information and behavior approach, OAI-ORE's aggregation approach, METS' referencing approach or BagIt's checksum algorithms approach.

3. Approaches

3.1. Communication: data and Information, partial or full interpretation

The three concepts of data, information, and knowledge are often regarded as the basic building blocks of information science field [29]. Nonetheless, the definitions and usage of these terms are not always consistent between leading scholars from different aspects of the information science academic community. As a starting point for building a systematic conception of communication among different fields, it is essential to achieve a standardized conceptions of data, information, and knowledge. OAIS defines these terms as below:

- **Data**: A re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.
- **Information**: Any type of knowledge that can be exchanged. In an exchange, it is represented by data.
- **Knowledge Base**: OAIS does not define Knowledge but instead defines knowledge base, which is a set of information, incorporated by a person or system, that allows that person or system to understand received information.

In an information system, data can (usually) be instantiated as a sequence of bits before it is processed. Proper interpretation by a human or machine process produces information which can be derived and used for communication. Different communities hold different