Exploring the impact of knowledge potential on preservation requirements within an OAIS Archive

Esther Conway, Brian Matthews and David Giaretta

Digital Curation Centre (DCC)
STFC Rutherford Appleton Laboratory
Chilton, Didcot, Oxfordshire OX11 0QX
EMail: e.a.conway@rl.ac.uk

ABSTRACT

The challenge of digital preservation of scientific data lies in the need to preserve not only the dataset itself but also the potential it has to deliver knowledge to a future user community. Thus the definition of knowledge that a data set is capable of imparting to any future designated user community has a profound impact on preservation requirements for an archive.

In this paper we identify key sources for the creation of this knowledge definition and describe how differing approaches to requirements gathering impact the characterisation of information encoded within scientific datasets. We then seek to explore how the definition of knowledge requirements from different sources impacts upon the requirements for an OAIS system.

The role of a data curator is not only to safely archive the dataset but to respond to preservation requirements, collecting and managing sufficient associated information to do this. Using illustrative examples from the scientific test beds identified in the EU project CASPAR, we demonstrate how the quality of the OAIS concepts of representation information, provenance and context is affected in response.

INTRODUCTION

The challenge of digital preservation of scientific data lies in the need to preserve not only the dataset itself but also the potential it has to deliver knowledge to a future user community. Thus the definition of knowledge that a data set is capable of imparting to any future designated user community has a profound impact on preservation requirements for an archive.

CASPAR[1] [1] is a major European project which aims to develop a methodology and toolkit to support digital preservation across a the whole lifecycle, based on the ISO OAIS standard, with use cases taken from a range of Scientific, Cultural, and Artistic domains. As a first stage of this development, the project in conjunction with the DCC[2] [2] undertook a major requirements gathering and analysis process on a number of case studies, including scientific data from atmospheric science and observations of solar interaction with the ionosphere.

The purpose of the process was to extract sufficient information to implement, extend, and validate the OAIS reference model [3], to enhance the techniques for capturing Representation Information and other preservation related information for content objects and to inform the CASPAR software development process. However, it became apparent that the nature of the information adequate to preserve the knowledge inherent in the data is dependent on the particular requirements of the stakeholder involved.

[1] Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR) is an Integrated Project co-financed by the European Union.

[2] Digital Curation Centre a centre funded by the Joint Information Systems Committee, an independent advisory that works with further and higher education establishments, and the UK e-Science core program.

As a result of this work we are able to identify key stakeholder types and how they impact on the definition of knowledge that a data set is capable of providing. Understanding this definition allows a curator to make an informed decision. If this definition is then accepted by the curator as a realistic preservation objective it will provides limits, setting the preservation scope for an archive. This preservation objective is what we consider to be the knowledge potential of an archive which a curator attempts to maintain for a future designated user community. We illustrate how these stakeholders, in addition to having a differing view on the nature of the information which needs to be preserved, also provide much of preservation description material and representation information. Using examples from the CASPAR scientific testbeds we then discuss how, with judicious analysis, an OAIS system is capable of preserving a wide range of knowledge that may be extracted from a scientific data set.

2 KEY STAKEHOLDERS IN A DIGITAL ARCHIVE

A number of key stakeholders are involved in a typical digital archive, each of which has different requirements which need to be accommodated.

2.1 Funding Bodies

Every digital archive will have some form of funding body associated with it to provide the resources to collect and maintain the data. During its lifetime, the custody of a data set may pass through several bodies generating rich documentation which explains the scientific purpose of the dataset and how it has evolved over time. These documents can take the form of experimental proposals which will explain the original intent of the experiment/observation, institutional reports which state the intent of maintaining supply of the data to a scientific community, and reports which show successful scientific output. It is worth noting the limits of such documentation as it will omit scientific potential outside the remit of the organisation. In the case of the CASPAR atmospheric science testbed we can see how different Research Councils are interested in different regions of the atmosphere resulting in the documentation of areas of scientific discovery being omitted from reports.

2.2 Scientific Organisations

Scientific organisations such as university departments or national and international institutes and laboratories, are frequently associated with datasets. They tend to work within a particular branch of science and can provide a great deal of detailed information on how a dataset can fulfil that particular area of scientific potential, providing for example software support materials and field specific bibliographies. However, these scientific organisations, whilst being an excellent source for support for that area of scientific discovery, will naturally neglect other disciplines. In the CASPAR test beds this was particularly evident for emerging and specialist areas of scientific investigation where much knowledge was still embedded in the data using scientific groups and did not have mature organisation supporting them.

