D15.1 Report on the survey of technologies, policies, metadata, semantics and ontologies

Workpackage	WP15	
Task	15.2, 15.3	
Author (s)	Andrew Riddick	NERC (BGS)
	Helen Glaves	NERC (BGS)
	Yannis Marketakis	FORTH
	Yannis Tzitzikas	FORTH
Authorized by	Name Surname	Company
Reviewer	Name Surname	Company
Doc Id		
Dissemination Level	CONFIDENTIAL/PUBLIC	
Issue		
Date	29/06/2012	

Abstract:
A user consultation exercise has been conducted to identify current practices with respect to long-term data preservation within the earth science domain. This survey has been conducted as two separate but related tasks and covered a wide range of topics including semantics, metadata, ontologies, polices and technologies.
This report contains the results of the direct user consultations and also includes information from a web-based independent search activity. It will be used as input to work package (WP) 33 which is concerned with the definition of data preservation policies and harmonised semantics, ontologies and metadata as well as the development of a preliminary architecture for a European earth science long-term data preservation infrastructure.

Document Log

Date	Author	Changes	Version	Status
22/05/2012	Helen Glaves		1	Draft
31/05/2012	Andy Riddick	Integration of additional sections appendices etc	1	Draft
22/06/2012	Andy Riddick	Improvements to sections 3 and 4 following partner feedback and completion of section 5	2.0	Draft
28/06/2012	Andy RIddick	Final draft	3.0	Draft
04/07/2012	Helen Glaves	Final corrections	3.1	Final
20/07/2012	Andy Riddick/Helen Glaves	Additional final corrections	3.2	Final

Executive Summary

Earth science communities use various ontologies and schemas to describe different kinds of data and also employ a range of policies, best practice and technologies for the purposes of long-term data preservation (LTDP). A survey of current LTDP practices has been conducted as part of the SciDIP-ES WP 15 with the objective of gaining a perspective on the current 'state of the art' with respect to long-term data preservation within the earth science domain.

The surveys have been carried out as two separate but related tasks, Task 15.2 *Survey of policies and technologies* and Task 15.3 *Survey of metadata, semantics and ontologies*, with the methodology for these surveys having been developed and documented as TN15.1 *Internal Technical Note on Survey Methodology*.

Each of the SciDIP-ES surveys has been conducted in two phases. An initial phase of user consultation has been carried out using a web-based questionnaire. The objective of the on-line questionnaire was to develop an overview of current LTDP activities and practices from as wide a range of users as possible within the earth science domain. This was then followed up with an in-depth consultation with a selected group of users which included respondents identified from the first phase of user consultation and also partners from the SciDIP-ES project consortium.

The results of the user consultation have also been augmented with information derived from an independent web-based search activity. This has been conducted by project partners with the objective of identifying resources currently available which are relevant to the SciDIP-ES project. Additional relevant information has also been derived from the results of user surveys conducted by current and previous projects.

The survey activities conducted have yielded useful results on the current state of the art with respect to system architecture, data discovery and access, preservation issues, data processing, knowledge extraction and management, as well as the use of metadata, semantics and ontologies, and these are detailed further in sections 4, 5 and 6 of this document.

The results of the work package 15 surveys will be used as input for the definition of common earth science data preservation policies, harmonised semantics, ontologies and metadata, and the preliminary architecture of the European earth science LTDP infrastructure being undertaken by work package 33. An in-depth analysis of the survey results from WP15 and the user requirements developed by WP12 will be used to identify the gaps and needs in current LTDP practices within the earth science domain.

TABLE OF CONTENTS

1 Introduction

The objective of work package (WP) 15 is to carry out a user consultation exercise to assess the current levels of understanding of the concepts of long-term data preservation in the earth science domain and identify existing approaches in the areas of:

- data preservation policies and guidelines;
- metadata, semantics and ontologies;
- technologies for data discovery, access, management (e.g. preservation, processing, knowledge extraction) and visualization.

The results of the complete user survey will form part of the input to subsequent work packages and in particular WP33 which is seeking to define common data preservation policies for the earth science domain and also develop a preliminary architecture for a European earth science long-term data preservation infrastructure.

For the purposes of the SciDIP-ES project the earth science domain has been defined in its widest terms and includes those domains as shown in Figure 2.02 (from LTDP/FIRST project, courtesy of ESA)

2 Methodology

The methodology for the surveys to be conducted in work package (WP) 15 has been developed in consultation with the relevant earth science user communities and their respective organisations. It also included an evaluation of any "lessons learned" from earlier projects in the earth science domain which have carried out user surveys as part of their project methodology e.g. Geo-Seas, OneGeology-Europe etc. and also the re-use of the data captured by previous projects with a similar interest or objectives in the field of long-term data preservation e.g. CASPAR, ParseINSIGHT, LTDP/ FIRST through a user survey.

The methodology for the SciDIP-ES surveys has been documented in TN15.1 *Internal technical note on survey methodology* and has been used as the basis for conducting the surveys undertaken by task 15.2 (Survey of policies and technologies) and task 15.3 (Survey of metadata, semantics and ontologies).

The surveys have been conducted by combining two separate approaches:

1) a web-based independent search activity carried out by project partners to identify current approaches to long-term data preservation and other related activities within the earth science domain

2) a direct user consultation exercise

2.1 Independent search activity

The purpose of the independent search activity was to conduct a web-based search in order to identify relevant projects, services and resources currently in use for the purposes of long-term data preservation within the earth science domain. These searches were conducted using generic search engines, for example, Google[1], Bing[2] etc., as well as dedicated semantic web search engines such as SWSE[3], SWoogle[4], etc.

The web based search activity was conducted by all of the partners engaged in the WP15 activities. Each partner was allocated a topic for the purposes of this activity as shown in Table 2.01 and also provided with a template for reporting of the relevant resources that were identified for each topic. The individual reports from this search activity are included in Annex E and will be used to provide detailed background information for the activities undertaken in WP33.

Partner	***Topic***
ESA	DATA PRESERVATION POLICIES
STFC	DATA PRESERVATION TECHNOLOGIES
ACS	KNOWLEDGE EXTRACTION
FORTH	ONTOLOGIES
DLR	DATA DISCOVERY
INGV	DATA MANAGEMENT
ICT	SEMANTICS
ISPRA	DATA PROCESSING
CAPGEMINI	OTHER RELEVANT INITIATIVES (OTHER)
CNES	OTHER RELEVANT INITIATIVES (EUROPE)
GIM	METADATA
TOR VEGATA	DATA ACCESS

Table 2.01: Independent search activity: topics allocated to each partner

The independent search activity has also included the analysis of the results of past projects which have conducted user consultations covering similar topics of interest (e.g. GIGAS[5], EuroGEOSS[6], PARSE.insight[7]).

[1] http://www.google.com
[2] http://www.bing.com
[3] http://www.swse.org
[4] http://swoogle.umbc.edu
[5] http://www.thegigasforum.eu
[6] http://www.eurogeoss.eu/
[7] http://www.parse-insight.eu/

Resource Report Form

Resource name:	**Domain**:
URL:	
Description of the resource: *include a brief outline of the resource, its purpose and the extent of its application within the earth science domain*	
Topic:	**Associated topic**:
Name: *name of person & partner completing report*	**Partner** : *affiliation of person completing report*

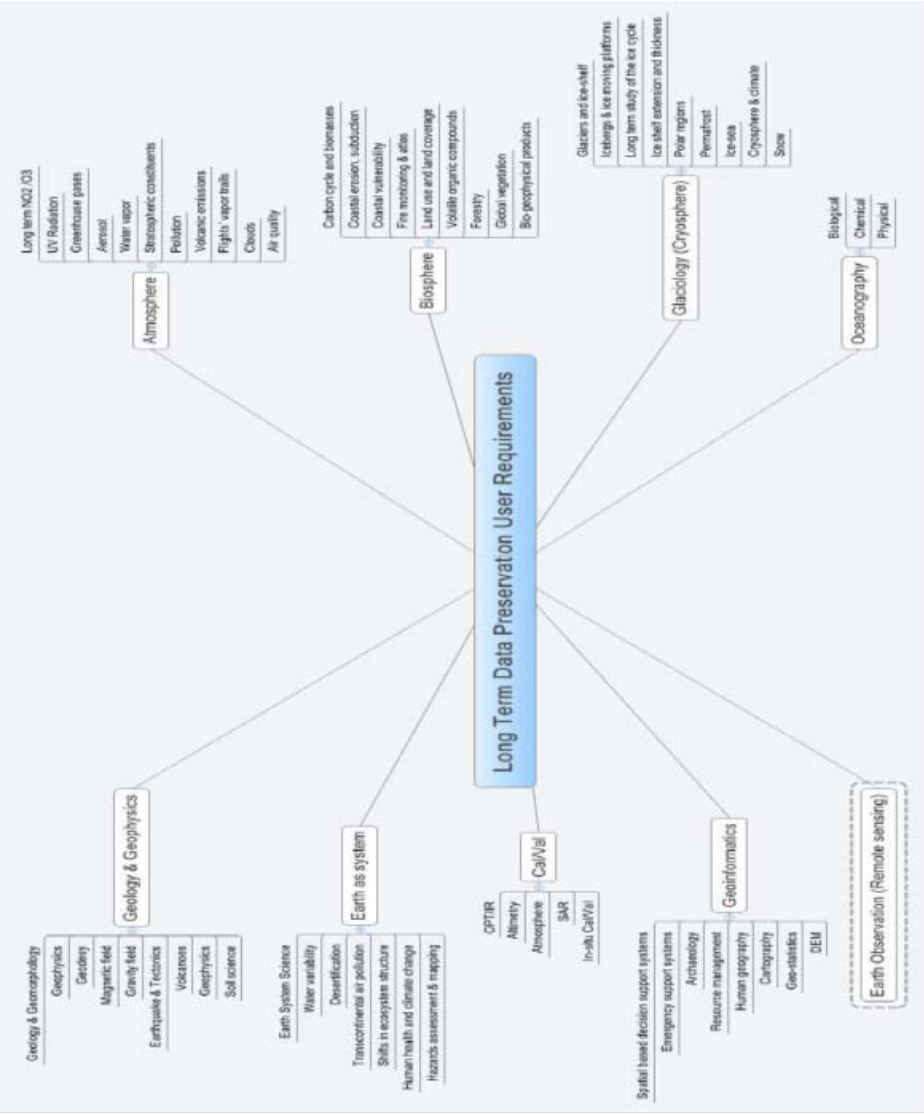


Figure2.01: Earth science domains as defined for the purposes of the SciDIP-ES project. Courtesy of the European Space Agency (ESA)

2.2 User survey

2.2.1 On-line questionnaire

The general approach for direct consultation with users in the earth science domain was to conduct the survey in two stages. The first of these took the form of an on-line questionnaire using a web based survey tool, SurveyMonkey[8] . This approach was selected because this on-line survey tool is free to use and allows user-defined questions which can include a number of structured questions and predefined answers (drop-down lists). The SurveyMonkey service also provides tools for the collection, collation and analysis of responses which can be used to evaluate the results of the surveys.

The questions in the survey were aimed at capturing information from a range of respondents including data producers, digital archive service providers and data users with differing levels of knowledge and expertise in the field of long-term data preservation. The questions for the on-line survey were compiled in such a way as to keep the complexity of the questionnaire to a minimum to ensure respondents were not discouraged from completing the entire survey whilst ensuring that the required information was captured. This was in part achieved by developing the questionnaire using a number of decision trees to ensure respondents were only presented with questions that were relevant (see Figures 2.1 & 2.2).

For both surveys the respondents were requested to provide some basic information about their affiliation, area of expertise and the function of their organisation. Figure 2.1 below shows the initial screen from the survey. All of the screens in the survey presented to the respondent following this first screen are dependent on the responses given by the user. Shown in Figure 2.2 is the screen presented to the respondent that has identified themselves as a 'digital archive service provider'.

[8] http://www.surveymonkey.com/

Figure 2.02 User Information screen from initial on-line survey

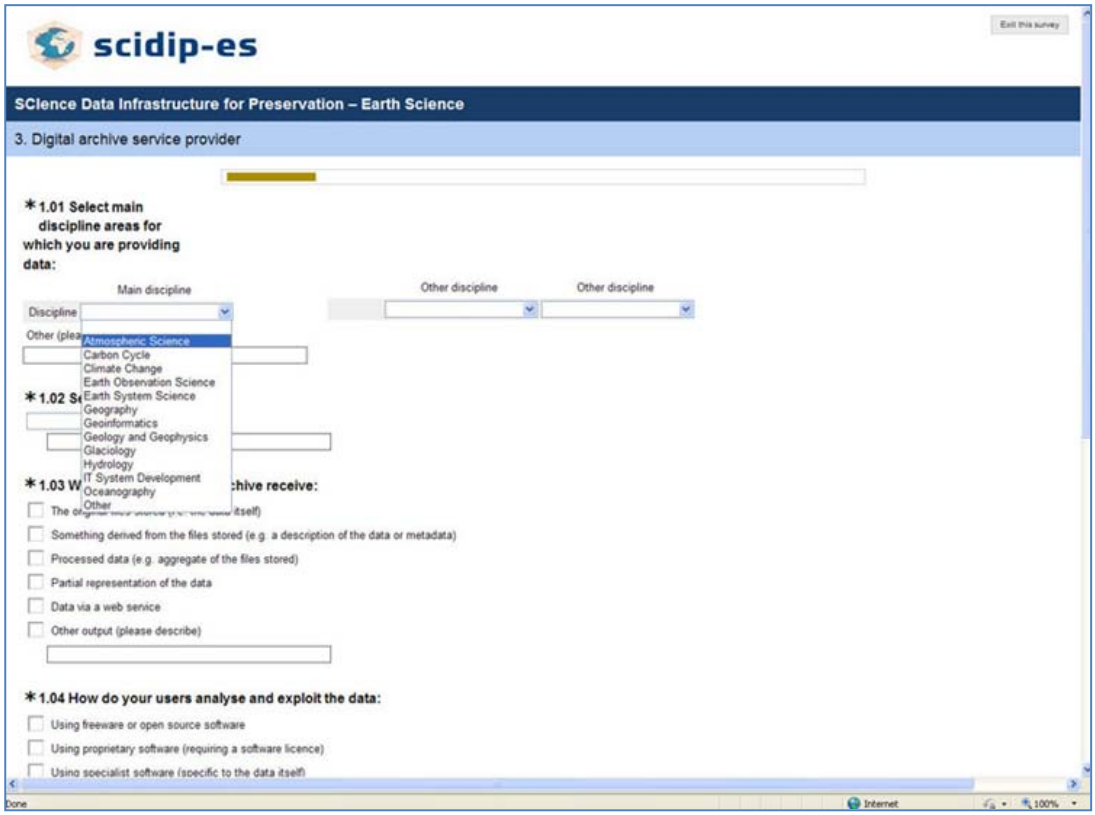


Figure 2.03 User specific in-put screen based on selection made in previous screen (see Figure 2.02)

The questions for this first phase of user survey were developed in consultation with the SciDIP-ES partners including those directly involved with the activities of WP33 to ensure that the results of

the survey would provide the necessary input to the definition of common policies, semantics, ontologies, metadata, architecture and governance being developed in WP33 as well as guiding the development of the tools and services being delivered by the SciDIP-Es project. The methodology for the WP15 user consultation was also closely aligned to that used for the definition of the use cases being developed by WP12 and in particular the capturing of the requirements of the SciDIP-ES user community for the services and tool kits.

In addition to capturing the basic information from as wide a range of respondents as possible it was intended that the responses to the first phase of user consultation would be used to identify a group of users that could take part in the second phase of more detailed consultation conducted either by face-to-face meetings or telephone/video conference interviews. The questions were therefore structured in order to identify a suitable cross section of respondents that could be approached for a more in-depth analysis of their current understanding and implementation of LTDP policies and procedures.

The SciDIP-ES user surveys are divided into two distinct tasks which cover separate areas of interest but the overall methodology for carrying out the surveys has been the same to ensure a consistent approach which provided a set of results that could by synthesised to provide a complete picture of the current 'state of the art' with respect to long-term data preservation within the earth science domain.

Dissemination of the user surveys was achieved by the project partners sending the links for the on-line surveys to their own user networks. In addition the link to the survey was also included on both the SciDIP-ES website and those of associated initiatives. There was also considerable dialogue between partners NERC and FORTH to ensure that the surveys for Task 15.2 and 15.3 were, as far as possible, not sent to the same users in an effort to maximize the response rate for both surveys.

2.2.2 Task 15.2 Survey of policies and technologies

Task 15.2 *Survey of policies and technologies* focussed on identifying existing architectures within the earth science domain including collecting information on:

1. the various aspects of interoperability
2. the tools and services used for the purposes of data discovery, access, preservation processing, knowledge extraction and management
3. the preservation policies and guidelines currently available

The survey was structured to ensure that users were presented with a relatively small number of relevant questions that were determined by the category of the user selected by the respondents at the start of the questionnaire. Each user was asked to select from one of the following categories:

- Archive service providers: individuals or organizations whose primary role is to manage and maintain an archive of data.

- Producers/creators of major data sets: individuals or organizations whose main function is to assemble and maintain major datasets for use by others.

- End users of archive data: this category includes anyone who uses archive data such as scientists and engineers who are actually using and interpreting data as well as the data managers whose role is to maintain the data retrieved from the archive.

A full list of the survey questions is available in Annex E. The T15.2 questionnaire was available on-line for a period of four weeks from 1 February 2012 until 29 February 2012.

2.2.3 Task 15.3 Survey of metadata semantics and ontologies

The survey conducted in Task 15.3 *Survey of metadata, semantics and ontologies* covered the various 'models' currently in use within the earth science domain. A user questionnaire covering these aspects was developed in consultation with the other project partners participating in WP15 the aim being to capture basic information only about the various metadata, semantics or ontologies ('models') that are currently used in the earth science domain. For each 'model' the respondent was requested to provide the following information:

- The name of the 'model'
- The URL/URI where the resource can be located
- The format (i.e. XML, RDF, etc.)
- Usage (how the 'model' is applied including the specific domain where relevant)
- Information about possible instances, or any other valuable information

The user survey of metadata, semantics and ontologies was launched on the 5 March 2012 and closed on 20 March 2012

2.2.4 In-depth user consultation

A second phase of more in-depth user consultation was undertaken by direct interviews with a selected group of users which included both the SciDIP-ES partners as well as a cross section of users identified from the responses provided in the first phase of the user survey. The purpose of this second stage of user consultation was to obtain additional more detailed information from selected users. The criteria used for the selection of the users consulted in this phase of the user survey was based on the need to ensure that no bias existed towards any one area of expertise or any particular earth science

domain. This in-depth user consultation of has been conducted by SciDIP-ES partners using a set of predefined open ended questions. (See Annex E.) which were also developed in consultation with those partners undertaking activities in WP15, WP12 and WP33. The reporting of these interviews with users was undertaken using a predefined template the purpose of which was to both guide the partner undertaking the interview regarding the information that should be captured from the interviewee(s) during the meeting as well as ensuring there was some level of consistency in the manner in which the results of the interview was reported to allow this information to be integrated and synthesized into a report.

3 Independent Search: results

The methodology used for the independent search activities is described above, and tables summarising the key information sources are provided within Annex C. The following sections summarise the key areas addressed by the independent search activities

3.1 Data preservation policies

3.1.1 Long term preservation of earth observation space data European LTDP common guidelines

These guidelines[9] provide a clear overview of the achievements made in the definition of policies for Earth Observation data preservation. This provides guidelines agreed between different SCIDIP-ES project partners and is the base to extend the harmonization on preservation policies to other Earth Sciences. The LTDP document will be enhanced according to the inputs coming from the other resources traced during the independent search activity.

Topics addressed in the European LTDP guidelines are listed as follows:

- Theme 1 – Preserved dataset composition : what to be preserved
- Theme 2 – Archive operation and organization
- Theme 3 – Archive security
- Theme 4 – Data ingestion
- Theme 5 – Archive maintenance
- Theme 6 – Data access and interoperability
- Theme 7 – Data exploitation and re-processing
- Theme 8 – Data appraisal and purge prevention

3.1.2 EUMETSAT - Policy for long term data preservation

This policy[10] which relates to EUMETSAT meteorological data provides a detailed description of the architectures and technologies used for data preservation in the EUMETSAT project.

The policy provides a clear description of a preservation program in the satellite operational context: performances and requirements (e.g. number of registered users, average daily

[9] http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_DraftV2.pdf
[10] http://earth.esa.int/gscb/ltdp/presentations/12.ltdp_approach_eumetsat.pdf

volume of data produced), operational aspects and constraints (e.g. max downtime), archive maintenance and data integrity (e.g. multi-copy, periodic migration to new media, etc.), data security, access and interoperability (e.g. HMA/OGC standards), reprocessing (e.g. API), standardisation (e.g. OAIS Ref Model, ISO, etc.).

3.1.3 USGS - National Cooperative Geologic Mapping Program Federal Advisory Committee

The above group has produced an implementation plan for their National Geological and Geophysical Data Preservation Program[11]. This policy relates mainly to preservation of physical materials (samples etc). An interesting aspect is that their general approach to preservation relies on the same pillars which are used in the digital domain:

- Identify protocols to select geological materials (data)
- Organizing the physical samples (data and metadata)
- Inventory and repository creation
- Dissemination to potential users
- User data access grant
- Ensure that the samples and data continue to be useful and reliable

3.1.4 National Geospatial Digital Archive – collection development policy

This policy[12] proposes user oriented preservation planning. The report begins by listing all the potential data users (designated communities) and then goes on to tailor the preservation program in that perspective: selection, evaluation and prioritization of the material to be preserved. Of particular interest is the section dedicated to the creation of awareness between the data potential users as part of the preservation program.

3.1.5 NOAA – policy

The following document[13] provides a general introduction to preservation principles, but is probably less useful for SciDIP-ES purposes

3.1.6 National Geospatial Digital Archive – data management policies

[11] http://datapreservation.usgs.gov/docs/2006DataPreservation.pdf
[12] http://www.ngda.org/docs/NGDA_Collection_Development_Policy.pdf
[13] http://www.nap.edu/catalog/11659.html

The National Geospatial Digital Archive has produced a document called "Libraries as Distributors of Geospatial Data: Data Management Policies as Tools for Managing Partnerships"[14]. This covers topics such as data sharing agreements, collection development policy, data management policy, end user licence agreements.

3.1.7 GEOSS – data sharing principles

The GEOSS project has produced a document entitled "Toward Implementation of the Global Earth Observation System of Systems Data Sharing Principles" [15]

The main GEOSS principles on Data Sharing are:
- There will be full and open exchange of data, metadata, and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation.
- All shared data, metadata, and products will be made available with minimum time delay and at minimum cost.
- All shared data, metadata, and products being free of charge or no more than cost of reproduction will be encouraged for research and education.

Further research on digital data policies comes from the digital library area[16] and deals with assessing the risks for loss of content due to technological changes such as the timing and likelihood of changes in technology environments and file formats (e.g. PDF, HTML, TIFF) that will affect accessibility and long-term preservation of digital content objects.

3.1.8 Interagency Working Group on Digital Data - National Science and Technology Council

This working group has produced a report entitled "Harnessing the power of digital data for science and society"[17] . The guiding principles provided in the document were derived from an analysis of the current digital scientific data landscape. These are based on the expertise of the members of the Interagency Working Group on Digital Data (IWGDD).

[14] http://muse.jhu.edu/journals/lib/summary/v055/55.2steinhart.html
[15] http://www.spacelaw.olemiss.edu/jsl/pdfs/articles/jsl-35-i-foreword.pdf
[16] http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf
[17] http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

These guiding principles are:

- science is global and thrives in the digital dimensions;
- digital scientific data are national and global assets;
- not all digital scientific data needs to be preserved and not all preserved data needs to be preserved indefinitely;
- communities of practice are an essential feature of the digital landscape;
- preservation of digital scientific data is both a government and private sector
- responsibility and benefits society as a whole;
- long-term preservation, access, and interoperability require management of the full data life cycle;
- dynamic strategies are required

The report also covers topics such as data life cycle and designated communities (organization, individuals, roles, sector and types).

3.1.9 United Nations, Educational Scientific and Cultural Organization – Guidelines for the preservation of digital heritage

The guidelines[18] contain a charter on the Preservation in the Digital Heritage prepared by UNESCO. It contains 20 articles covering the UNESCO main pillars for digital preservation: Digital Heritage as Common Heritage, Guarding against loss of heritage, Measures required, Responsibilities, plus complete guidelines on digital preservation written for a wide technical and political audience.

Particular areas of interest within this document are digital continuity (continuity of survival, access and production), also the responsibilities functions and characteristics of reliable digital preservation programmes, plus preservation planning and management.

3.1.10 Inter-university Consortium for Political and Social Research - Principles and Good Practice for Preserving Data

The guidance in the paper[19] defines the rationale for preserving data and the principles and standards of good practice as applied to data preservation, documents the development of a digital preservation policy and uses digital archive audit principles to suggest good practice for data. Section 4 focusses on formulating a data preservation policy, including

[18] http://unesdoc.unesco.org/images/0013/001300/130071e.pdf
[19] http://www.ihsn.org/home/index.php?q=focus/principles-and-good-practice-preserving-data

issues such as OAIS compliance, Administrative responsibility, organisation – roles and responsibilities, financial sustainability, technological and procedural suitability, system security, and procedural accountability.

3.1.11 Joint Information Systems Committee (JISC) – digital preservation policies

The JISC data preservation policy[20] contains a model for institutional digital preservation topic, developed around the following topics:

- Principle statement: Address how the digital preservation policy can serve the needs of the organisation and the benefits it will bring.

- Contextual Links: Highlight how this policy integrates into the organisation and how it relates to other high level strategies and policies

- Preservation Objectives: Information about the preservation objectives and how they will be supported.

- Identification of Content: Outline what the policy's overall scope is in terms of content and its relationship to collection development aims.

- Procedural Accountability: Identify high level responsibilities for the policy and provide recognition of the most important obligations faced in preserving key institutional resources

- Guidance and Implementation: Guidance and implementation clauses on how to implement the preservation policy and/or identification of where additional guidance and procedures are available in separate documentation or from staff.

- Glossary: List of definitions, if required

- Version Control: History and bibliographic details of the version. Add date of the policy, and its intended duration and review process

[20] http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf

3.1.12 Library of congress –Facing off with digital preservation policy

This research report[21] is based on a study of the preservation policies of thirteen university libraries, and identified fifteen key topics to be considered in the development of such policies:

- Content Scope
- Selection/Appraisal
- Accessioning/Ingest
- Preservation Model/Strategy
- Storage, Duplication and Backup
- Security Management
- Mandates
- Rights and Restriction Management
- Access and Use
- Financial Planning
- System Parameters
- Metadata/Documentation
- Staffing and Training
- Roles and Responsibilities
- Glossary/Terminology

3.1.13 Online Computer Library Centre (OCLC) – digital archive preservation policy[22]

This covers a number of aspects of digital preservation including:

- Assessing the risks for loss of content posed by technology variables such as commonly used proprietary file formats and software applications.
- Evaluating the digital content objects to determine what type and degree of format conversion or other preservation actions should be applied.
- Determining the appropriate metadata needed for each object type and how it is associated with the objects.
- Providing access to the content.

This policy also deals with data format risk assessment which attempts to detect the timing and likelihood of changes in technology environments and file formats (e.g. PDF, HTML, TIFF) that will affect accessibility and long-term preservation of digital content objects.

[21] http://blogs.loc.gov/digitalpreservation/2011/07/facing-off-with-digital-preservation-policy/
[22] http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf

3.1.14 UK Data Archive - preservation policy[23]

The specific aims of the preservation policy are to:

- provide authentic, reliable instances of data collections to the designated user community;
- be a "trusted repository";
- maintain the integrity and quality of the data collections;
- ensure that digital resources are managed throughout their lifecycle in the medium that is most appropriate for the task they perform;
- ensure that all data collections are protected;
- ensure that the relevant level of information security is applied to each data collection;
- instil good practice in active preservation management;
- improve the speed and efficiency with which information is preserved and retrieved;
- develop and maintain systems of low-cost storage, with appropriate location and with regular review;
- optimise the use of the archive's space for storage purposes.

3.1.15 Data preservation policies - summary and main trends

The EUMETSAT policy is based on the LTDP framework and on the Open Archival Information System (OAIS) standard, and a strong element of this policy is defining the data access policy. Outside the earth observation domain the LTDP guidelines are less applied directly, though a number of the important elements of LTDP are applied in other earth science data preservation policies.

A consistent element amongst LTDP and other earth science data preservation policies is the strategy for selection of data to be archived. The policy of the National Oceanic and Atmospheric Administration in the U.S. (NOAA) mentions using data quality as a means of assessing whether data should be included in the archive. Clearly the aims and objectives of the organisation undertaking the archiving and also of the user community are also important and this is evident in several of the above policies outside the earth sciences (a number of which are concerned with digital libraries). In particular the policy of the National Science and Technology Council (described in section 3.1.8) highlights that in some cases certain data may be cheaper to re-produce than to archive (for example in the case of some model generated data) but other data (e.g. sensor data from satellites) may represent a one time only opportunity to capture such data.

[23] http://www.data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

Another important feature of a number of the other earth science related policies (e.g. the USGS) is the need to include provision to make potential users more aware of the data available, in order to encourage exploitation of the data, and in some cases to ensure continued funding. This concept is taken further in the GEOSS data sharing guidelines where promoting the use of GEOS data for research and education is a key element, and developing metrics to monitor the effectiveness of these data sharing activities are also important.

3.2 Data Preservation Technologies

Available technologies have been grouped into data analysis tools, database archiving tools, data description languages, data dictionary/semantic technologies, tools for emulation, file readers/browsers, format identification tools, metadata tools, tools for planning the management of information assets, repository management and storage tools, software archiving tools, and also workflow management tools.

3.2.1 Data analysis tools

A wide variety of data analysis tools are available ranging from tools such as Microsoft Access and Excel within the readily available Microsoft Office software suite, to more specialised tools such as MatLab, Mathematica and GrADS for scientific data analysis across a range of disciplines. Other tools are more relevant to specific disciplines, for example ENVI for processing geospatially located images is used in the remote sensing domain. There is clearly also a wide range of programming languages available to assist data analysis, to create customised tools, including commonly used programming languages such as C/C++/C#, Java and visual basic, to the R language for statistical analysis of data. Again these programming languages can be applied across a range of disciplines both within and outside the ES domain.

Tools available for visualising data include HDFview for viewing, managing and editing HDF4 and HDF5 (hierarchical data format) binary files; the java based Integrated Data Viewer (IDV) for analysing and visualising geoscience data; and the Ocean Data View for desktop visualisation of oceanographic, atmospheric and other geo-referenced and time series data. There are also a number of tools available for manipulating data in NetCDF format, a common format used within the earth sciences.

3.2.2 Data description languages

EAST is a data description language that supplies complete and non-ambiguous information about the format of the described data, including the logical structure of the data and the physical representation of the individual data items, allowing for

variations in operating system and machine representation of numeric data. EAST was developed by the Consultative Committee for Space Data Systems (CCSDS) and has applications for earth observation as well as other types of earth science data.

DFDL is a language for describing data formats and allowing them to be converted to an XML document. Data can also be written from an instance of an information set and converted back to its native format. DFDL uses the W3C XML schema definition language (XSDL). This is an established approach that is already being used in commercial systems such as IBM's WebSphere Message Broker and Microsoft's BizTalk flat file.

3.2.3 Data dictionary/semantics technologies

The Data Entity Data Specification Language (DEDSL provides a standard method of specifying the attributes within a data dictionary and their values, using the Parameter Value Language (PVL). The recommendation for developing DEDSL comes from the Consultative Committee for Space Data Systems (CCSDS) and it seems to relate mainly to earth observation data.

The Web Ontology Language (OWL) and the Simple Knowledge Organisation System (SKOS) support the creation of ontologies, whilst the Resource Description Framework (RDF) provides a general methodology for the conceptual description or modelling of information that is implemented in web resources, using a variety of syntax formats. OWL has been endorsed by the World Wide Web consortium (W3C) and has been used across a number of disciplines including earth science. These technologies are applied in the Protégé open source ontology editor, allowing ontologies to be created, populated and visualised.

Of particular relevance to digital preservation is the Open Archives Initiative Object Re-use and exchange protocol (OAI-ORE) which defines standards for exchange and aggregation of web resources. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation of data.

3.2.4 Emulation

A number of emulators are available to emulate older software and also to emulate operating systems and to some extent hardware components. Of particular note is the KEEP emulation framework which allows older files and programs to be accessed using emulation, without needing to obtain and install older software. Dioscuri is a hardware emulator designed by the data preservation community. This tool is built in Java and can therefore be ported to a number of different platforms. Other tools are available to emulate legacy operating systems and processors, including QEMU ("Quick Emulator") which allows applications complied for one architecture to be run on another.

3.2.5 Format Identification

The format of a digital object must be known in order to interpret its information content. Format typing is therefore fundamental to the effective use, interchange and preservation of all digital content. In terms of the Open Archival Information System (OAIS) Reference Model, the format typing of a digital object is representation information about that object; that is, it provides "information that maps the data object into more meaningful concepts." However, in order to implement that mapping it is necessary to have complete representation information about the format itself: its syntactic and semantic rules for encoding information into digital form. This has lead to the development of “format registries” and tools to interrogate the format of digital objects. Format registries in particular constitute an on-going topic of research in the field of digital preservation.

Work on format identification has also been undertaken by the UK National Archives, who have developed a public file format registry called PRONOM. Other projects to develop technical registries, including the UK Digital Curation Centre's Representation Information Registry, and the Global Digital Format Registry project at Harvard University, are now in progress.

3.2.6 Metadata Tools

NASA's Earth Observing System (EOS) Clearinghouse (ECHO) is a metadata registry and order broker that allows query and access to data from a large number of repositories, primarily NASA repositories, though any repository can request to have their metadata included in the ECHO database, and stores data from a variety of science disciplines.

There are also several tools to assist the capture, cataloguing and retrieval of metadata in XML format, including the open source data management system – eXist; the metadata authoring tool, MATT; the Mercury web based system to retrieve metadata and associated datasets; and the open source metadata catalogue METACAT. The latter system is in use throughout the world to manage environmental data.

Another widely used geospatial metadata catalogue system is GeoNetwork OpenSource which is an open source geospatial data catalogue service host, metadata creation and management system, and basic web mapping platform. Another widely used system is the THREDDS Data Server (TDS) - a web server that provides metadata and data access for scientific datasets, using OPeNDAP, OGC WMS and WCS, HTTP, and other remote data access protocols.

3.2.7 Planning

Tools are available for an organisation to identify its digital assets (Data Asset Framework (DAF)), and to audit the contents of a repository (Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)). The later methodology was developed by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), and allows organisations to assess the objectives, assets and risks associated with digital preservation. The planning tool PLATO supports the migration aspect of preservation planning.

3.2.8 Repository Technologies

Fedora (Flexible Extensible Digital Object Repository Architecture) is a modular architecture built on the principle that interoperability and extensibility is best achieved by the integration of data, interfaces, and mechanisms (i.e., executable programs) as clearly defined modules, and is often used in the digital library community.

EPrints is a free and open source software package for building open access repositories that are compliant with the Open Archives Initiative Protocol for Metadata Harvesting. It shares many of the features commonly seen in Document Management systems, but is primarily used for institutional repositories and scientific journals. EPrints is a Web and command-line application based on the LAMP architecture (but is written in Perl rather than PHP). It has been successfully run under Linux, Solaris and Mac OS X . A version for Microsoft Windows was released in May 2010.

D-Space is an open source tool aimed at organisations with minimal resources. The DSpace architecture is a straightforward three-layer architecture, including storage, business, and application layers, each with a documented API to allow for future customization and enhancement. The storage layer is implemented using the file system, as managed by PostgreSQL database tables.

Of relevance to the earth science community is the National Geospatial Digital Archive (NGA) which aims to create a new national federated network for archiving geospatial imagery and data, as well as collecting and archiving important digital geospatial data and images.

3.2.9 Storage technologies

The JASMIN&CEMS cluster includes 4.6 Petabytes of usable fast access Panasas® parallel file storage (http://www.stfc.ac.uk/eScience/news+and+events/38663.aspx). The important aspects of the data storage design are the 1 Tb/s aggregate bandwidth from data to processors which supports the processing of very large data volumes, and the lower total cost of ownership than competing solutions due to less need for manual

intervention by operators to manage and expand the system. The 1133 data blades constitute the second largest configuration that Panasas® have provided to a single installation.

Hierarchical storage management (HSM) is a data storage technique which automatically moves data between high-cost and low-cost storage media. HSM systems exist because high-speed storage devices, such as hard disk drive arrays, are more expensive (per byte stored) than slower devices, such as optical discs and magnetic tape drives. While it would be ideal to have all data available on high-speed devices all the time, this is prohibitively expensive for many organizations. Instead, HSM systems store the bulk of the data on slower devices and then copies data to faster disk drives when needed. The following link: http://www.stfc.ac.uk/e-Science/services/atlas-petabyte-storage/22459.aspx provides details of an STFC based example.

3.2.10 Software and database archiving

There are a number of tools and repositories to assist software archiving, particularly for storage and version control of the source code including: SourceForge (http://sourceforge.net) which acts as a centralised location for software developers to control and manage open software development; RubyForge (http://rubyforge.org)- a site dedicated to software developed in the RUBY programming language; and Subversion (http://subversion.apache.org/) which is a software versioning and revision control system distributed under an open source license. Developers use Subversion to maintain current and historical versions of files such as source code, web pages, and documentation.

In addition to these tools there are also various software development communities which provide support for developing open source and other software. Of particular relevance to archiving software in the earth science domain is Java Forge (http://www.javaforge.com/project/11) a non-profit and free open source software development community with a hosting portal for open source projects. It hosts software development services such as project related web hosting, document management, wiki, forum, online chat and issue tracking.

Tigris.org is another open source software development community. It provides services such as web hosting, mailing lists, issue tracking, wiki, download, and revision control using Subversion or Concurrent Versions System (GNU CVS). It is hosted by CollabNet, the initiators and stewards of Subversion, and runs CollabNet Enterprise Edition. Portions of the Subversion project itself are hosted on Tigris. Tigris competes with the better-known SourceForge, although it is primarily focused on projects for collaborative software development.

In terms of web archiving OAIster (http://oaister.worldcat.org) is a freely accessible search engine for open access web resources, available from OCLC. OAIster uses the Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) to harvest records from websites. A set of tools called WARC (http://code.google.com/p/warc-tools/)

provide an open source software library, a set of command line tools, web server plug-ins and technical documentation for manipulation and management of web archive files, or WARC files.

Much geoscience data resides in relational databases and the changes in these database systems due for example, to new versions of the database software being developed present preservation challenges. The DeepArc tool developed by the National Library of France with XQuarck transforms relational database content into XML for archiving purposes. This also includes tools to map between an existing database schema and XML data models, using the XQuery technology.

One of the challenges in archiving relational databases is to archive the database structure, its content and also any built-in validation and constraint scripts The DeepArc tool described above claims to be able to cope with these requirements.

The Xing archive enquiry tool generates a web based search and browse interface for XML content, and can therefore by used to interrogate database content archived as XML.

3.2.11 Data preservation technologies – summary and main trends

A number of the tools described in section 3.2 are applicable to digital preservation in the broadest sense, for example, the MATLAB software is used to manipulate preserved data and extract additional knowledge, and the Microsoft Office suite of products produce a variety of file formats in which data can be stored). However certain technologies such as data description languages (section 3.2.2) the semantics technologies (section 3.2.3) emulation (section 3.2.4), format identification (3.2.5) and the metadata tools described in section 3.2.6 are directly applicable to digital preservation, and are likely to be of more specific relevance to WP33.

Many of the software tools which are directly applicable to digital preservation are relevant to a wide variety of science (and sometimes also non-science) disciplines. Few are specific to the earth sciences, but a number of these technologies are concerned with the basic elements of files and their representation in computer systems. Hence they should be applicable to the types of file format commonly found in earth science archives. For example the EAST and DFDL data description language would potentially provide ways of describing a wide variety of data formats. Considering the aim of increasing the level of interoperability between different earth science disciplines the data dictionary (e.g. Data entity Data specification language) and semantic languages such as OWL and SKOS (see section 3.2.4) will be important in documenting data dictionaries and establishing new ontologies to ensure this interoperability.

The availability of emulators both for software and operating systems will be important. The Dioscuri emulator was designed by the digital preservation community and being java based can be ported to a number of platforms, and therefore seems a particularly

useful tool. Important metadata tools (some of which are also referenced in the user surveys) include the open source metadata catalogue MERCAT which is widely used to manage environmental data and also the GeoNetwork metadata catalogue system which is widely used within the earth science community.

In terms of software archiving, a number of the available tools are also those commonly used by software developers during the development phase (e.g. SorceForge, and SubVersion), since these provide mechanisms for documenting and version control of the code. Open source development communities (e.g. Tigris.org) also fulfil a useful function in digital preservation in that they provide a means for users to track and be informed about changes to their software, and often methods of upgrading oopen source applications as new versions of the underlying software become available.

Considering the technologies available for storage and archive repository development, FEDORA (Flexible Extensible Digital Object Repository Architecture) has been mentioned in the survey responses, and therefore is clearly used by the earth science community to some extent. Products such as EPrints and D-Space are probably more applicable to the digital library and academic publishing worlds, but may have some relevance to SCIDIP-ES. Repository planning tools such as the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) tool, did not come up in any of the user survey responses, but given the importance of auditing repositories and establishing the criteria for including certain data (and risks in not doing so) would seem to have a potential application in the earth science domain.

3.3 Data discovery and access

The functionality provided by various web portals has been examined. The facility for data discovery is often combined with a facility for on-line access to data or to provide access to the data via off-line ordering.

The portals examined and listed in Annex C fall into two main types, those which provide a federated search across multiple archives and those which provide a dedicated search of a specific archive system. Frequently the database behind a specific portal can be accessed by federated search systems using OGC compliant standards and metadata. There is a strong indication that the facilities for federated searches across multiple archives are generally well developed.

The relevant OGC compliant standards include OGC Catalog Services (CSW) specification, Web Map Service (WMS), Interface Implementation Specification, Web Feature Service (WFS) Implementation Specification, Web Coverage Service (WCS). These standards have been widely implemented to provide access to potentially very detailed and rich sets of geospatial information.