2.3 Data Producers

Every dataset will have an individual scientist, or group of scientists responsible for its production. In addition to the scientific intent recorded in an experimental proposal, they will in addition hold information and observations at the time of the experiment/observation which can expand the knowledge potential of the data. This could be event associations with other phenomena for example lighting strikes and ionisation of a region of the atmosphere or identification of recurrent patterns which merit further investigation. In the case of the CASPAR MST testbed we see how the project scientist discovered that a signal signature due to precipitation was present in a dataset traditionally used for wind profiling. In this instance the scientist was able to study this and publish his finding in his paper on “VHF signal power suppression in stratiform and convective precipitation” [4]. This type of material has a tendency not to be formally recorded, sometimes manifesting itself in wiki and web logs. We do however majority of this type of information is at high risk of loss.

2.4 Scientists in the Community

This collection of scientists is the most diverse and distributed. Indeed other groups may be considered to be a subgroup of scientists as their opinion will have been ultimately informed by the larger scientific community. Except in the circumstance of highly specialised datasets with discrete user communities, we would also expect a full survey of the wider data using to be completely unrealistic. The ability to capture such information from an active data using community would be greatly enhanced by the developed of annotation systems such as AstroDAS [5] which permit the annotation of astronomical data allowing for scientific assertions to be captured. Projects such as CLADDIER [6] and OJIMS [7] are also developing ways of referencing and kite marking datasets which will potentially ease the discovery of knowledge associated with datasets.

2.5 Data Archivist

The archivist is the group or individual who is the current custodian of the data. The extent to which they have interacted with other stakeholder groups and extracted knowledge requirement with its associated information will be highly dependent on the resources available to, the motivations, background and personal bias of the individual archivist

The relationship between the areas of knowledge as required by the various stakeholder groups is illustrated in Figure 1.

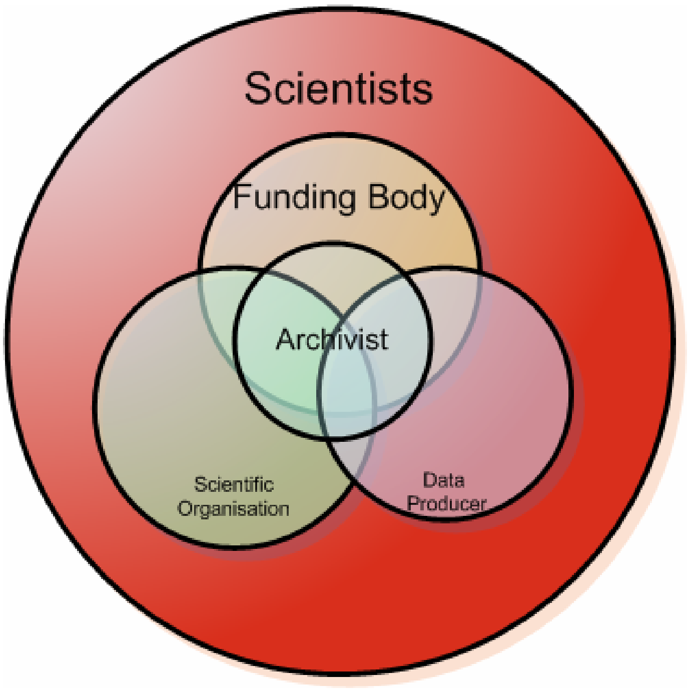

Figure 1: The overlapping area of knowledge as perceived by stakeholder groups

3 KNOWLEDGE REQUIREMENT IMPACT ON OAIS INFORMATION

The OAIS reference model specifies that within an archival system, a data item has a number of different information items associated with it, each performing a different role in the preservation process. For a particular archive, the nature of the information within these information types depends on the requirements of the stakeholders involved. We examine these information types in turn.

3.1 Preservation Description Information

OAIS specifies that information be provided to describe the data set with properties required for preservation. Such Preservation Description Information comes in four types.