Of particular relevance in a European context is the **INSPIRE Geoportal** (http://inspire-geoportal.ec.europa.eu/discovery/) which is the central discovery portal for the European geospatial data infrastructure (EU-GDI) providing a front end to an OGC compliant data catalog, and also the GEO portal. The GEO Portal (http://www.geoportal.org/web/guest/geo_home) is the central portal and clearinghouse for **Global Earth Observation System of Systems (GEO-GEOSS)** providing access to geospatial and earth observation (EO) data. The GEO portal allows the user to discover, browse, edit, create and save geospatial information from GEO members around the globe. This data discovery portal accesses the OGC compliant catalogues, viewing and download services of various organizations worldwide through the use of standardized OGC-compliant protocols.

Another important project concerned with data access is **GENESI-DEC** (http://www.genesi-dec.eu/). The project has established open data access services allowing European and worldwide Digital Earth Communities to seamlessly access, produce and share data, information, products and knowledge. This will create a multi-dimensional, multi-temporal, and multi-layer information facility of huge value in addressing global challenges such as biodiversity, climate change, pollution and economic development. GENESI-DEC evolves and enlarges the platform developed by the predecessor GENESI-DR project by federating to and interoperating with existing infrastructures.

GENESI-DEC involves key partners of ESFRI projects and collaborates with key participants of Digital Earth and Earth Science initiatives, including the International Society of Digital Earth and GEO-GEOSS to ensure the efficient use of already existing and planned developments.

The **INSPIRE**, **GEO-GEOS**, and **GENESI-DEC** portals are front ends to large complex systems which allow data producers to upload data and metadata to the portal and also for users to retrieve their data. Key under-pinning technologies are the web service protocols discussed above. These systems have a key European significance and therefore SCIDIP-ES will need to take account of how our tools and services at least fit with the services provided by these systems.

The **NERC Data Grid** (http://ndg.badc.rl.ac.uk/) provides a gateway to find data and explore what is known about the datasets. The data themselves remain located with the data providers, and this provides a multi-archive search for discovering data. In a similar manner the Earth System Grid (ESGF - http://www.earthsystemgrid.org/) provides a gateway to scientific collections which may be hosted at sites around the world.

In some cases, in addition to the functionality to discover and access data, tools are also made available within the data discovery/access portal to enable visualisation of data, although it appears that this integration of visualisation and analysis tools is not currently a common feature.

The **Heterogeneous Missions Accessibility (HMA)** project aims to establish harmonised access to heterogeneous Earth Observation mission data from multiple missions ground segments, including national and ESA Sentinel missions. The project partners who already have a direct contractual relationship with ESA in the framework of HMA are: ASI (Italian Space Agency), CNES (French Space Agency), CSA (Canadian Space Agency), DLR (German Space Agency), EUSC (European Union Satellite Centre).

Other web portals examined are aimed at the discovery and access of earth observation data, and in many cases it is clear that the domains which these portals support are quite diverse. For example the **Global Land Cover** facility at www.landcover.org is commonly accessed by users from a diverse range of communities including from science (geography, earth science, ecology, climatology, conservation, education) environmental policy (global warming, sustainable development, risk management) and resource management (biodiversity assessment, forestry, protected area management). In other cases e.g. the **SPOT catalogue** and maps store (http://catalog.spotimage.com) and the "**GMES Land Monitoring Portal**" (http://www.land.eu/portal/) the portal provides access to a specific dataset or range of data sets.

As would be expected, data is generally provided in formats (e.g. GIS files or images) which are appropriate to the predominant user community. There is not a great deal of evidence of users from one discipline being able to access and use relevant data from disparate domains. In fact the form based search facilities frequently provided allow searching on the basis of terms such as location, sensor, data type and time, some of which require a knowledge of earth observation data, and so may not encourage users of other disciplines to make use of it. This is clearly one area where the development of tools and services in the SCIDIP-ES project can contribute to making data more interoperable between disciplines.

3.4 Metadata and Data exchange formats

Relevant metadata standards are listed in Annex C (under metadata in the "topic" column).

3.4.1 ISO metadata standards

The independent search activities have focussed on all relevant metadata standards to geographic information as well as more discipline based standards. One of the key standards relevant to data preservation is ISO19115 which provides a structure for describing geographic data, this also contains an extension (Part 2) providing specific metadata elements for describing imagery and gridded datasets for earth observation. The survey activities in WP15 have indicated that these standards are widely applied in describing earth science data sets, and underpin a variety of data catalogue systems for data discovery. Thus the ISO19115 standard will be key in describing a wide range of earth science data. Related to this are ISO19110 which allows description of feature catalogues providing detailed

description of geographic vector data, and also ISO19119 which although not strictly a metadata standard, defines architecture patterns for service interfaces to geographic information.

3.4.2 Other metadata standards relevant to earth science data

The heterogeneous Missions Accessibility (HMA-I) project led by ESA has modelled the metadata for earth observation products as geographic features encoded in geographic mark-up language (GML). Typical attributes required in the metadata for EO products may include items such as date of acquisition of the data, location, and factors affecting the clarity of optical imagery (e.g. presence of cloud etc). The focus here is on metadata used by catalogue interfaces, rather than all metadata relevant to EO products, and therefore very relevant to discovering EO data.

In the United States the Federal Geographic Data Committee has adopted the “Content Standard for Digital Geospatial Metadata (CSDGM), Version. 2 (FGDC-STD-001-1998)” as the US Federal Metadata standard. All US Federal agencies are ordered to use this standard to document geospatial data created as of January 1995. The standard is often referred to as the 'FGDC Metadata Standard' and has been implemented beyond the federal level with State and local governments adopting the metadata standard as well. This standard was used in Europe as well but has now mostly been replaced by ISO19115.

The PREMIS (Preservation Metadata: Implementation Strategies) data dictionary scheme provides a metadata schema specifically for the data preservation activities of repositories. The aim is to the support the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context; and also represent the information most preservation repositories need to know to preserve digital materials over the long-term. An important aspect of the PREMIS scheme is that it aims to be neutral with respect to the actual technologies used for preservation, and thus may be worth further investigation as part of WP33 in SCDIP-ES.

3.4.3 Data transfer and exchange formats

There are a variety of XML based mark-up languages used for a variety of earth science, environmental and earth observation that are relevant to achieving the interoperability between data sets which will be important in SCIDIP-ES. Of particular relevance in the earth observation community is SensorML developed by the open geospatial consortium (OGC) to describe the geometric, dynamic and observational characteristics of sensors and sensor systems. The geography markup language (GML) provides a means of encoding geospatial information, and has been extended to create the GeoSciML (geoscience markup language) which can encode basic geological features using GML.

The GeoMS (Generic Earth Observation Metadata Standard) developed jointly by ESA and NASA facilitates the exchange of earth observation validation data among investigators and missions. In an archiving context the Encoded Archival Description (EAD) provides a standard for describing collections held by archives, and enables the encoding of archival finding aids into searchable records that are platform independent. Whilst EAD is not specifically configured for earth science data, it may be a useful tool in increasing interoperability between earth science data sets, particularly given it is platform independent. In a similar context the XML formatted data unit (XFDU) provides for packaging data and metadata including software into a single package to facilitate information transfer and archiving, and thus may have some application in packing earth science data.

3.4.4 Semantics and Ontologies

Relevant semantic technologies are listed in Annex C (see "semantics" topic), and some of they key resources are described below.

The semantic web for Earth and environmental terminology (SWEET - http://sweet.jpl.nasa.gov) is an investigation in improving discovery and use of Earth Science data, through software understanding of the semantics of web resources. Semantic understanding is enabled through the use of ontologies, or formal representations of technical concepts and their interrelations in a form that supports domain knowledge. The ontologies within the SWEET system are implemented using the OWL ontology language. Currently 6000 concepts are included within 200 ontologies. The ultimate vision of the semantic web consists of web pages with XML namespace tags around terms, enabling search tools to ascertain their meanings by following the link to the defining ontologies.

Another relevant project concerned with ontologies is the ESA OTEG project (open Access ontology/Terminology for the GMES space component). This is a multi-domain thesaurus and vocabulary with a web interface (http://gmesdata.esa.int/OTE/navigateInfoDomain). The system is able to display a graphical view of the inter-relationships between different terms. It is also possible to link through to details of individual datasets starting from the vocabulary terms.

A further relevant ontology is "GeoNames" (http://www.geonames.org/ontology/documentation.html), this makes it possible to add geospatial semantic information to the World Wide Web. Overall 6.2 million geonames toponyms now have a unique URL with a corresponding RDF web service. Other services describe the relations between toponyms. The Features in the GeoNames Semantic Web are interlinked with each other, and the system allows users to search for a particular concept and be linked to documents about that concept.

The General Multilingual Environmental Thesaurus (GEMET) is a multi-lingual thesaurus (http://www.eionet.europa.eu/gemet) managed by the European Environmental Agency (EEA) that contains a wide variety of environmental terms, including terms relevant to earth science with their translations in more than 20 languages. It has been developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA), Copenhagen. This project is probably less relevant for SCIDIP-ES than for example the SWEET ontology system or the ESA OTEG project described above.

3.5 Data Management

The process of data management is concerned with ensuring that datasets are easily accessible, stored in a secure environment and contain data that is up to date.

The Rasdaman project seeks to extend standard relational database systems with the ability to store and retrieve multi-dimensional raster data (arrays) of unlimited size through an SQL style query language. Key features are fast access to multi-terabyte objects, and also scalability from laptop to high performance computer systems. Thus the system aims to deal with "big data" and also supports open standards such as WMS, WCS, WCS-T, and WCPS. The software tools produced will be open source and therefore very relevant to dealing with large amounts of geospatial data in SCIDIP-ES.

EarthServer is a related ongoing EU FP7 project (http://www.earthserver.eu/) aimed at open access and ad-hoc analytics on Earth Science (ES) data, based on the OGC geo service standards Web Coverage Service (WCS) and Web Coverage Processing Service (WCPS). EarthServer will use some of the technologies developed by Rasdaman. The outcome will be open-source server and client packages, field tested in main areas of Earth science: atmosphere (climate modelling), hydrosphere (oceanography), lithosphere (geology), and cryosphere (snow & ice mapping). This service will allow interoperable,standards-based, integrated large-scale retrieval on the coverage data themselves, aligned and combined with metadata retrieval.

Rasdaman and EarthServer then are projects developing the technologies for rapid querying of large amounts of spatial data ("big data") in the earth sciences. Given that EarthSErver is running concurrently with SCIDIP-ES there would be oportunities for further collaboration possibly as part of WP33.

A number of systems provide front ends to large relational or other database systems. Within the SCIDIP-ES consortium the MOIST ("Multidisciplinary Oceanic Information System") developed by INGV is aimed at hosting multidisciplinary data and metadata from sea floor observatories. Data are collected at the end of a mission or in real time by the Data Centre at INGV in Rome. MOIST's internal architecture is a single RDBMS that stores both data and metadata. Another example of a data management system using relational database with a web front end is the ESWUA electronic space weather upper atmosphere website also developed by INGV. This system contain petabytes of data and provides great flexibility in data retrieval.

There are a variety of smaller web/database systems listed in Annex C. One of the issues in retrieving data across multiple database systems is of course the likelihood of differences in the database structure, and the existence of common metadata standards.

With recent developments in cyber- infrastructure there is an increasing tendency for the creation of large systems for data management, which may also provide facilities

for discovery, access and processing of data within a coherent system, such as for example the EarthCube and GEON initiatives currently being developed in the United States, which seeks to integrate data and information for knowledge management across the geosciences within a cyber-infrastructure network.

3.6 Data Processing

Various web resources relevant to data processing are summarised in Annex C. These include technologies and frameworks for processing data as well as projects seeking to develop better facilities for data processing, and facilities linked to the INSPIRE initiative.

3.6.1 Technologies and frameworks for processing data

●These include the Web Processing Service (WPS) interface standard which provides rules for standardising inputs and outputs (requests and responses for geospatial processing services. Through WPS a generic user gains access to geospatial data processing tools provided by third parties. WPS can be seen as a way to perform standardized geospatial computations in a distributed environment. In the context of LTDP it can be used as a tool to preserve data processing algorithms and procedures in the geospatial domain as long as adequate data preservation policies are implemented on the infrastructure providing the service itself.

● The OpenGIS® Web Coverage Processing Service (WCPS) Interface Standard (http://www.opengeospatial.org/standards/wcps) defines a protocol-independent language for the extraction, processing, and analysis of multi-dimensional gridded coverages representing sensor, image, or statistics data. Services implementing this language provide access to original or derived sets of geospatial coverage information, in forms that are useful for client-side rendering, input into scientific models, and other client applications.

● Open virtualisation format (OVF) represents a standard vendor independent representation of virtual machines which, in turn, are a common component of data preservation strategies. A virtual machine containing all the processing chain components of a given dataset can be used to reproduce and analyse the procedures and algorithms used in data processing.

● Earth System Modelling Framework (ESMF) defines an architecture for composing complex, coupled modelling systems and includes data structures and utilities for developing individual models. The ESMF framework is emerging as a standard among the modellers in the earth science domain. The standards and software tools defined by ESMF might be useful to support LTDP of model related data. Moreover, its components can be used as standardized data processing tools. ESMF is supported mainly by US organizations, universities and research centres.

● Open Modelling interface (OpenMI) was developed within the EU funded projects HarmonIT and OpenMI-Life. OpenMI evolved to become a generic solution to build

software components that can be applied to linking any combination of models, databases and analytical/visualisation tools. As an emerging standard in the domain of earth science will play a major role in preservation of data processing capabilities. Open MI has a similar role to the Earth Modelling Framework (ESMF) described above, although a key feature is that it is able to pass variables between models at run-time. A framework of open source components are used to “wrap” components of models and to this extent OPenMI may represent a useful means of preserving linked environmental models.

3.6.2 Initiatives related to INSPIRE

- The INSPIRE Directive aims at creating an infrastructure for geographical information interoperability in Europe. In this context data holders should publish their geographic datasets through a range of Network Services. INSPIRE Transformation services provide a means to transform a given dataset through the invoking of a service implementing a standardized procedure on a remote machine. Typical examples of transformation services are the schema transformation which transforms the structure of the input dataset and the Coordinate Reference System (CRS) transformation which can be used to bring together datasets based on different CRS.
- The “Invoke Spatial Data Service” service allows definition of both the data inputs and data outputs expected by the spatial service and define a workflow or service chain combining multiple services. It also allows the definition of a web service interface managing and accessing (executing) workflows or service chains.
- In the context of LTDP the INSPIRE Transformation and Invoke Services can be used as a tool to provide distributed data processing capabilities which can be preserved independently of the datasets.

3.6.3 Other projects concerned with processing data

- The Kepler Project supports the use of the free and open source scientific workflow applications. Kepler has been designed to help scientists, analysts, and computer programmers create, execute, and share models and analyses across a broad range of scientific and engineering disciplines. Kepler can operate on data stored in a variety of formats both locally and over the internet, and is an effective environment for integrating disparate software components, such as merging "R" scripts with compiled "C" code, or facilitating remote, distributed execution of models. Using Kepler's graphical user interface, users simply select and then connect pertinent analytical components and data sources to create a "scientific workflow"—an executable representation of the steps required to generate results. The Kepler software helps users share and reuse data, workflows, and components developed by the scientific community to address common needs. Kepler’s main developers are US Universities (UC Davis, UC Santa Barbara, and UC San Diego). As an emerging standard data analysis and processing tool Kepler is likely to play a role in LTDP.

- KEEP (Keeping Emulation Environments Portable) is a project that will develop emulation services (KEEP Emulation Services) to enable accurate rendering of both static

and dynamic digital objects: text, sound, and image files; multimedia documents, websites, databases, videogames etc. The overall aim of the project is to facilitate universal access to our cultural heritage by developing flexible tools for accessing, manipulating and storing a wide range of digital objects using emulation tools either to reproduce the original environment in which they were created or to enable those objects to be migrated accurately to another environment.

In addition to the development of a KEEP Emulation Framework, within which 3rd party emulators are hosted, the project is also supporting the development of a virtual machine which will permit other environments to operate independently of the actual software and hardware environments.

3.7 Knowledge Extraction

Several systems which allow extraction of additional knowledge from archived material (as oposed to simply retriving data from a database) are described below. These systems provide additional facilities and services to interact with the data – commonly image files or GIS data.

The Knowledge-Centred Earth Observation (KEO) system (http://keo-karisma.esrin.esa.int/keo-home/Welcome.html) developed by ESA provides facilities for automatic extraction of information from earth observation images including spectral signature, textural information, geometric parameters and discrete cosine transform. The Component Based Processing Environment (CPE) within this system allows the user to graphically chain together processing components and to work with EO data from different sources. The CPE provides calibration and classification of single images, as well as object/feature detection from single images, and also signal processing.

Another European system providing facilities for knowledge extraction is the Service Support Environment (SSE - http://services.eoportal.org/). The SSE service directory offers access to a continuously expanding set of basic and more complex earth observation and GIS services. The main EO contributors such as space agencies, data processing centres, data providers, educational establishments, private companies and research centres have chosen to actively participate in the SSE initiative enabling the SSE portal to give access to a large variety of services. The objective of the Heterogeneous Mission Accessibility activity is to define the interoperability concept across the ground segments of the European, Canadian and EUMETSAT missions which will contribute to the initial phase of GMES. The SSE portal is offering access to the prototypes and services for demonstration resulting from the HMA activities.

The CORINE programme (http://www.eea.europa.eu/publications/COR0-landcover) provides a geographic information system for collating standardising and exchanging data on the environment. The land cover project is part of the CORINE programme and is intended to provide consistent localized geographical information on the land cover of

the 12 Member States of the European Community. The project addresses the problem that in all the countries of the European Community, the information on land cover available at national level is heterogeneous, fragmented and difficult to obtain. The CORINE system is able to provide extraction of data at EU Community, national and regional level, and thus provide information at a variety of scales as well as deal with differences in nomenclature between different countries etc.

3.8 Other Relevant Initiatives

Table 3.01. Other relevant initiatives in Europe

Name	Description	Domain
SCAPE (Scalable Preservation Environments)	The main goal is to assess the large scale applicability of the **SCAPE** Preservation Platform and the preservation components developed within the project. Using these software components, it creates test environments for the different application scenarios and complex large scale preservation workflows. As part of the test bed evaluation methodology, the automated planning tool Plato will be used to evaluate the strengths and weaknesses of the action components in several scenarios. (http://www.scape-project.eu/about/project)	Domain independent

Name	Description	Domain
EUDAT	A European e-Infrastructure ecosystem is currently taking shape, with communication networks, distributed grids and HPC facilities providing European researchers from all fields with state-of-the-art instruments and services that support the deployment of new research facilities on a pan-European level. However, the accelerated proliferation of data – newly available from powerful new scientific instruments, simulations and digitization of library resources –, has created a new impetus for increasing efforts and investments in order to tackle the specific challenges of data management, and to ensure a coherent approach to research data access and preservation. EUDAT aims to address these challenges and exploit the opportunities using its vision of a Collaborative Data Infrastructure.(http://www.eudat.eu/)	Scientific data – not specific to earth science
ENSURE	The ENSURE initiative is driven by the need to guarantee long term usability of the immense amount of data produced or controlled by organisations with commercial interests. Guided by real world use cases in health care, financial data, and clinical trials ENSURE seeks to extend the state of the art in digital preservation with particular focus on cost and value of preservation, preservation life cycle issues, and identity management over decadal time intervals http://ensure-fp7-plone.fe.up.pt/site	Mainly health finance and other non-science areas, but work on cost, value and preservation lifecycle management may be relevant to SciDIP-ES
OpenAIRE (Open Access Infrastructure for Research in Europe)	OpenAIRE, is a three-year project, that will establish the infrastructure for researchers to support them in complying with the EC OA pilot and the ERC Guidelines on Open Access. The project will establish and operate an electronic infrastructure for depositing and handling peer reviewed publications produced as part of FP7 projects (http://www.openaire.eu/)	Domain Independent

It is clear from the above table that there are a considerable number of relevant initiatives in Europe particularly concerned with knowledge integration and preservation in the earth observation, atmospheric science, and oceanography domains.

Some of the key initiatives which extend beyond Europe are highlighted in Table 3.02. below (a full list of all initiatives gathered is provided in ANNEX C):

Table 3.02. Relevant initiatives outside Europe

Name	Description	Domain
CNPDI (Canadian Network for Polar Data Infrastructure)	The purposes of the network include building a secure network housing the infrastructure needed to provide long-term preservation of research data. The long-term preservation of digital information requires an archival information system that constantly verifies data integrity, and that upgrades to new standards over time. The network also provides access to diverse data sets generated by Arctic and Antarctic Researchers via a metadata system following international standards (http://cnpdi.ca)	Glaciology
DART (Digital Archiving and Retrieval Tool)	The project is working to develop techniques to represent and package engineering data for long-term digital storage. Current activities and accomplishments of the project include the development of an initial prototype based on the Open Archive Information System (OAIS) Information Consumer interface. The prototype allows users to perform content-based search of their archived CAD objects. (http://gicl.cs.drexel.edu/wiki/Digital_Archiving_and_Retrieval_Tool)	Domain independent
FEDORA (Flexible Extensible Digital Object Repository Architecture)	Fedora (Flexible Extensible Digital Object Repository Architecture) was originally developed by researchers at Cornell University as an architecture for storing, managing, and accessing digital content in the form of digital objects. Fedora defines a set of abstractions for expressing digital objects, asserting relationships among digital objects, and linking "behaviors" (i.e., services) to digital objects (http://www.fedora-commons.org/)	

The above list highlights a number of cyberinfrastructure initiatives which are seeking to link data from different disciplines using emerging technologies such as cloud computing and advanced visualisation capabilities. The emerging capability to link data and models to produce useful interpretations will also present data preservation challenges.

3.9 Summary of independent search activities

A number of key under-pinning technologies which support digital preservation (directly or indirectly) are evident, particularly various metadata standards, and also web services protocols, and to some extend the use of relational database formats both proprietary and open source. These standards are clearly well adopted in the earth science community and will be of benefit in interfacing SciDIP-ES tools and services with existing archive systems. A wide variety of technologies pertinent to digital preservation were identified (section 3.2) many of these are not specifically intended for earth science data, but nevertheless may be relevant. A number of the archiving technologies have arisen from the digital library world, and these would merit further investigation, particularly tools for data description and emulation.

Clearly a number of frameworks and protocols have been developed using these underlying building blocks (e.g. the OpenMI standard for integrated modelling, and various metadata schemas for earth observation data, and an understanding of how these frameworks work will be of advantage in SciDIP-ES.

One of the trends apparent in the development of technologies to support digital preservation is that there is an increasing tendency for the development of large cyber-infrastructure initiatives which include provision for archiving in large relational databases or other means, and which frequently adopt recognised standards for example for metadata or the use of web services to access the data. Particular examples are the GENESI-DEC, Euro-GEOSS systems in Europe, and the EarthCube initiative in the United States. Often smaller but nevertheless important systems (e.g. the MOIST system) also take advantage of emerging standards e.g. for web services, and this enhances the potential to interface with these systems.

4 User Surveys: Results, Analysis, and Conclusions

4.1 Introduction

The user consultation was carried out in two phases as described in section 2 above. The first phase was conducted as an on-line survey with the objective of collecting high level information about the way various infrastructures, technologies, standards and services are used for data preservation within the earth science domain to get an overview of current trends and to also identify individuals who could potentially contribute to a subsequent, more detailed phase of user consultation.

The in-depth consultation was conducted with a group of users who were selected from both within the SciDIP-ES consortium and also from the respondents to the on-line user survey. This consultation process was carried out either by one-to one interviews with individual users or through small working groups. The objective of the second phase of the user survey was to build on the information captured by the web-based survey through direct consultation with users with expertise in the relevant areas.

The results of the various elements of the user survey are described below and include a number of high level conclusions and recommendations. In interpreting the survey results we have sought to address a number of key areas relevant to data preservation including: system architecture and infrastructure (section 4.3), discovery of and access to archived data (section 4.4), preservation issues (section 4.5), data processing, knowledge extraction and management (section 4.6), and finally a section on metadata semantics and ontologies (section 4.7). An in-depth analysis of the results of the survey will be conducted by WP33 as part of the development of the common policies and architectures for long-term data preservation.

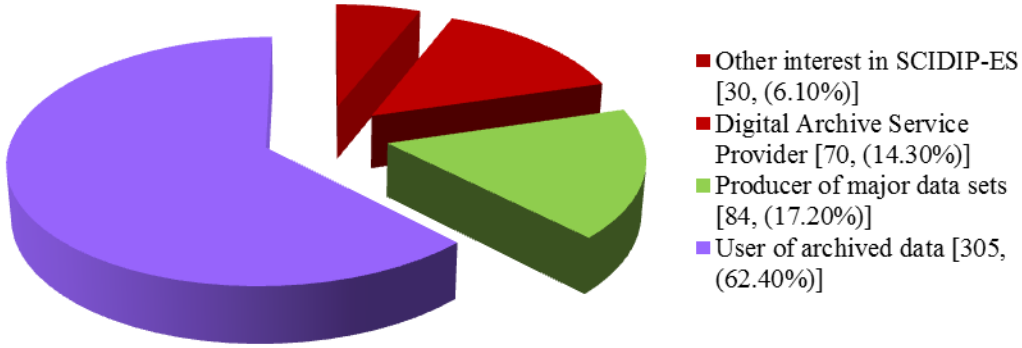
4.2 Overview of the data collected

4.2.1 Task 15.2 Survey of policies and technologies

A total of 489 respondents completed the survey of policies and technologies, providing a good basis for statistical analysis. Figure 4.1 below shows the overall number of responses in the three main categories of users (archive service providers, producers of major datasets, and end users of archive data).

Of those that responded the users of the data (“Users of archived data”) was the largest category of respondents with 62% of the responses being from that group. Of the rest, the digital archive service providers and the data producers were equally represented. A fourth category of respondent was also identified which are those that have another interest in SciDIP-ES. This category was provided for those people who do not wish to answer the technical questions but wanted to indicate an interest in the project and to be added to the project mailing list. The response rate from the three main categories of user has provided sufficient information for a number of statistically valid conclusions to be drawn from the user survey.

Figure 4.1 Relative proportions of the main user groups identified



The geographical distribution of respondents indicates that the majority were based in Europe which is as expected since the project consortium is made up entirely of European organisations (See Figure 4.22). The results also show that all of the major continents are represented with a significant proportion of the respondents being dispersed across a wide range of countries around the world.

Figure 4.2 Distribution of Respondents by Continent

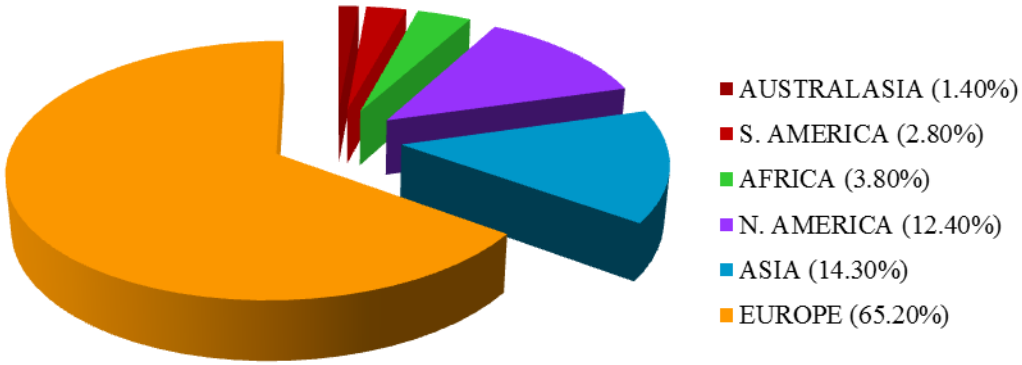
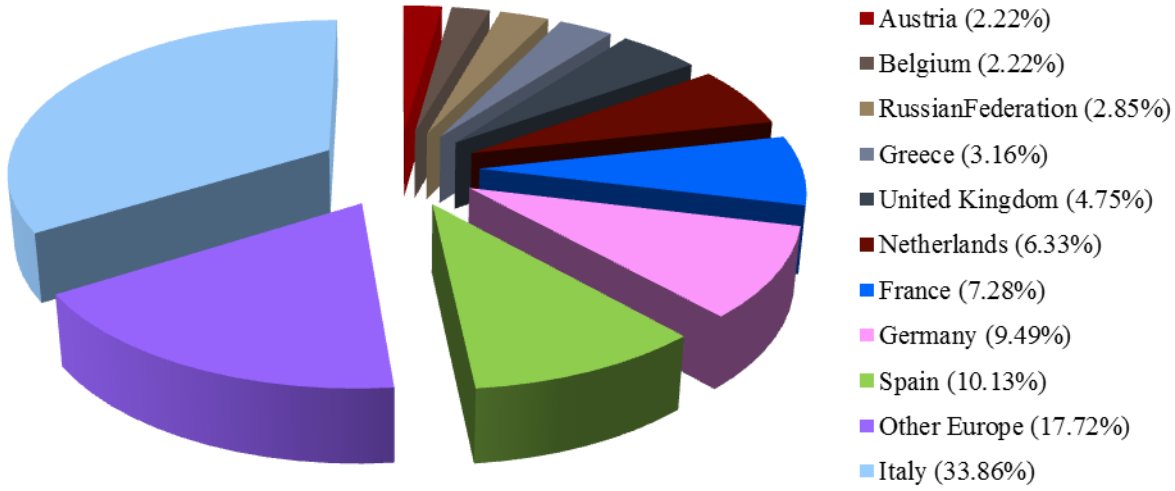


Figure 4.03. Distribution of on-line survey respondents in Europe



The breakdown by country for those respondents based in Europe indicates that users were dispersed across a range of countries. There were respondents from all of the countries represented in the SciDIP-ES consortium and the large proportion of respondents based in Italy is likely to be due to the fact that 7 of the 17 partners in the consortium are based in Italy. The level of respondents from each country may also be representative of the level of knowledge and expertise of long-term data preservation and this is a trend that should be explored further in WP33 since there may a need for a greater level of knowledge exchange with those European countries that are poorly represented in the results of the user survey.

It should also be noted that the geographical spread of respondents will also to some extent be skewed by the mechanisms used for the dissemination of the user surveys which was largely achieved by the project partners forwarding the survey to their own user networks. It has previously been shown that a user receiving a request to complete a survey is more likely to participate in this type of user consultation if it has been received from someone that is known to them. It is therefore likely that there will be more responses from users in those countries represented by partners in the SciDIP-ES consortium.

4.2.2 Task 15.3 Survey of metadata, semantics and ontologies

The online survey of metadata, semantics and ontologies had 62 responses from 57 distinct users. A small number of the respondents participated in the questionnaire more than once in order to provide additional examples of 'models' due to the constraints of the survey which allowed only 5 examples of models to be included in each response. The number of respondents is relatively small due to the fact that there are a limited number of experts in the field of metadata, semantics and ontologies within the earth science domain

Figure 4.04. Respondents to metadata semantics and ontologies survey by country

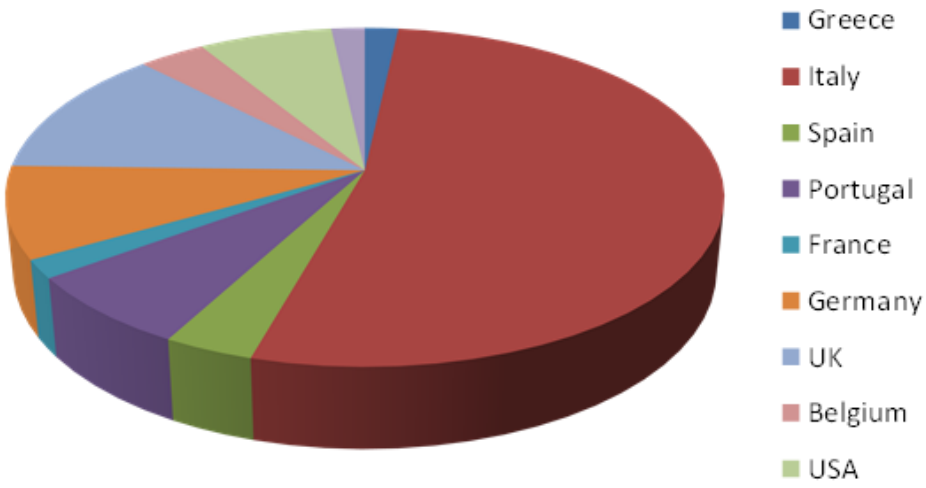


Figure 4.05 (below) shows the different roles of the respondents. It is important to note that almost half of the responses have been provided by non-SCIDIP-ES partners and furthermore many of the answers were derived from users of the various data models. This is important because the one of the purposes of the surveys was to recognize the state of the art in semantics, metadata and ontologies that are currently in use in the earth sciences, which means that the survey should not be restricted to the SCIDIP-ES consortium. This is important because the feedback from these users allows identification of the real weaknesses and gaps in the various systems. The following tables give an overview of the organizations that responded the questionnaire. The Table 4.01 below contains the organizations that responded and are inside SCIDIP-ES project while the second Table 4.02 contains those outside of the initiative.

Table 4.01 SCIDIP-ES partners

SCIDIP-ES Partners	
Organization	Short Name
European Space Agency	ESA
Science and Technology Facilities Council	STFC
Advanced Computer Systems	ACS
Foundation for Research and Technology Hellas	FORTH
Deutsches Zentrum Fuer Luft	DLR
Natural Environment Research Council	NERC

Instituto Nationale di Geofysica e Vulcanologia	INGV
Forschungsinstitut Für Telekomminikation e.V.	FTK
Instituto Superiore per la Protezione e la Ricerca Ambientale	ISPRA
Jacobs University Bremen	JUB
Centre National d'Etudes Spatiales	CNES
Geographic Information Management	GIM
Universita degli Studi di Rome – Tor Vergata	UTV

Table 4.02 Non-SCIDIP-ES partners

Non-SCIDIP-ES partners	
Organization	**Country**
University of Pavia	Italy
Consiglio Nazionale delle Ricerche (CNR)	Italy
Roffer's Ocean Fishing Forecasting Service Inc	USA
Institut De Ciénces De La Terra Jaume Amlera	Spain
Institute and Geophysical Observatory of Antananarivo (IOGA)	Madagascar
Rensellaer Polytethnic Institute	USA
Scripps Institution of Oceanography	USA
Instituto Espaniol de Oceanografia	Spain
European Maritime Safety Agency	Portugal
National and Kapodistrian University of Athens	Greece
Fern Universitat in Hagen	Germany

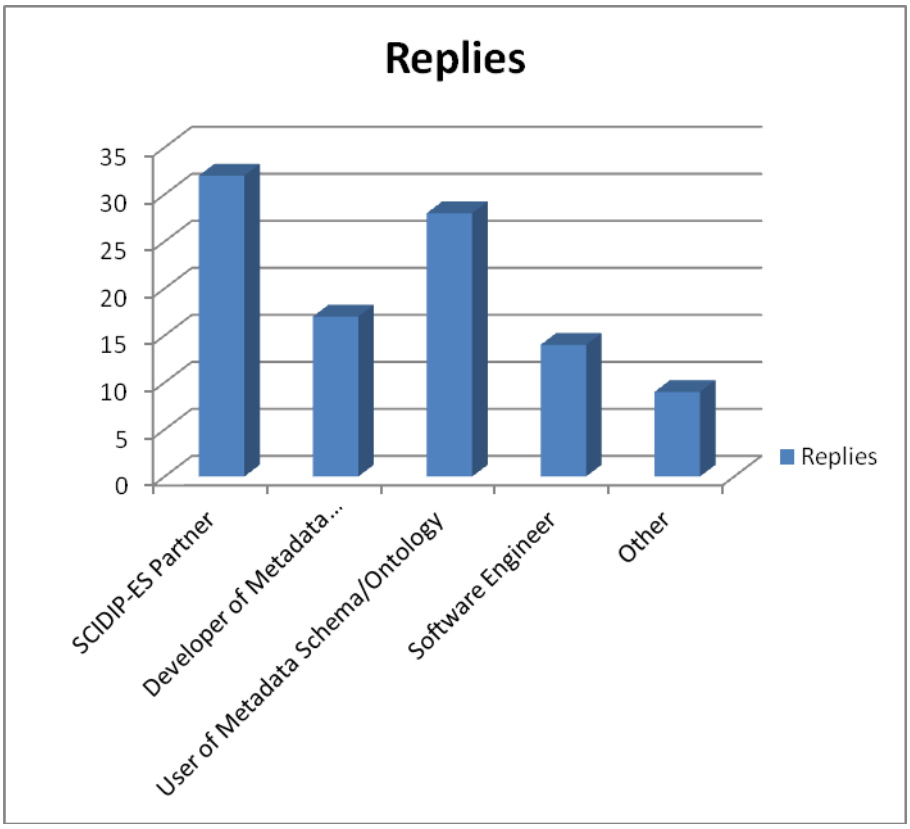


Figure 4.05. Responders by organisation activity

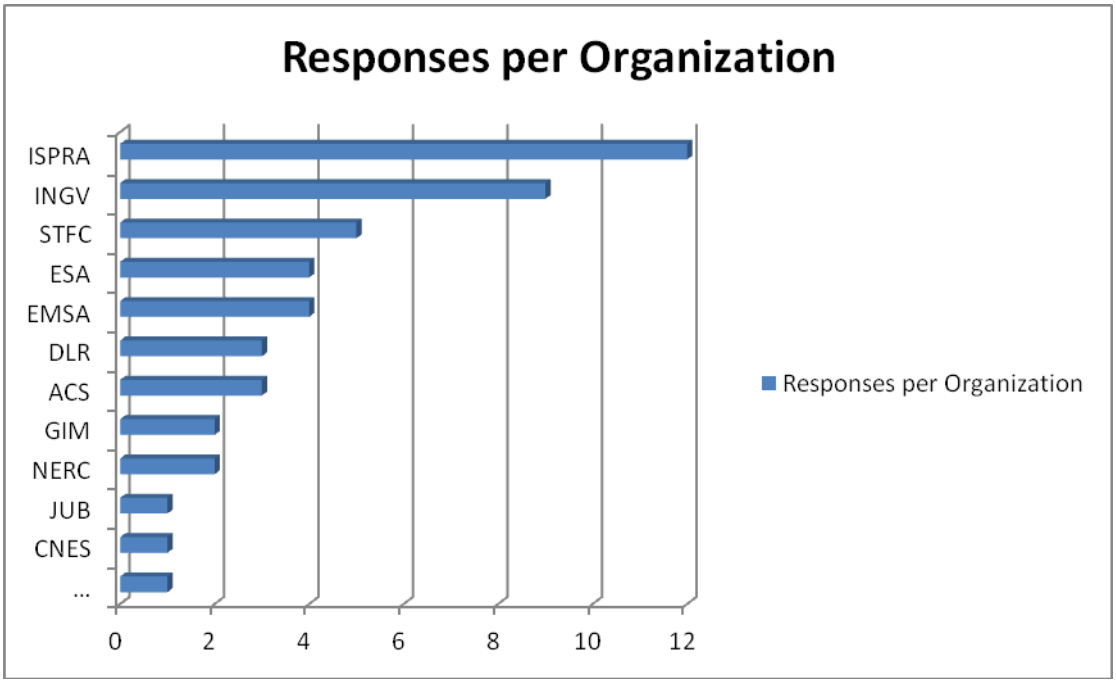


Figure 4.06. Number of respondents per organisation

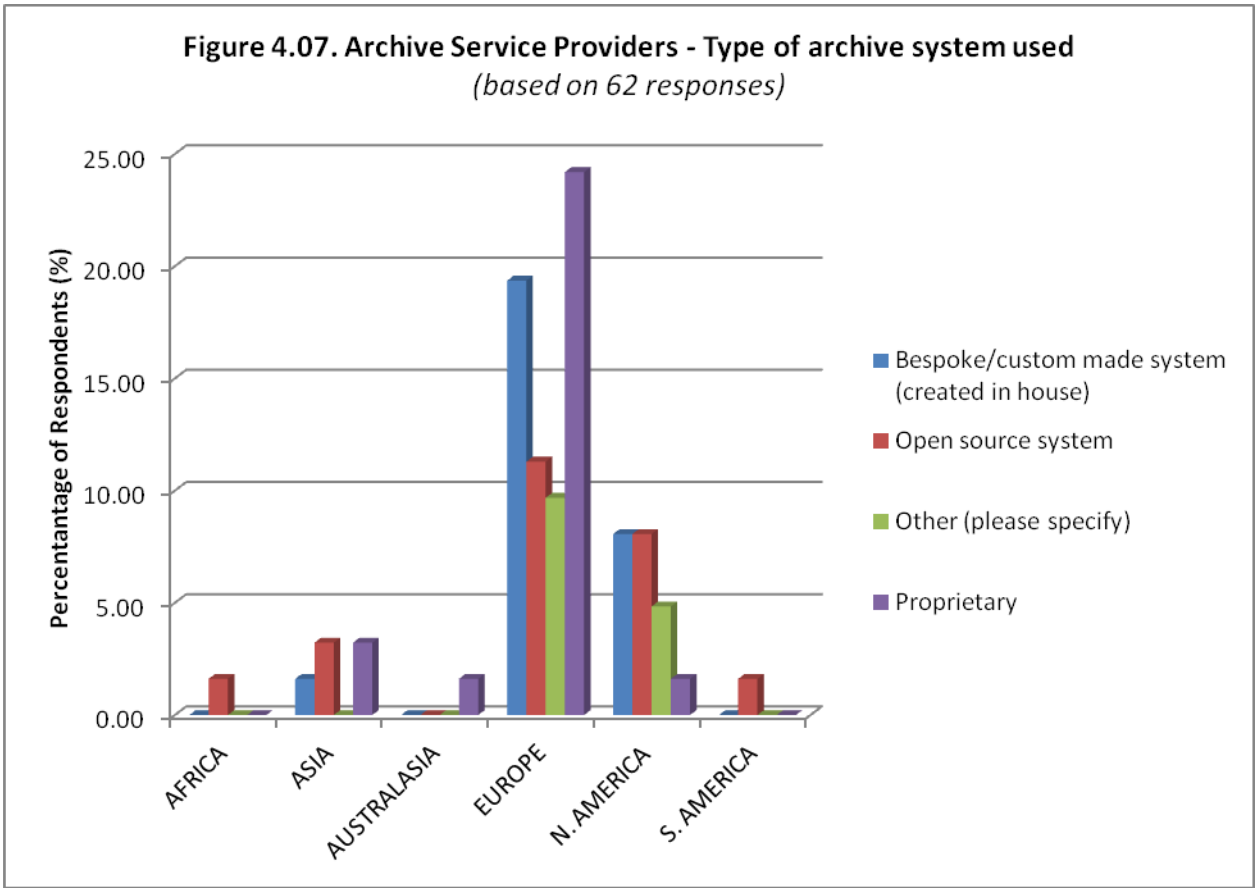
4.3 System architecture and infrastructure

4.3.1 Results from the on-line survey

4.3.1.1 Archive Service Providers

Respondents were asked to indicate whether the archive system they used was a proprietary system, based on open source components or a custom developed system for their specific needs. The results indicate a significant variation in the type of system used, even for example within Europe or North America (see Figure 4.07). The use of proprietary (commercially developed) systems is particularly important in Europe with nearly 25% of respondents in total falling into this category. However nearly 30% of respondents indicated a European origin and the use of open source or custom systems Where users selected the "Other" category, there was generally little information about what systems they were using, though some users described using a combination of custom developed, open source and proprietary systems.