Reference Information: Reference information assigns identifiers within identifier schemes to the data, and is independent of the knowledge you wish to extract

Context: Context describes the relationship between data and its environment. The most basic knowledge requirement would need very basic context information where for example a snapshot of the atmosphere at a particular time and location is required. However when this expands, where for example tracking large scale atmospheric phenomena, one may need to establish temporal and spatial relationships between files in a dataset in order to do this. For example the a funding institution such as the Met office requires this type of snapshot information to feed into predictive models but an individual scientist studying Mountain waves would wish to study a type phenomena which takes time to pass over the MST radar site and are approximate 8 miles in length. This may expand again if the information in the data set is required to interact with external data sets, which may for example mean mapping to a different co-ordinate system for example heliocentric and geocentric systems in astronomical data to allow for such interoperability.

Provenance: Provenance information documents the history of the data, what actions were performed on the data, by whom and when. Provenance may for simple requirements be viewed as a special type of context information in the case of snapshot type data for the Met office. However more detailed provenance information may be required if factors such manual or automated scaling may affect the data quality. It was noted that the appearance of a particular phenomena such as the occurrence of a sporadic E-Layer, was most reliably identified by a skilled manual scaler visually inspecting ionograms. This means that future scientists conducting research related to the appearance of the sporadic E should use only data which had been through this type of process. Physical factors such as the type of instrument or its mode of operation may allow for easy identification of different phenomena. In the CASPAR EISCAT testbed we see how the scientific objective of special program experiments influence the instruments mode of operation and in some cases the EISCAT instrument has been operated to respond to events of geophysical interest such as proton events or earthward directed coronal mass ejections. If the preservation scope of an archive encompasses investigation of such events or scientific objectives, adequate province is necessary for the discovery and use of the data. The data may additionally need to be from trusted institutions to ensure a desired level of authenticity. Currently the ingest of data into the Rutherford Appleton Laboratory WDC [8] archive is highly reliant on the archivists appraisal of trusted producers. Maintenance of such an archive needs the addition of provenance information relating to these producers to demonstrate the required authenticity to future users.

Fixity: Fixity information documents the authenticity mechanisms for the data. Fixity information may also be considered to be independent of knowledge requirements except in the case where a specified level of authenticity is part of that requirement. For example if an atmospheric data were required as evidence of a pollution event in a legal case it may be necessary to demonstrate the data had not been tampered with by means of digital signature.

3.2 Representation Information

Representation information in OAIS describes how the information is signified by the data, including what the semantic content is being represented and how that is physically rendered in the data format. Representation Information has three main types.

Structural Information: The required structural information is the minimum information needed to extract and correctly identify the required parameters. The knowledge which is needed to be extracted from the data set determines which level of processed data and parameters from that level of data. These parameters may form part of the content or indeed part of the provenance information for example in the case of station identification or mode of operation for an instrument.

Semantic Information: The level of semantic information required varies according to the level of understanding, interpretation and authenticity which is needed to be attached to the extracted parameter. This ranges from simple definition from communities such as in the ionospheric science use case, CF naming conventions [9], URSII parameter definitions [10] to extensive documents such as URSII handbook of Ionogram interpretations [11]. All of these semantic definitions will additionally evolve over time as user community vocabularies shift.

Other including Higher Level Knowledge: The amount of additional "other" materials needed tends to be the most explosive in reaction to the expansion of knowledge requirements. Typical examples include

- Software including scientific models
- Code documentation, description of algorithms
- Support materials for operation of software
- WebPages including support materials, educational materials, non technical documents for consumption by general audience, information packs and background documents
- Subject specific bibliographies and texts

Variation in knowledge requirement presents a curator with differing preservation options.   It affects the type of information you need to preserve, the types of organisation you will be required to source it from, it's classification and time dependencies. The curator must perform a cost benefit analysis in reaction to this. The cost to an archive of actively managing and maintaining the required OAIS information needs to assessed against the benefit gained from preserving the knowledge one is capable of extracting from the data set. OAIS does not set the scope of preservation but rather helps clarify what you are aiming to preserve and how you may wish to accomplish this.

Figure.2 illustrates some of the typical sources of OAIS information from stakeholders, together with their dependencies and how they may be potentially classified to the key information types within an OAIS system.