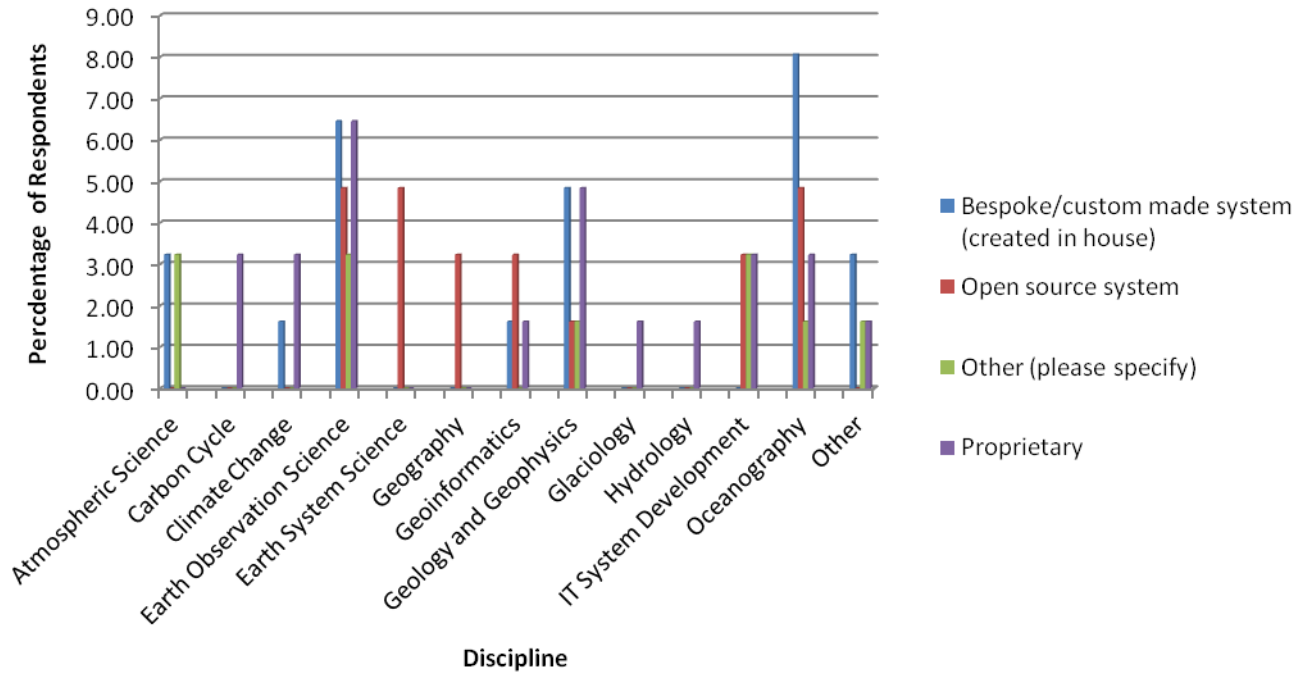
Figure 4.07. Archive Service Providers - Type of archive system used
(based on 62 responses)



Comparing this with the breakdown of the type of storage system used between disciplines (Figure 4.08) indicates a mixture of types of storage system within the three disciplines having most respondents (Earth Observation Science, Geology and Geophysics, and Oceanography), with custom developed, proprietary, and open source systems present. The seventy or so

archive service provider respondents are split between a number of disciplines, and therefore the number from each discipline is small. Due to the small numbers in each discipline it is difficult to make definitive conclusions about the variation of archive storage system with discipline, but tentatively Earth Observation Science and geophysics are dominated by custom and proprietary systems, with some use of open source whilst oceanography is dominated by custom and open source systems.

Figure 4.08. Archive Service Providers – variation of archive system type between disciplines *(based on 62 responses)*

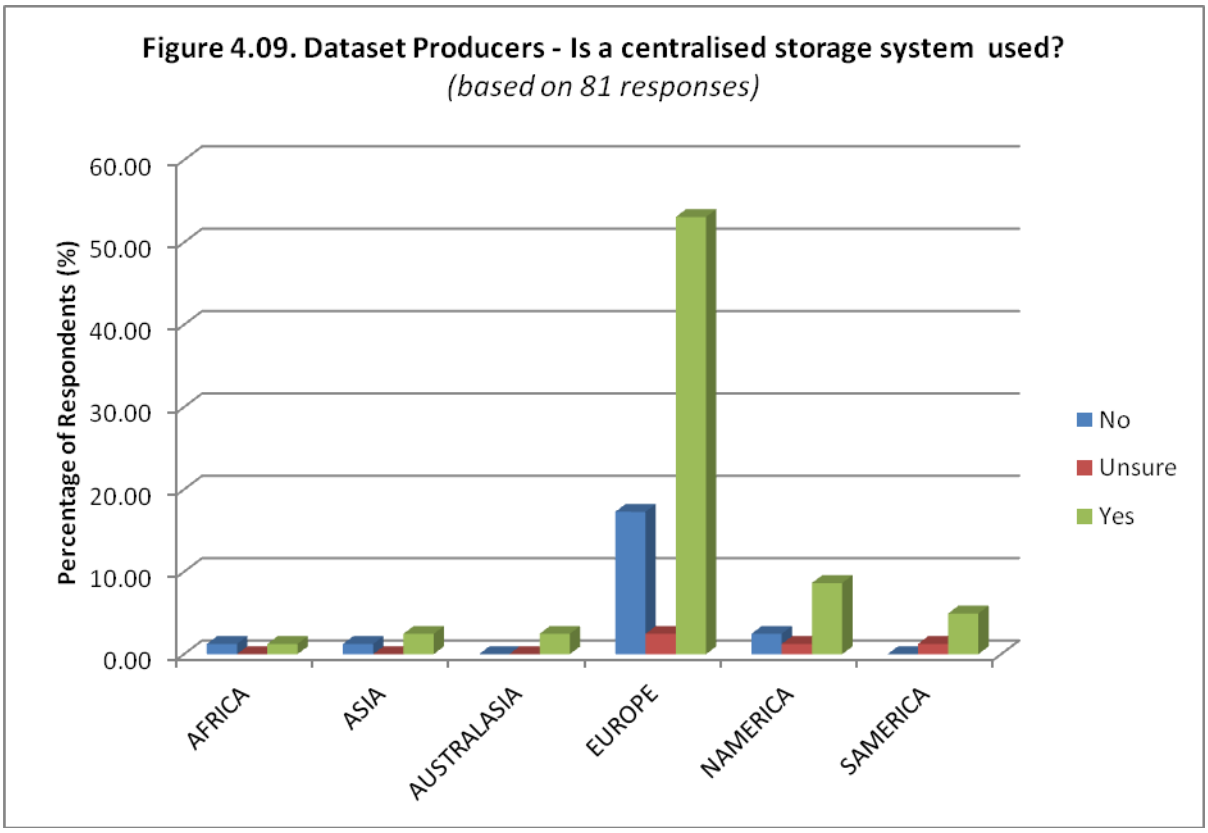


System Name
Earth Observing System Data and Information System (EOSDIS) –NASA system
Linux OpenSUSE 12.1
MySQL 5.5
GeoNetwork
GeoServer
Integrated Rule-Orientated Data System (iRODS) (open source storage middleware)
Postgres SQL database
Oracle
Cassandra
Microsoft Access
Arc Map
MERCI for ATSR and MERIS RR MODIS
Rasdaman

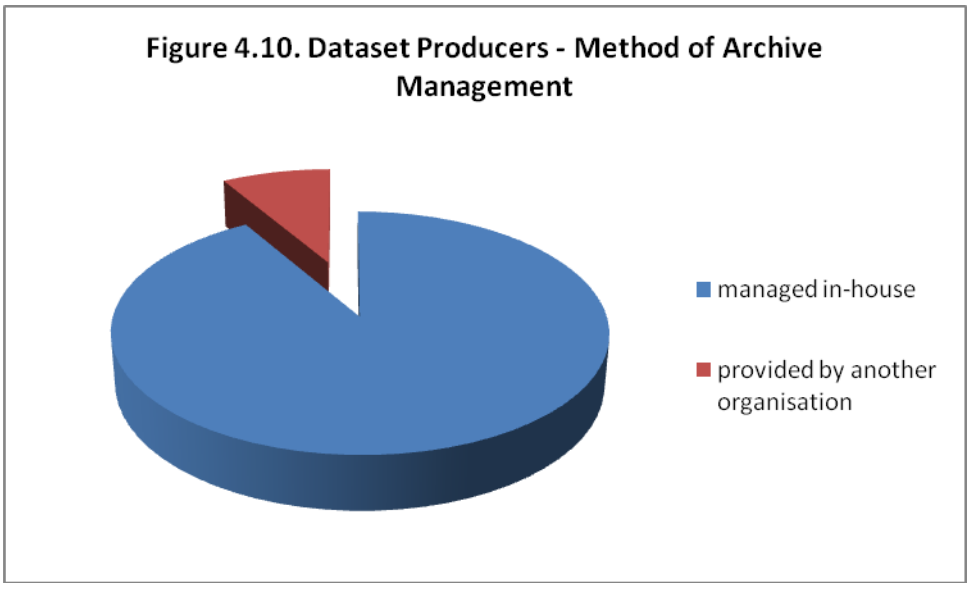
Table 4.03. Specific systems/technologies cited

4.3.1.2 Producers of Major Datasets

The results confirm that (as was expected) the majority of dataset producers have a centralised storage system of some kind for data preservation (Figure 4.09)



In addition the pattern is predominantly for the archive to be managed in house (in over 90% of cases – Figure 4.10)



Considering the relative importance of custom made, open source, and proprietary systems the dataset producer category tends to be dominated by custom developed, and proprietary systems with less prevalence of open source systems (Figure 4.11). Also the 52 respondents to this question are located predominantly within Europe, and so it is difficult to make definitive conclusions for other

areas, other than the fairly low numbers of North American respondents are dominated by custom developed systems. Comparing this with the distribution by discipline (Figure 4.12) shows a tendency towards the use of custom/bespoke developed systems particularly in Earth Observation Science and Atmospheric science with lower frequency of proprietary systems and little significance of Open Source systems .This plot (Figure 4.12) also shows that the dataset producer category is very strongly dominated by representatives of the Earth Observation Science Discipline.

Figure 4.11. Dataset Producers - variation in type of archive system *(based on 52 responses)*

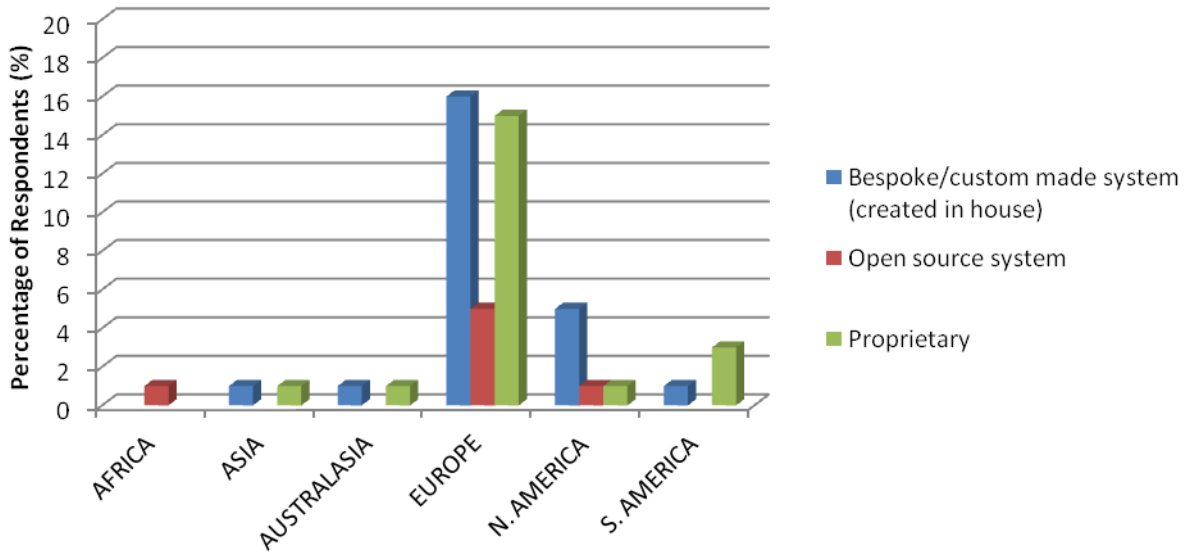
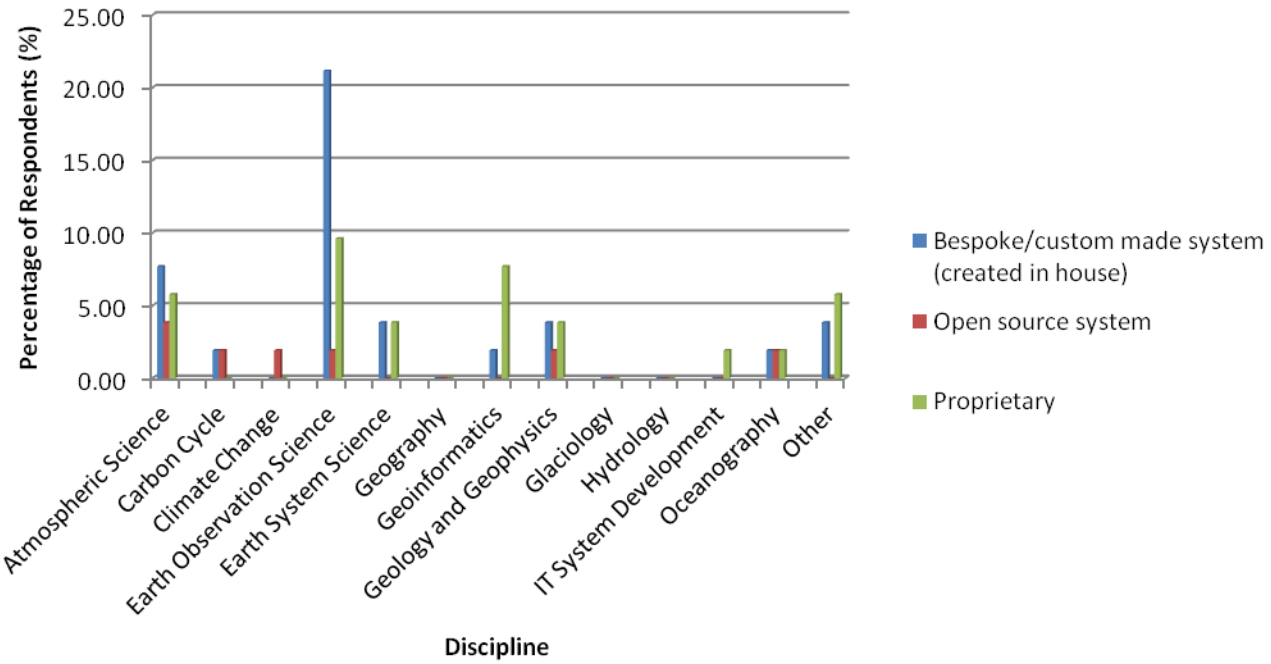


Figure 4.12. Dataset Producers - variation of system type with discipline *(based on 52 responses)*



4.3.2 Results from the direct user consultation - system architecture and infrastructure

The trend for organisations involved in Earth Observation is to use archive systems based on proprietary software sometimes with customisation to their specific needs. In some cases open source technologies such as iRODS (Integrated Rule Orientated Data System) are also employed. There is a wide variation in system architecture ranging from complex systems managing large distributed archives (e.g. the CEDA system), to databases running on one application server and accessed via a web domain as in the some of the earth science domains concerned with in-situ data. The implication particularly in the earth observation domain is that the SCIDIP-ES tools and services will need to interface with a wide variety of systems.

Outside the earth observation domain, particularly where the data archived are held in relational database formats, open source software tends to be used for the building archive systems. Typical open source technologies such as Postgres or MySQL are used to create the underlying database, with PHP and Python used for the development of the scripts to retrieve and process data. Relational database systems ranging from Oracle and SQLServer to open source formats are the most common platforms used for the storage of marine and other environmental data. The common use of these technologies outside the earth observation domain may provide some constraints on the types of technologies which the SciDIP-ES tools and services will need to interface with. Here the cost advantages of open source, plus the practice of maintaining the source code through a user community, are seen as compelling advantages over proprietary systems.

4.3.3 Conclusions - system architecture and infrastructure

4.3.3.1 Summary

In general the archive systems used within the earth observation community tend to be proprietary commercial systems. These are often based on tape archives with disk storage for those data sets requiring more rapid access and may also be customised to meet the precise requirements of the archive. It is common for open source technologies e.g. Python, web services technologies to be used in the development of tools for accessing the archive, and in some cases open source components are also involved in development of the archive structure itself.

The majority of dataset producers indicate that they have a centralised storage system for archiving data, and by far the predominant trend is for the archive to be managed in-house.

Data on the type of system used is available from the archive service provider and data set producer categories. Most data comes from Europe. Here the Earth Observation and Geology and Geophysics domains tend to be dominated by the use of custom developed systems (developed in-house) and also proprietary commercial systems.

There is also a tendency for open source technologies to be used more frequently for the creation of archives developed for other earth science disciplines (e.g. for holding in-situ data in open source database systems such as PostGres and MySQL). There is thus quite a variation in the type of archive storage system both within and outside the earth observation community.

At the same time it was apparent particularly from the direct user consultation exercise in task 15.2 (and described in Section 5 above) that even archive service providers sometimes see their archive system as a “Black Box” to hold their data. Such users typically have a much more detailed understanding of the systems and services which are used to discover and retrieve data from their systems than of the archive architecture itself.

4.4 Discovery of and Access to Archived Data

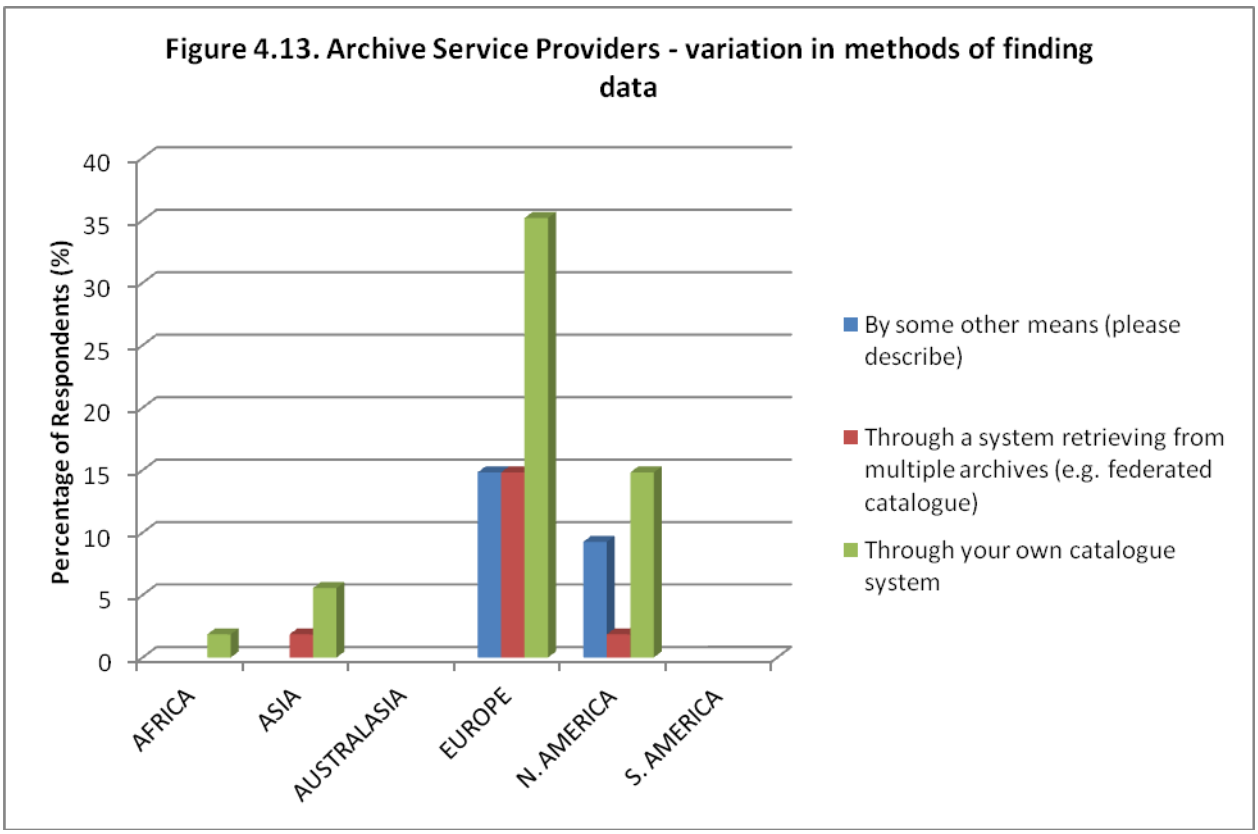
4.4.1 Results from the on-line survey

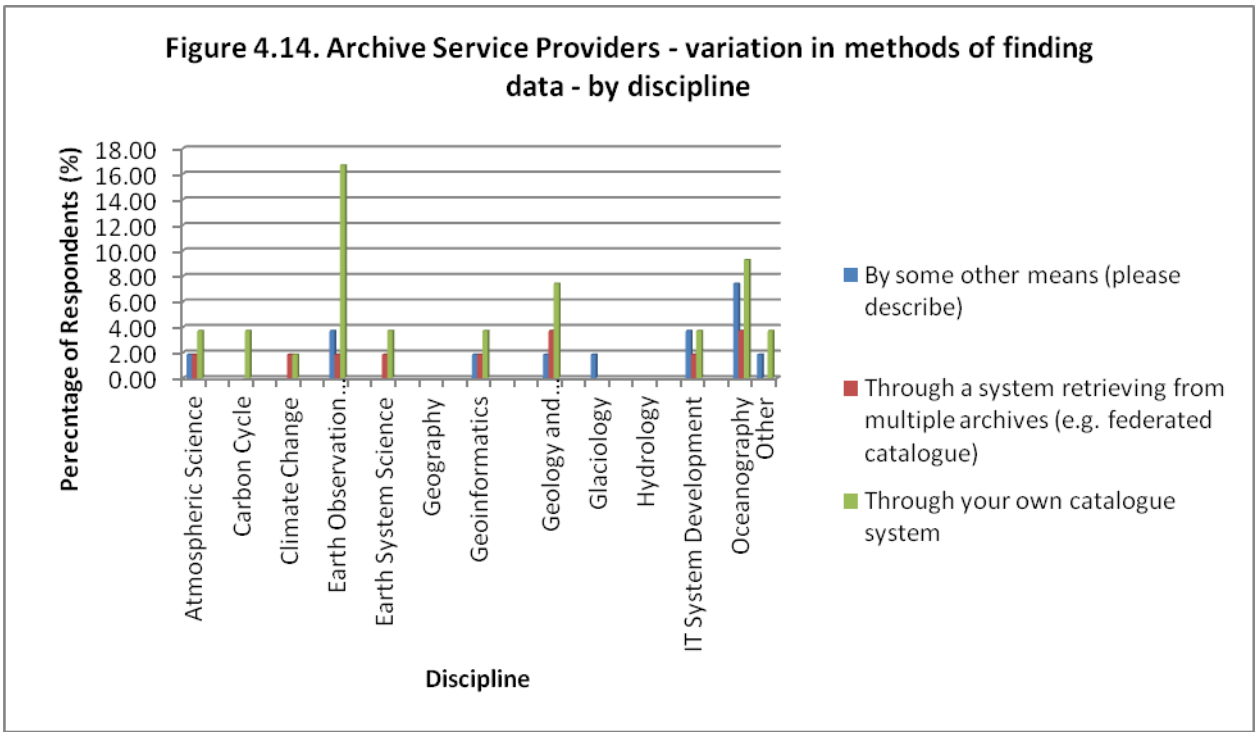
4.4.1.1 Archive Service Providers

Methods of finding data

One of the questions the on-line survey set out to address was to what extent archive service providers provided their own catalogue system and to what extent they provided access to their data through a federated system accessing multiple archives. The results are shown in Figure 4.13, and it is clear that within Europe a significant proportion of all respondents (c.35%) use their own catalogue system, whilst about 15% use a system accessing multiple archives. The pattern in North America indicates that most respondents either use their own catalogue system or some other means of access, with the use of federated catalogues being less important here.

Comparing the method of accessing data with discipline (Figure 4.14) indicates that for the archive service provider category use of the providers own catalogue system tends to be the main means of access, and provision of federated catalogues tends to be less important. In the disciplines which have generally small numbers of respondents there is a mixed pattern with the providers own catalogue system being commonly used, but also some use of federated catalogues.





Where respondents had indicated that some other means was involved they frequently described using a combination of their own and federated catalogues. The overall impression is therefore that whilst the overall trend seems to be for providers to use their own catalogue, there is nevertheless considerable use of federated catalogues.

How data is made available to users

Considering how the archived data itself is made available to users, the overall trend is for the use of on-line and web service based delivery systems, and these methods account for nearly 80% of respondents, with off-line ordering systems accounting for only a small proportion. This trend is common across Europe and North America (Figure 4.15)

Overall off-line ordering systems constitute a relatively low proportion of the mechanisms used by the archive service providers surveyed. However, this is still a significant method used in the earth observation science and geology and geophysics disciplines as shown by the plot of method of delivering data versus discipline in Figure 4.16. This possibly may be a reflection of the extremely large size of some datasets in the Earth Observation area, which would be difficult to deliver over the web. In the other science disciplines the use of off-line ordering systems is conspicuously low.

Figure 4.15. Archive Service Providers - how data are made available to users

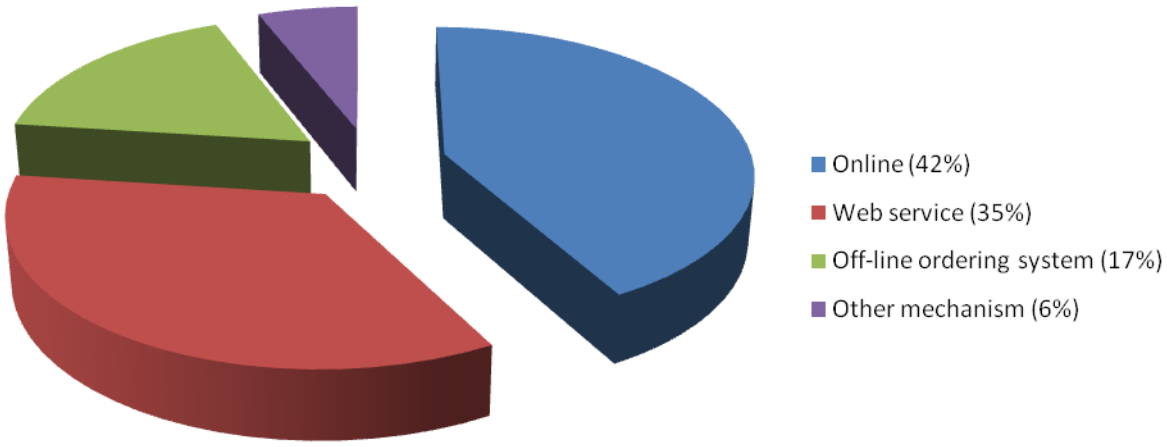
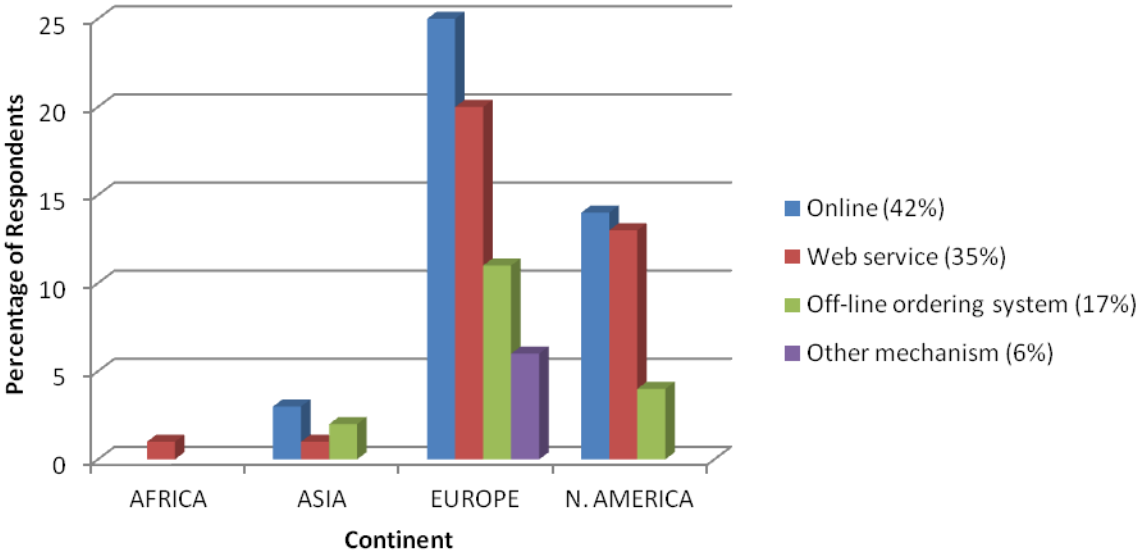
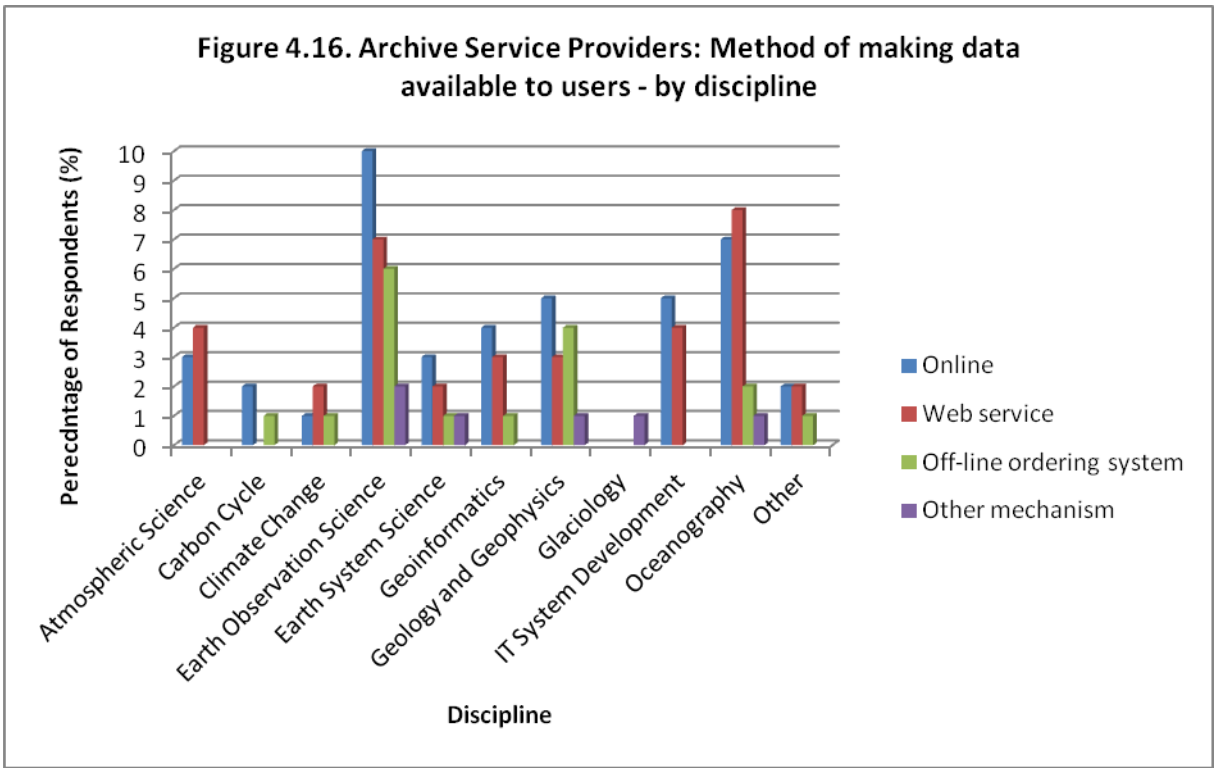


Figure 4.15a. Archive Service Providers - how are data made available to users





In what format is data received from the archive?

Respondents were also asked what users actually received from the archive, for example whether just the original files, or whether any processing or other value adding procedures were employed. The results indicate a fairly mixed response, with the distribution between these categories being fairly similar between Europe and North America and the provision of data via web services and processed data being most common, with derived products also often being delivered (Figure 4.17). Comparing the types of data delivered by discipline (Figure 4.18) suggests that the provision of data by web services is important in the three most dominant disciplines represented (Earth Observation Science, Geology and Geophysics, and Oceanography) as is the provision of processed data, though providing processed data and partial representations of the data tend to be more important in Oceanography. Otherwise across the other science disciplines there is no clear trend between the type of data delivered and discipline.

Figure 4.17.Archive Service Providers - How users receive data *(based on 64 respondents)*

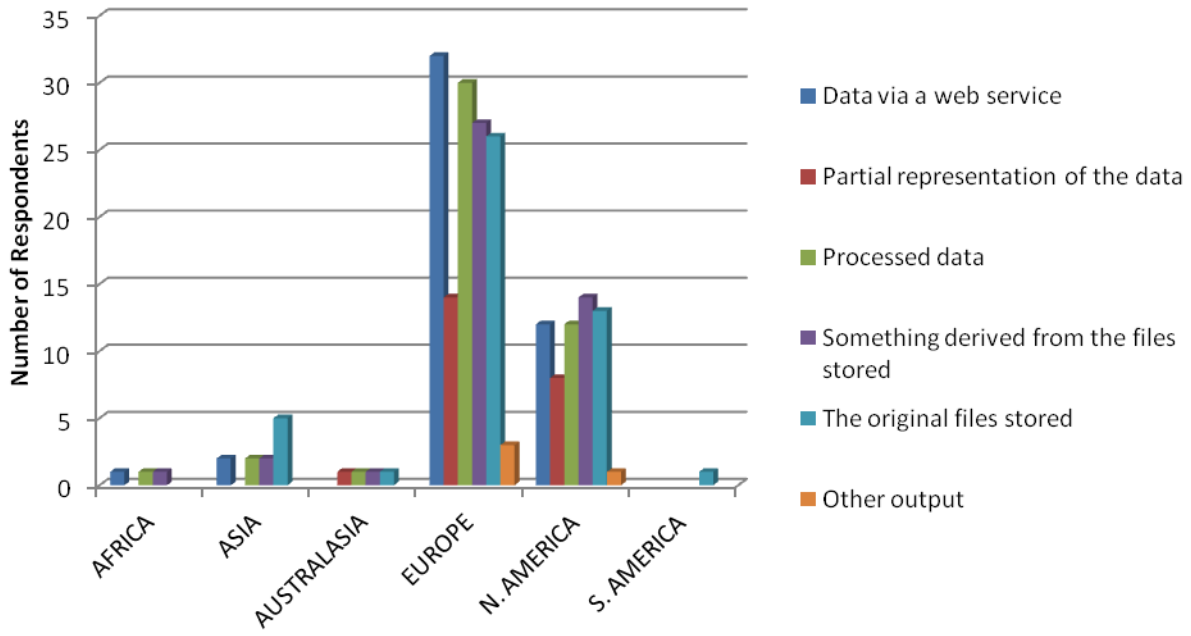
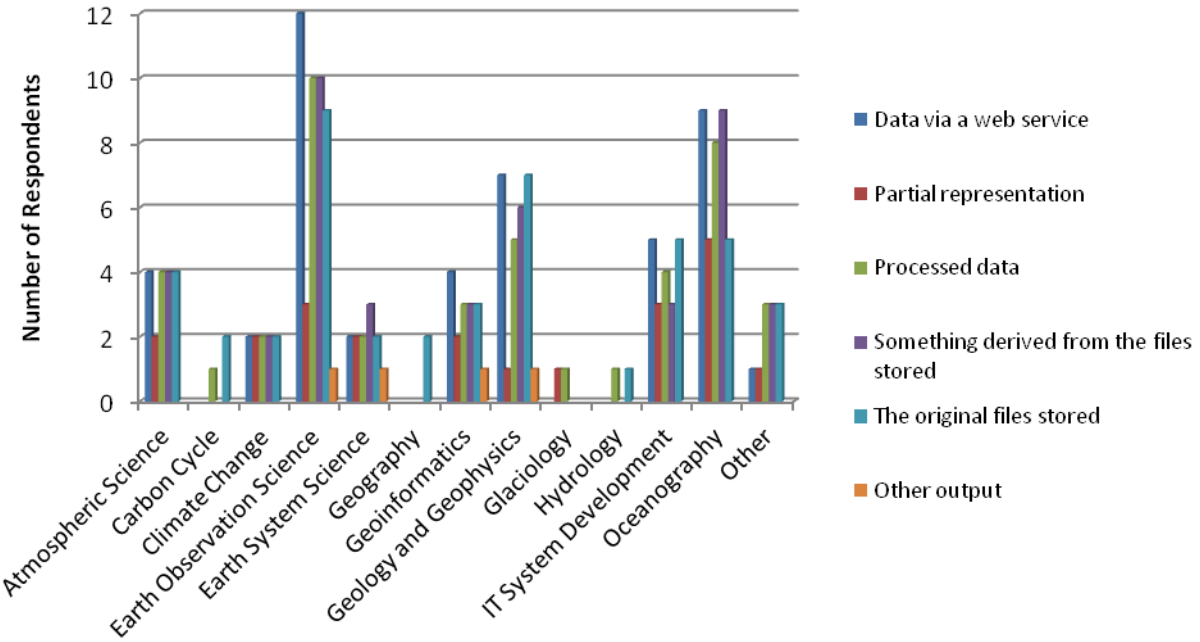


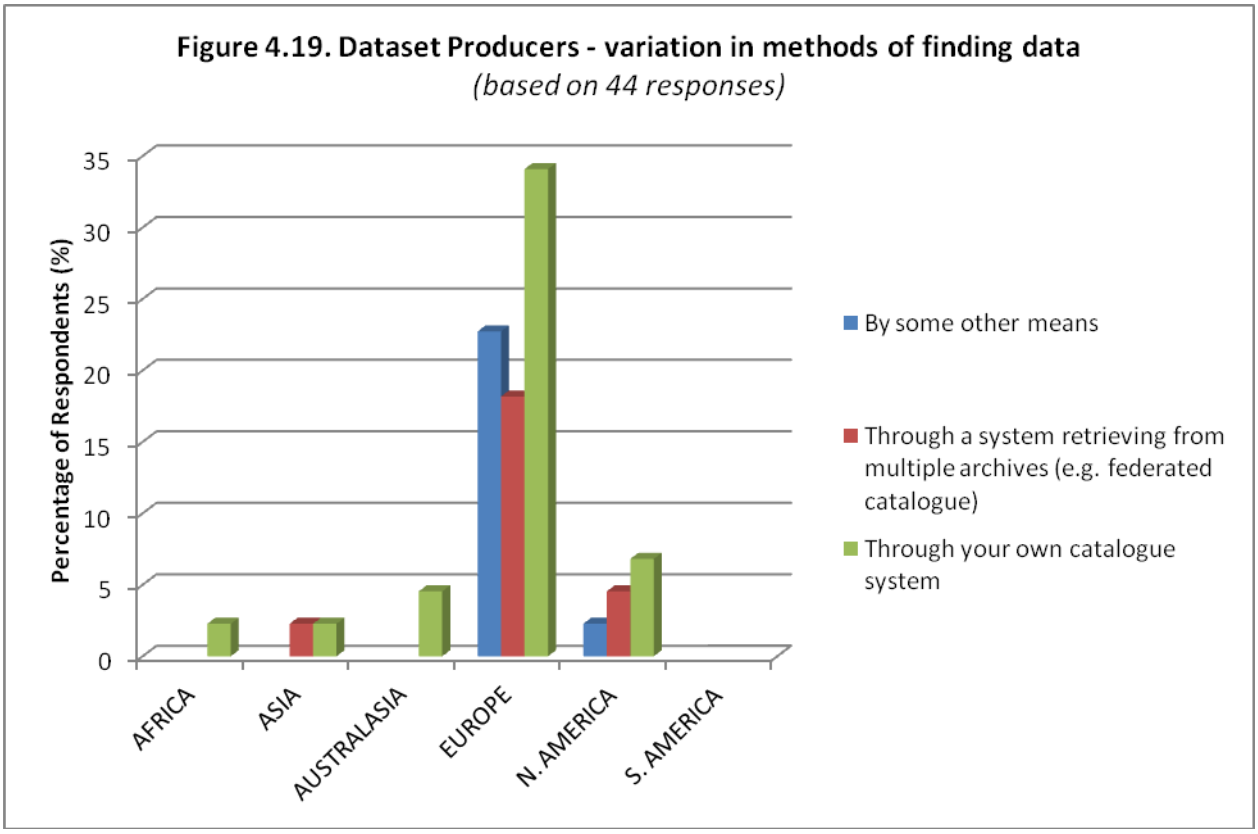
Figure 4.18. Archive Service Providers - How users receive data by Discipline *(based on 64 respondents)*



4.4.1.2 Producers of Major Datasets

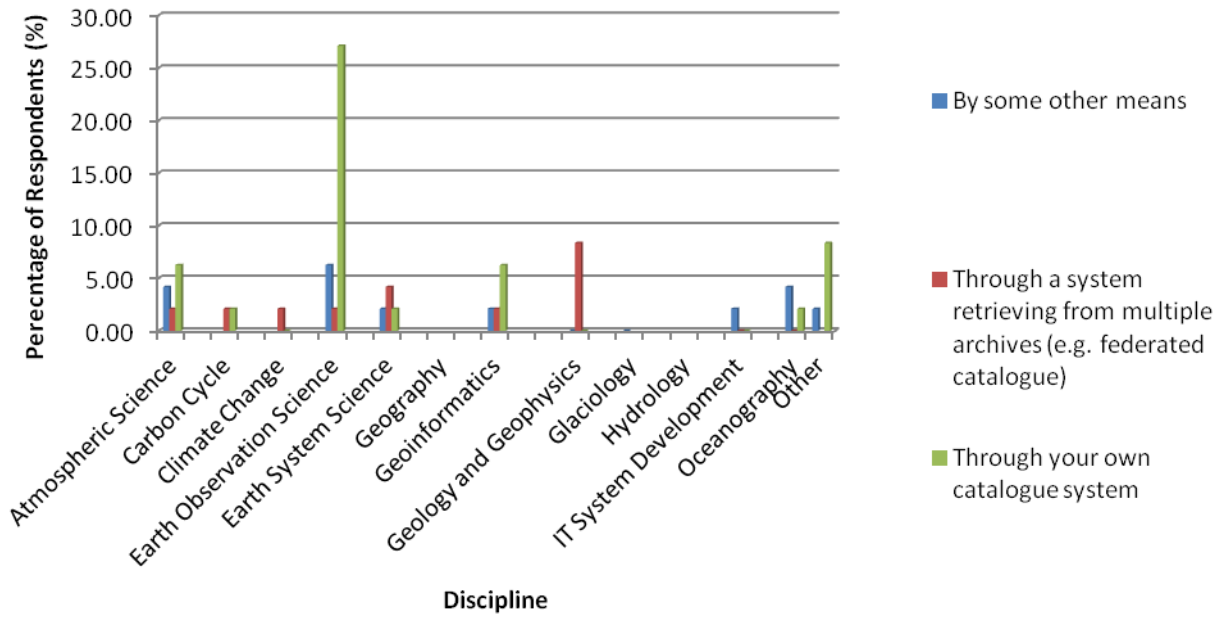
Methods of finding data

The distribution of methods of finding data (e.g. through users own catalogue system, or through a federated system etc) for dataset producers indicates a predominance of users own catalogue systems being used, with a strong tendency also to use federated catalogue systems, particularly in Europe where most of the dataset producers who responded are located.