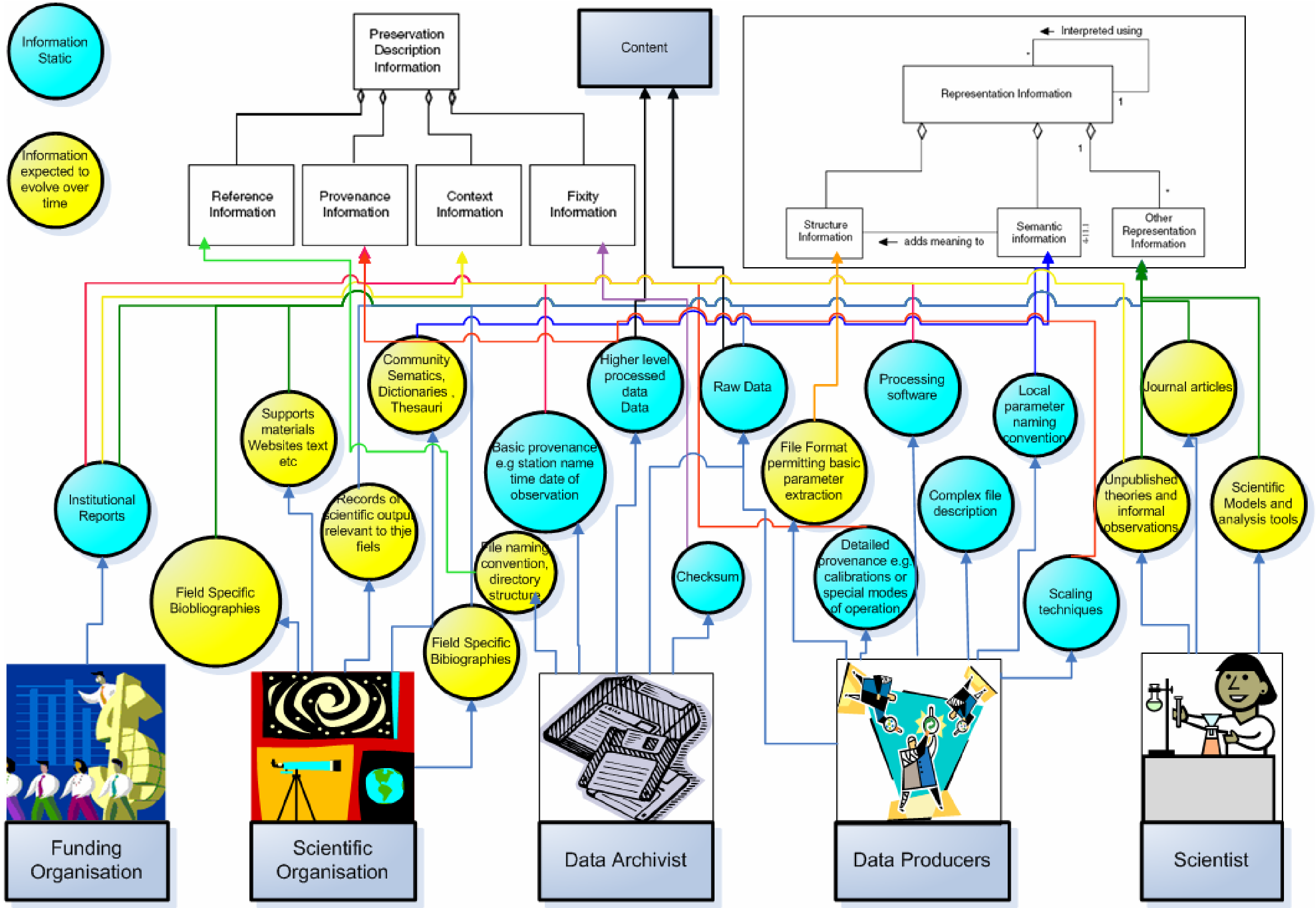
Information Static
Information expected to evolve over time
Preservation Description Information
Content
Interpreted using
Representation Information
Reference Information
Provenance Information
Context Information
Fixity Information
Structure Information
adds meaning to
Semantic information
Other Representation Information
1
1
*
*
Community Sematics, Dictionaries , Thesauri
Higher level processed data Data
Raw Data
Processing software
Local parameter naming convention
Journal articles
Supports materials Websites text etc
Basic provenance e.g station name time date of observation
File Format permitting basic parameter extraction
Institutional Reports
Records of scientific output relevant to thje fiels
Complex file description
Unpublished theories and informal observations
Scientific Models and analysis tools
File naming convention, directory structure
Checksum
Detailed provenance e.g. calibrations or special modes of operation
Scaling techniques
Field Specific Bioblographies
Field Specific Bibiographies
Funding Organisation
Scientific Organisation
Data Archivist
Data Producers
Scientist