Considering variations in how data is discovered by discipline (Figure 4.20), the use of the users own catalogue is the most important method in Earth Observation Science. Whilst there are lower numbers of respondents in the Geology and Geophysics discipline here, there is an indication that these are predominantly using a catalogue system retrieving from multiple archives.

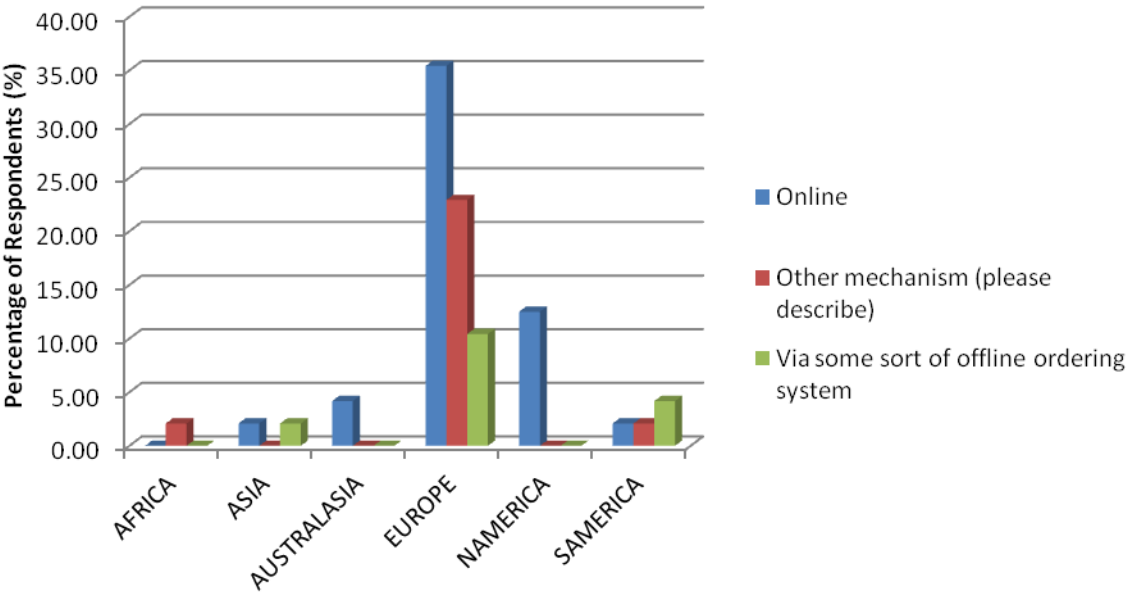
Figure 4.20. Dataset Producers- variation in methods of finding data - by discipline *(based on 44 responses)*



Where users have indicated "by some other means" this generally refers to the use of some sort of internal system to locate data, which is generally web based.

How is data made available to users

Figure 4.21. Dataset Producers - mechanisms of data delivery *(based on 48 responses)*



The tendency is predominantly for use of on-line delivery, with offline ordering indicated only by a small proportion of respondents (Figure 4.21). On-line delivery is also the predominant mechanism for delivery of data in most disciplines. In earth observation science (Figure 4.22) the "Other" mechanism of delivery is generally via FTP or via a combination of on-line and off-line delivery.

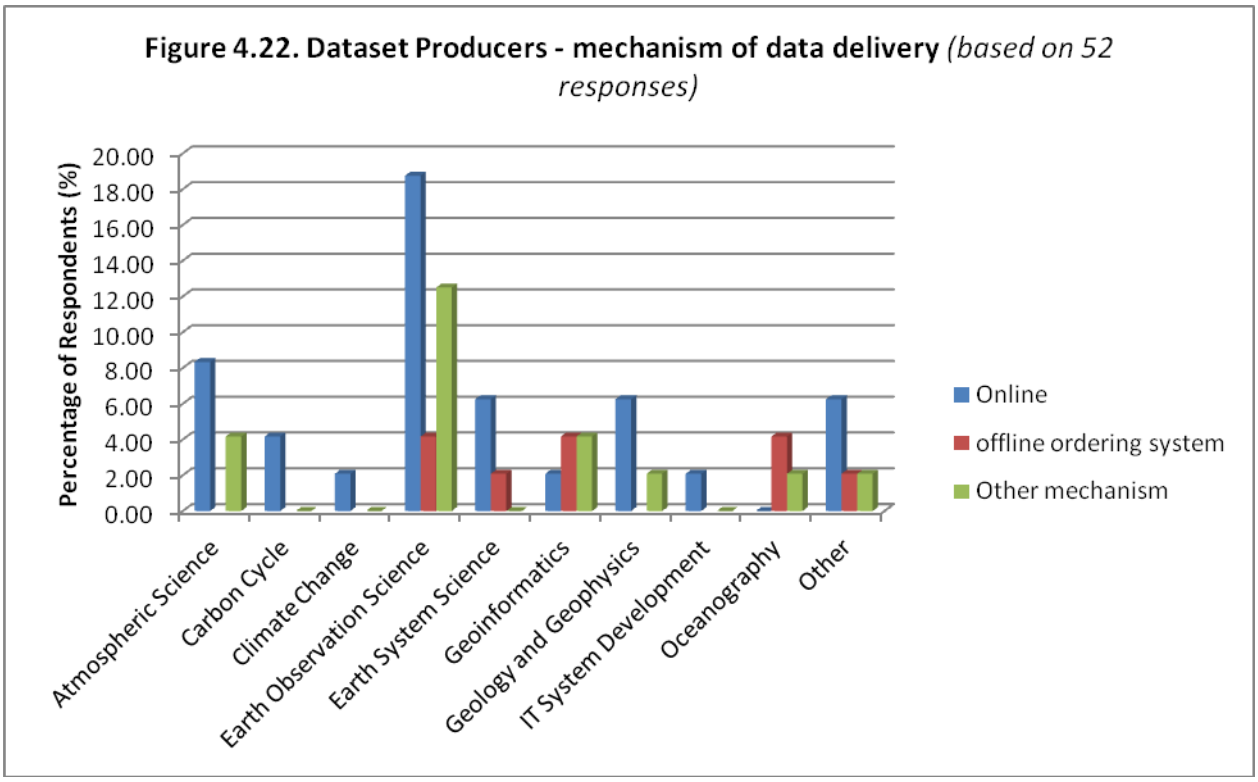
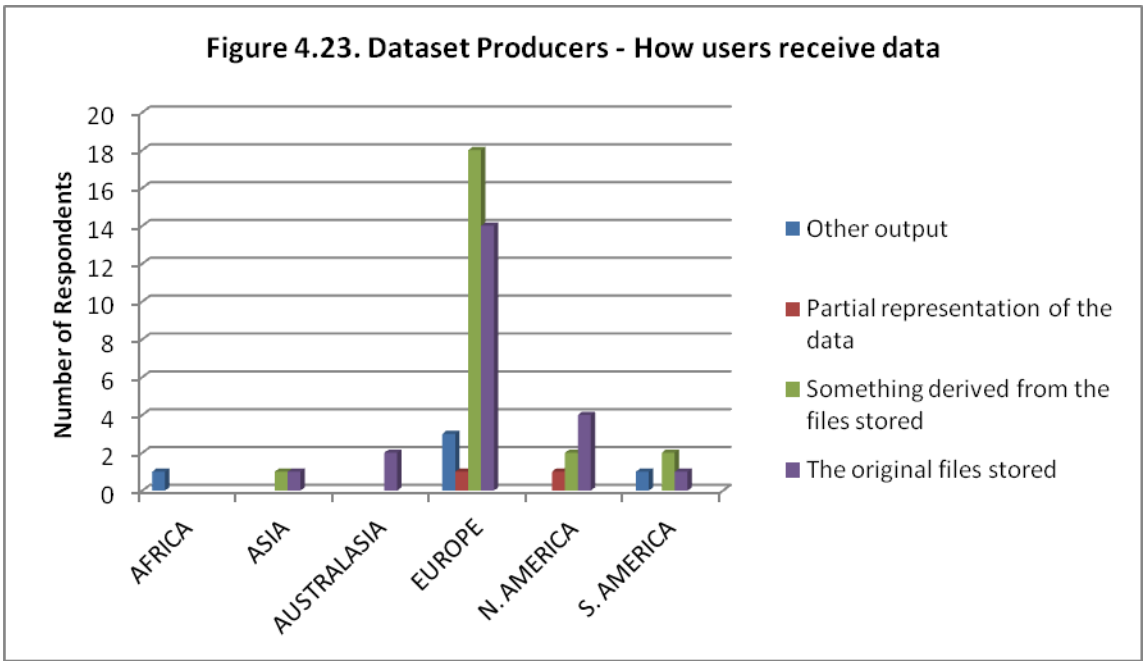


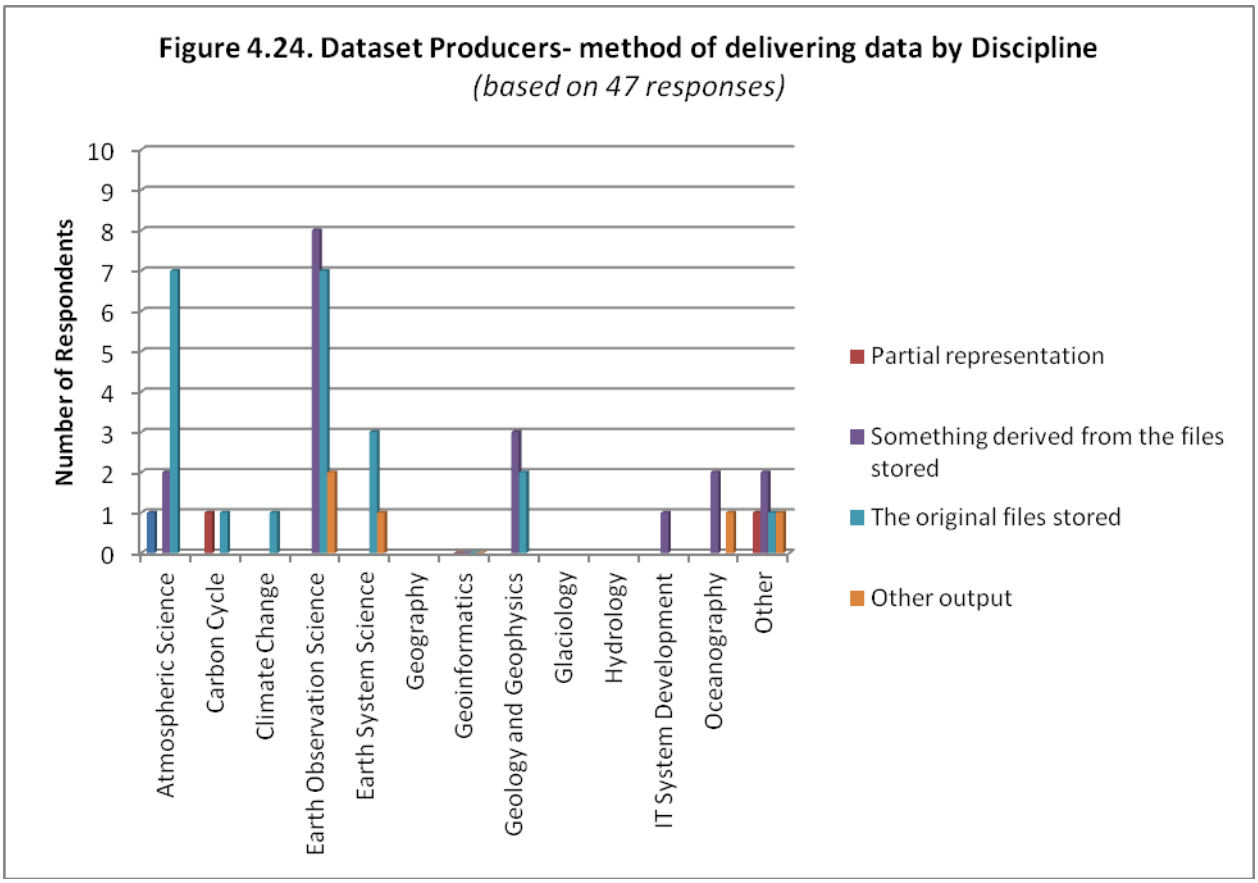
Figure 4.22. Dataset Producers - mechanism of data delivery *(based on 52 responses)*

In what format is data received from the archive

In terms of in what form users actually receive data the most important mechanisms in Europe are to provide the original files stored or something derived from the files (Figure 4.23).



Providing users with the original files is the predominant format for delivery of data to users in the Atmospheric, and Earth System sciences (Figure 4.24) In Earth Observation and Geology and Geophysics providing the original files, as well as delivering derived products from the original data are the main mechanisms of data delivery. Figure 4.24 also indicates that producing partial representations of the data to deliver to users is not an important mechanism for the data set producer category.



4.4.1.3 End Users of Archive Data

Methods of finding data

Considering the search mechanisms users prefer, the pattern for Europe, North America and Asia is fairly similar (Figure 4.25) with the use of a specific catalogue and searching through a multi-archive search generally being equally important. Breaking the European component down by country (Figure 4.26) also indicates that overall multi-archive searches and search by a specific catalogue are generally about equally popular, though for a number of countries the multi-archive search is slightly more common than searching through a single catalogue.

Figure 4.25. Archive Users - How users find data *(based on 274 responses)*

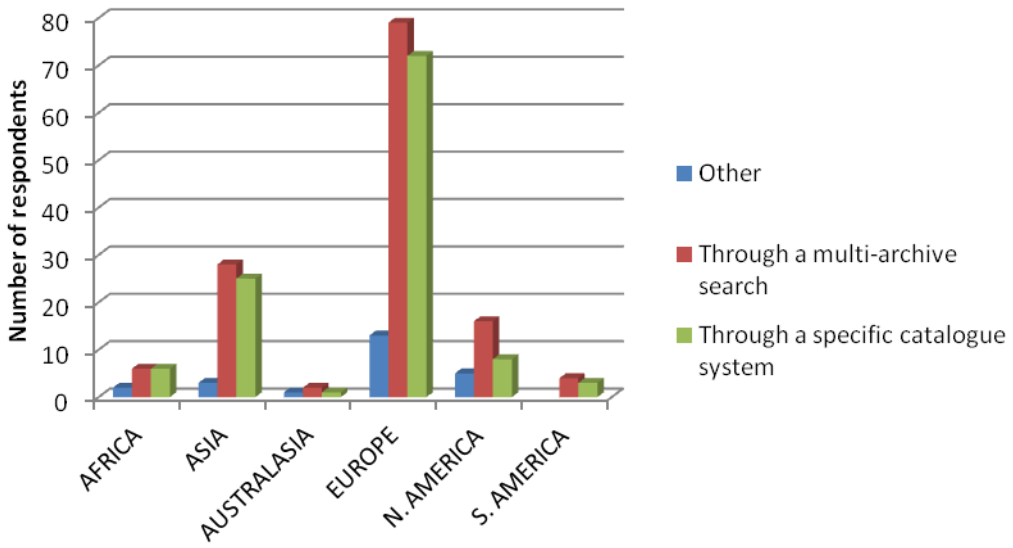
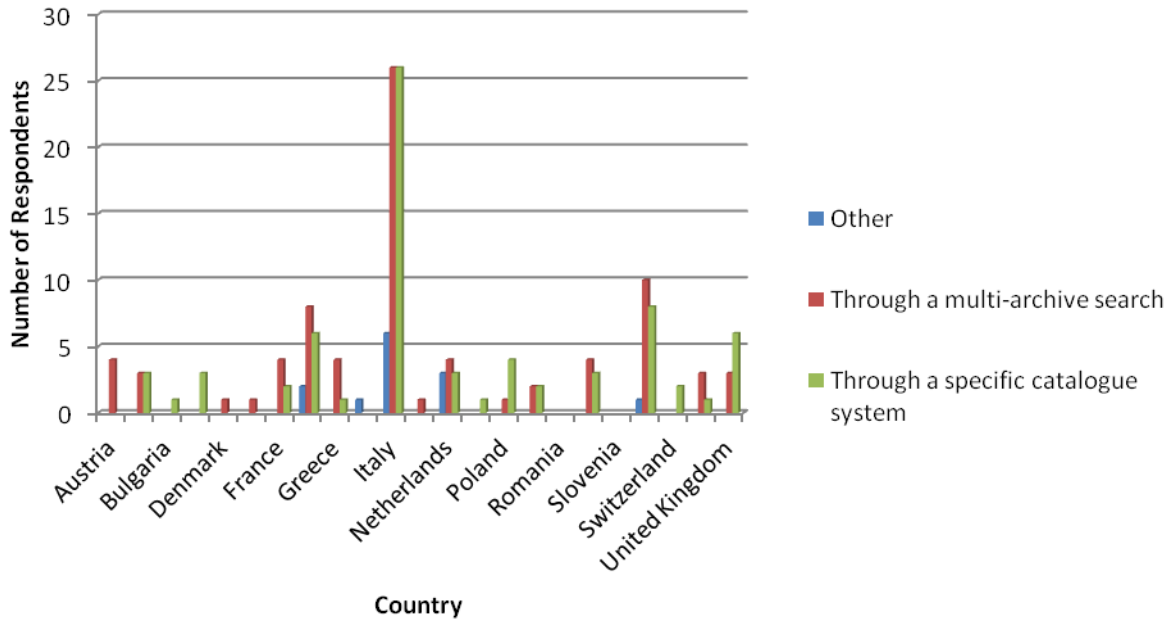
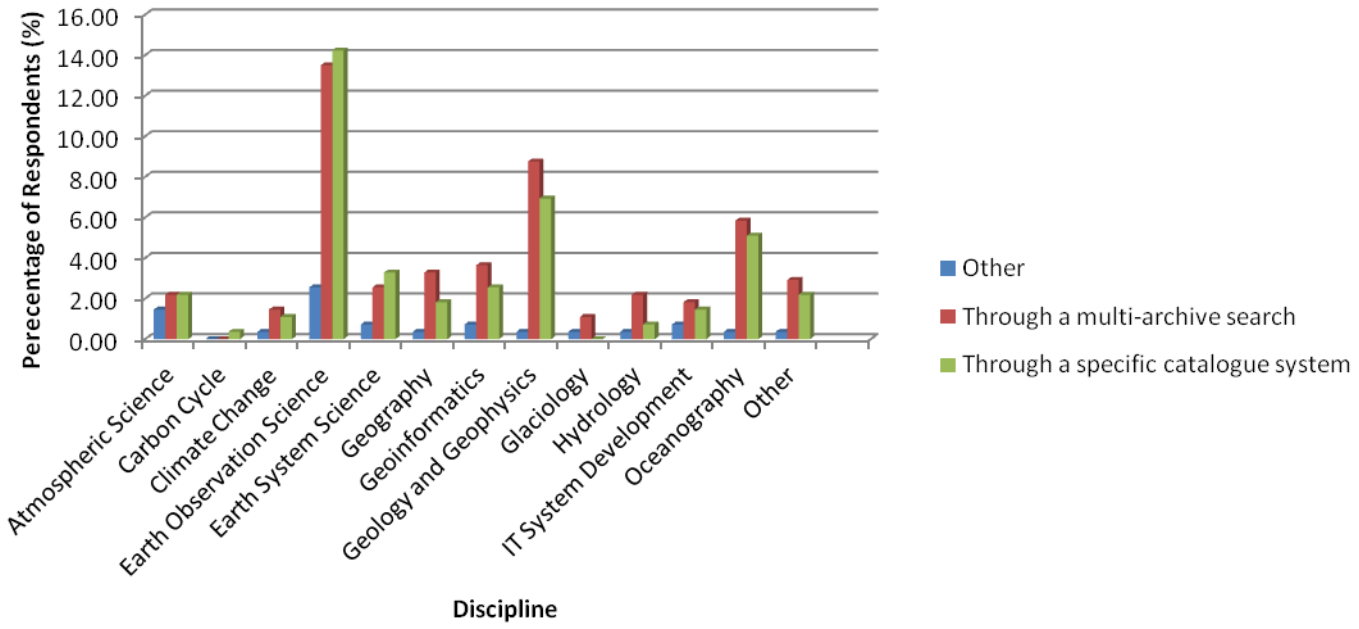


Figure 4.26Archive Users in Europe - method of finding data *(based on 274 responses)*



Considering the various scientific disciplines there is not really a significant difference in the searching mechanisms used between disciplines, and in most cases there is an almost equal proportion of searching through a single archive and multi-archive search (Figure 4.27).