Figure 2: Typical sources of OAIS information, their dependencies and how they may be potentially classified within an OAIS system

4 CONCLUSIONS

This paper has shown how the potential of a single dataset to supply knowledge may be seen differently by dataset stakeholders. These varying requirements present curators of scientific data with a complex challenge. It is possible to create an archive with the minimum of preservation description and representation information to ensure the delivery of basic information to a future user community. The CASPAR case studies show taking this approach will result in risking the loss of a great deal of potential knowledge. Safeguarding this expanded potential in an OAIS system involves greater investment from a curating institution that needs to actively understand the full potential of its data holdings. The scope of archive will in all likelihood be limited by the logistics and cost of ingesting and actively managing this additional information, which in many cases evolves over time. This raises the challenge to CASPAR of producing software tools and strategies which reduce the cost to institutions, ameliorating this risk.

REFERENCES

[1] - CASPAR, Cultural, Artistic and Scientic knowledge for Preservation Access and Retrieval
http://www.casparpreserves.eu/

[2] - Digital Curation Centre http://www.dcc.ac.uk/

[3] - Consultative Committee for Space Data Systems. ReferenceModel for an Open Archival Information System (OAIS). CCSDS 650.0-B-1. Jan 2002 http://public.ccsds.org/publications/archive/650x0b1.pdf

[4] - A. J. McDonald, K. P. Monahan KP, D. A. Hooper, and C. T. Gaffard. VHF signal power suppression in stratiform and convective precipitation. *Ann. Geophys.*, 24:23-35, 2006.

[5] - R. Bose, R. Mann and D. Prina-Ricotti, 2006. AstroDAS: Sharing Assertions across Astronomy Catalogues through Distributed Annotation (PDF).

[6] - CLADDIER project
http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_claddier.aspx

[7] - OJIMs Project
http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/repositories_sue/ojims.aspx

[8] - NetCDF Climate and Forecast Metadata convention http://www.cfconventions.org/

[9] - Ionospheric Parameter Codes http://www.ukssdc.ac.uk/wdcc1/ionosondes/ursi_codes.html

[10] - W. R. Piggott and K. Rawer . URSI handbook of ionogram interpretation and reduction. UAG 1972
http://www.ips.gov.au/IPSHosted/INAG/uag.htm

BIOGRAPGPHY

Esther Conway

STFC Rutherford Appleton Laboratory, UK

Esther originally trained in Physics at Imperial College London after which she worked as a professional librarian specializing in scientific information, during which time she took a Masters in Information Systems and Technology at City University. Esther has also worked for several years in the area of commercial publishing as a Solutions Consultant for Thomson Learning. In her most recent position she works as a technical analyst for the DCC based at the Rutherford Appleton Laboratory.

Brian Matthews

STFC Rutherford Appleton Laboratory, UK

Brian Matthews is group leader of the Information Management Group in the STFC e-Science Centre, which conducts R&D in areas including data management, institutional repositories, semantic web and digital curation. Brian has over 20 years of experience in computer science with some 80 papers in areas including formal methods, web systems, controlled vocabulary and metadata for scientific data, and data security.

David Giaretta

STFC Rutherford Appleton Laboratory, UK

Dr David Giaretta has had extensive experience in planning, developing and running scientific archives and providing and managing a variety of services to large numbers of users. He has made fundamental contributions to the OAIS Reference Model that forms the basis of much digital preservation work far beyond repositories of scientific data, and contributes still to developing the follow on standards. He is a member of the RLG/ NARA Digital repository Certification Task Force. He has published a number of scientific papers in refereed journals and given presentations at many international conferences, scientific as well as technical. In addition he has broad experience in e-Science. In 2003 he was awarded an MBE for services to Space Science. Having been involved from the start of the successful DCC consortium, Dr Giaretta is Associate Director for Development in the DCC and now also the Project Director of CASPAR.

AKNOWLEDGEMENTS

In particular we would like to thank our colleagues based at the Rutherford Appleton Laboratory Mathew Dunckley, Stephen Rankin, Steve Crothers, Chris Davis, Rita Blake, Sam Pepler, David Hooper Simon Lambert, Brian Matthews, Mathew Wild and Ruth Bamford