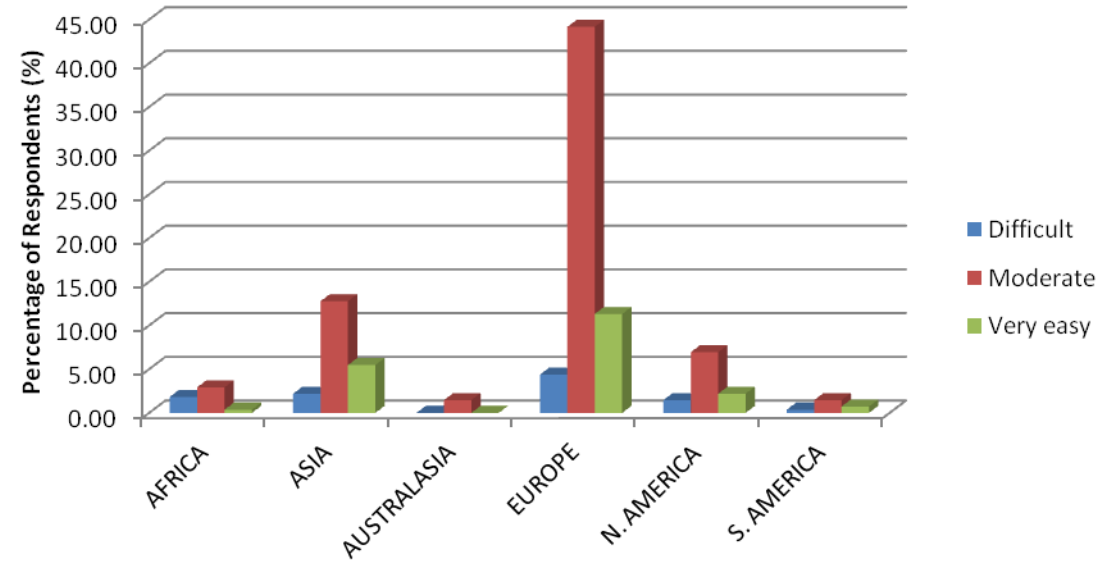
Figure 4.27. Archive Users - methods of finding data by discipline *(based on 274 responses)*

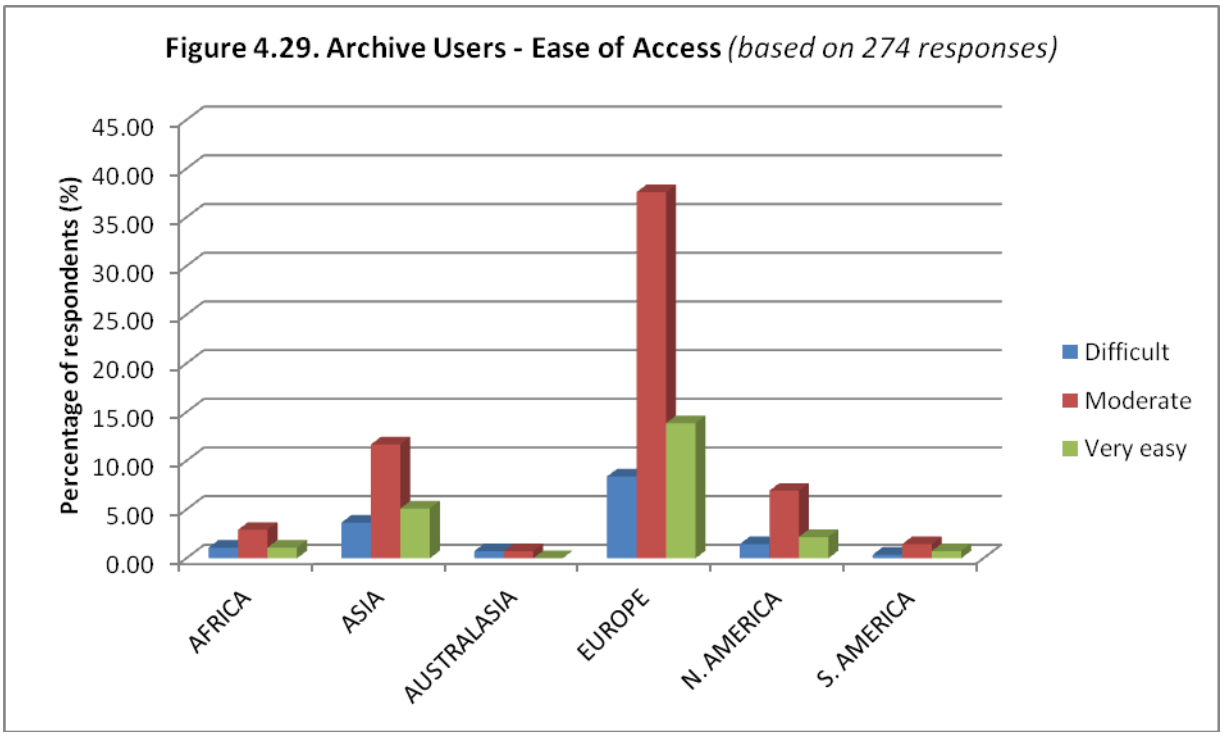


Ease of locating and accessing data

The survey explored both the ease of locating data and the ease of access to data. The results indicate that in both cases the majority of respondents indicate that both operations are moderately easy (Figures 4.28 and 4.29), and the same overall trend occurs across Europe North America and Asia.

Figure 4.28. Archive Users - Ease of Locating Data *(based on 274 responses)*





The variation of both ease of locating and accessing of data is fairly similar across the various disciplines (Figures 4.30 and 4.31), though for Earth Observation Science there is more spread in the ease of access data suggesting that some users in Earth Observation Science and in Geology and Geophysics find accessing data slightly easier than locating it, and some also find accessing data more difficult than locating it.

Figure 4.30. Archive Users - Ease of Locating Data by Discipline *(based on 274 responses)*

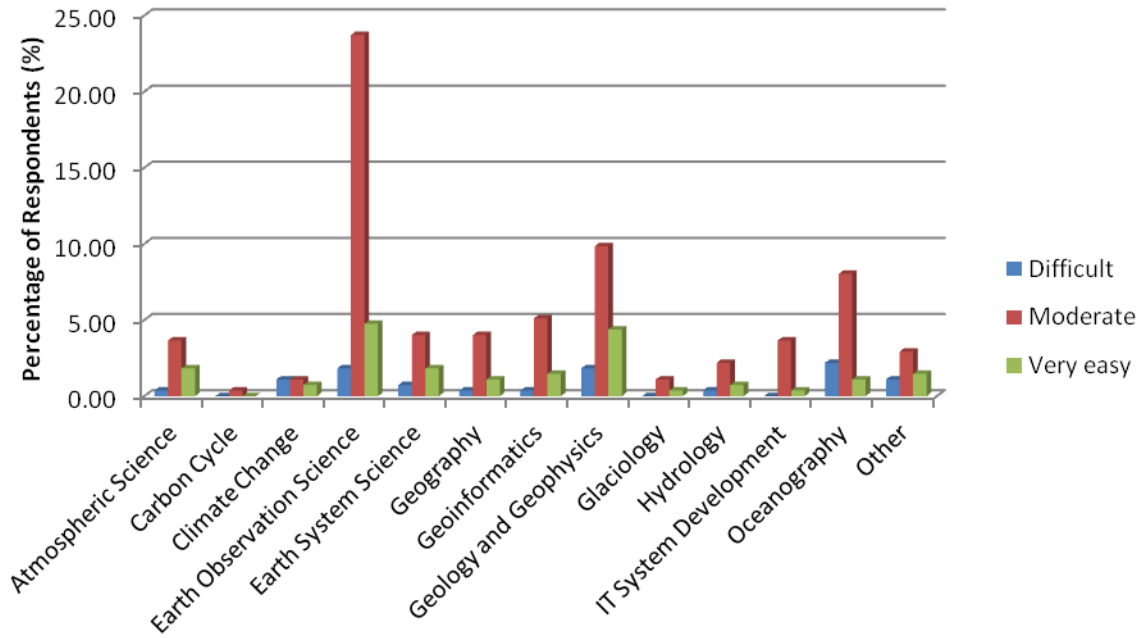
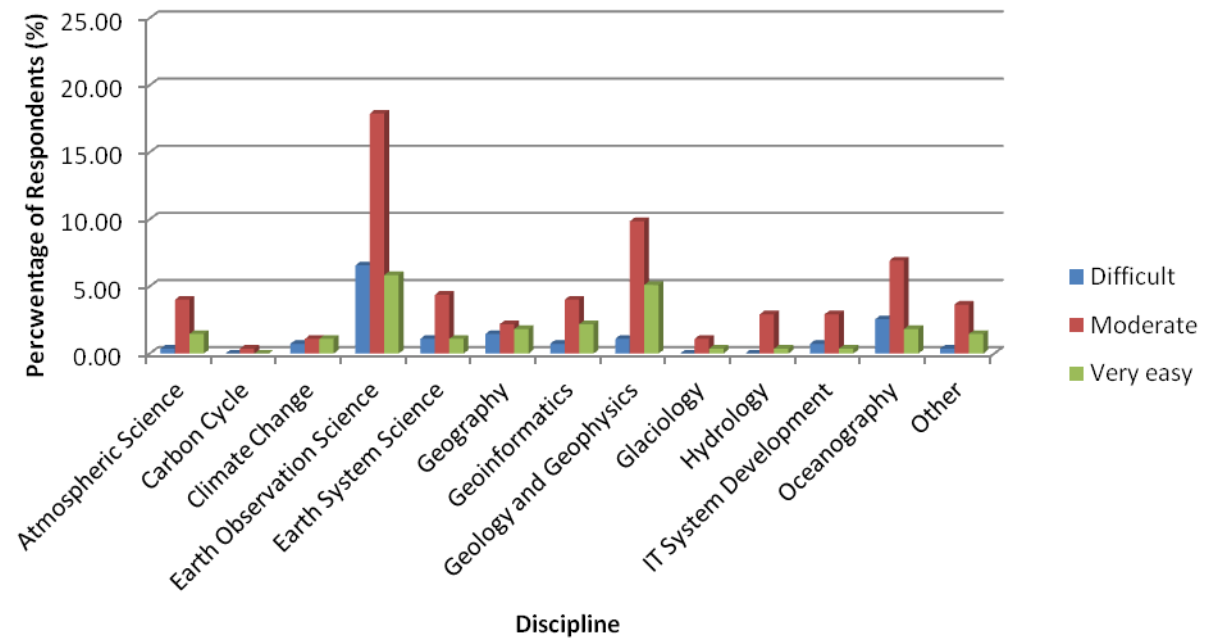


Figure 4.31. Archive Users - Variation in Ease of Accessing Data by Discipline *(based on 274 responses)*



Users were asked how ease of locating data can be improved, the suggestions received fall into a number of themes as described below:

● Increased use of multi-archive searches, to avoid trying to find products from different providers across different archives.

● There is also some interest in being able to access multiple catalogues from within the software used to perform the data analysis.

● Provision of better designed web interfaces to facilitate faster access to data, a Google style search interface was cited by some users.

● More detailed metadata, for example the processing levels available etc. and allowing data for different sensors in different archives Search by keyword as well as by spatial co-ordinates and bounding box.

● Improvements in common dictionaries and ontologies to assist in establishing semantic relationships between different datasets.

When asked how data access can be improved the following main themes were as follows:

● More use of direct ftp downloads, rather than delivery of data by email

● Making web portals more user friendly

● Simplify the processes for obtaining authorisation to access data

● More application of open data policies to Earth Observation data

● Improving tools for visualisation such as the NASA Glovis program

● Centralisation/federation of catalogues with links to the resources to be downloaded

● Establishment of uniform methods for accessing data through standardised services

Is additional information required to discover data?

The issue of whether the user would need to have access to additional information to help them discover data was also explored. Overall there was strong trend towards users requiring such additional information (Figure 4.32). Considering how this requirement varies by discipline, most users within Europe, and North America indicate that they need such additional information across all disciplines. Though for Geology and Geophysics around half the respondents reported that additional information was not required.

Figure 4.32. Archive Users - additional information required to discover data *(based on 244 responses)*

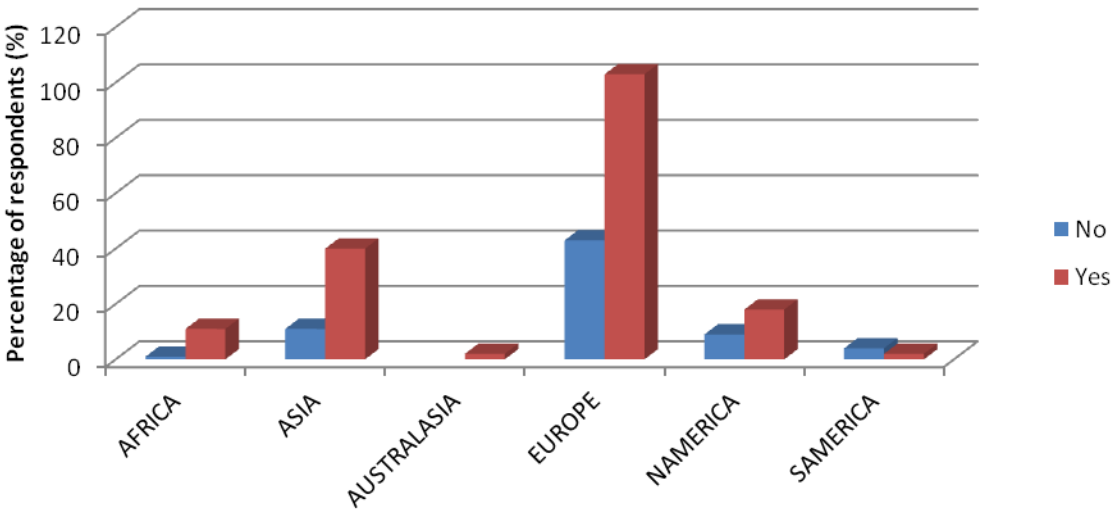


Figure 4.33. Archive Users in Europe: Whether additional information is needed to discover data *(based on 244 responses)*

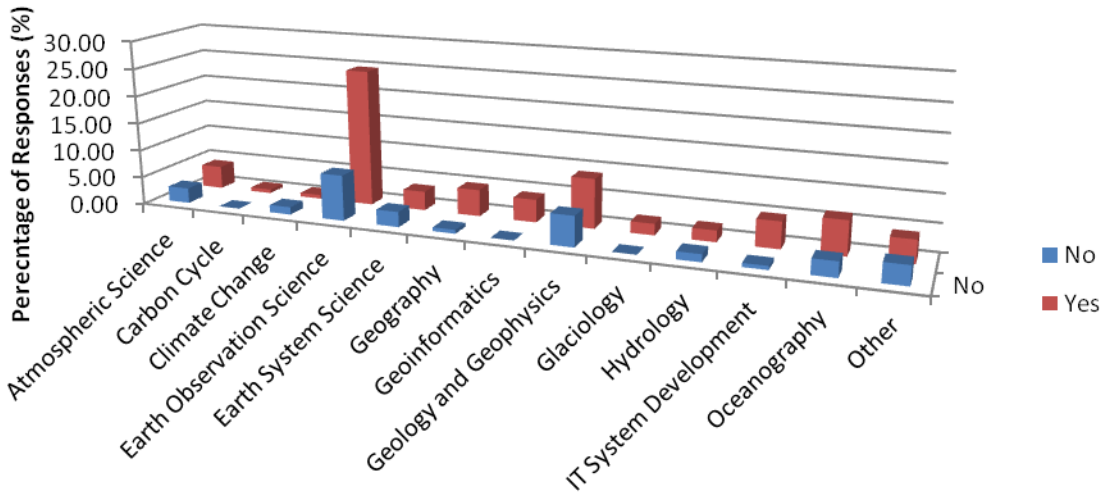
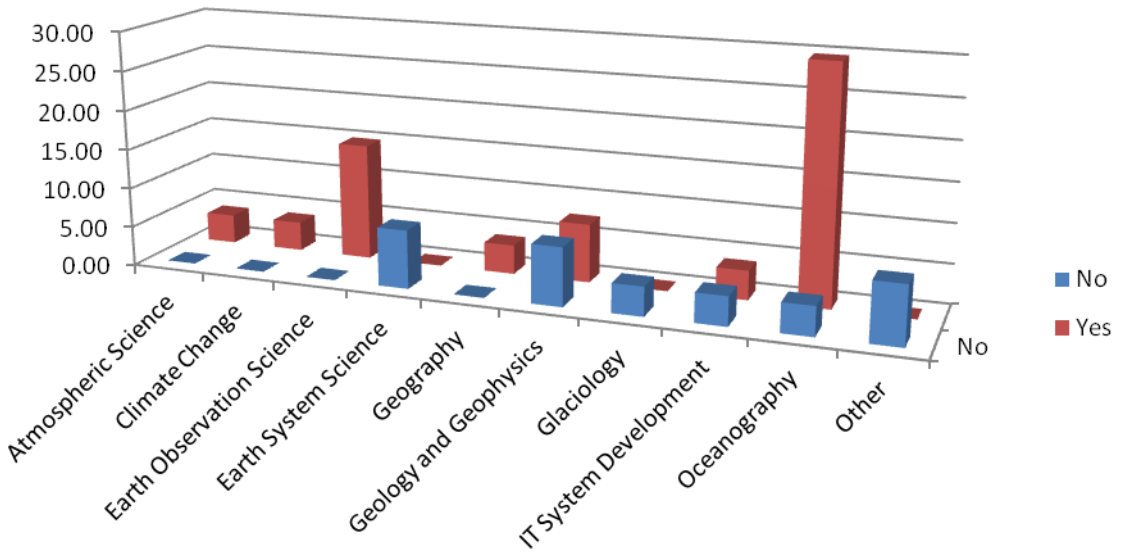


Figure 4.34. Archive Users in North America: Whether additional information is required to discover data *(based on 244 responses)*



Users were asked whether the data received is in a suitable format for analysis, or whether conversion was required. The results were compared between Europe and North America (Figures 4.35 and 4.36). Generally in Europe the trend is for the user being able to analyse the data without conversion, though in Earth Observation, Geology and Geophysics there are also a significant proportion of users who need to perform conversion. Whereas in North America (Figure 4.36) there is a trend towards more conversion required particularly in Geology and Geophysics and Oceanography, though this is based on smaller numbers of respondents than in Europe.

Figure 4.35. Archive Users in Europe - Whether data is retrieved from the archive in a useful format *(based on 244 responses)*

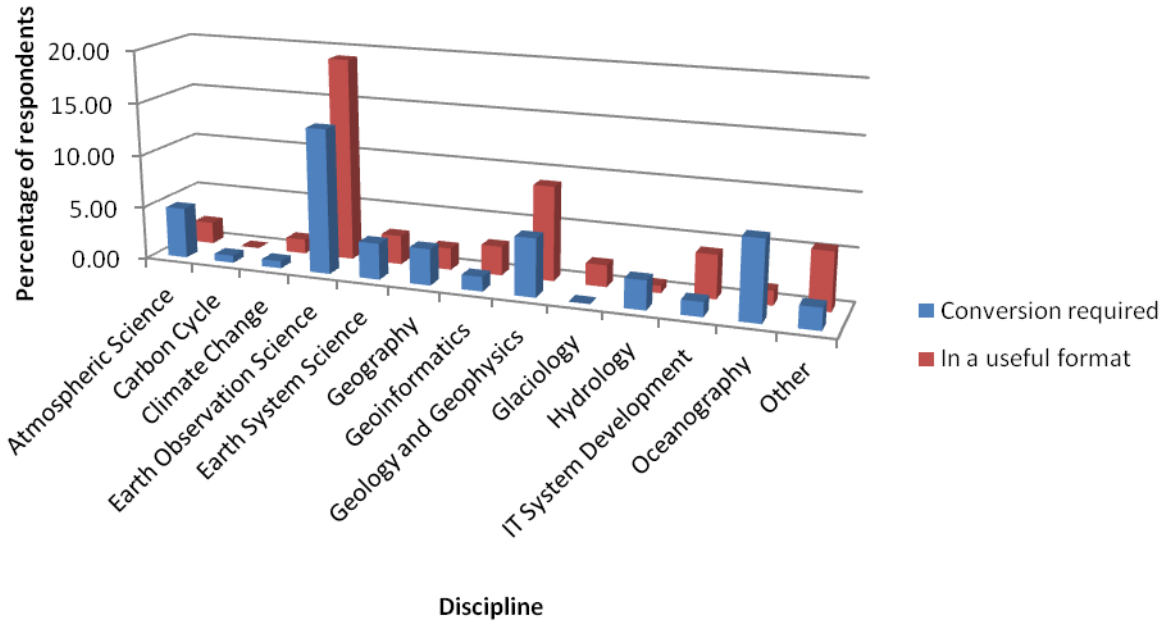
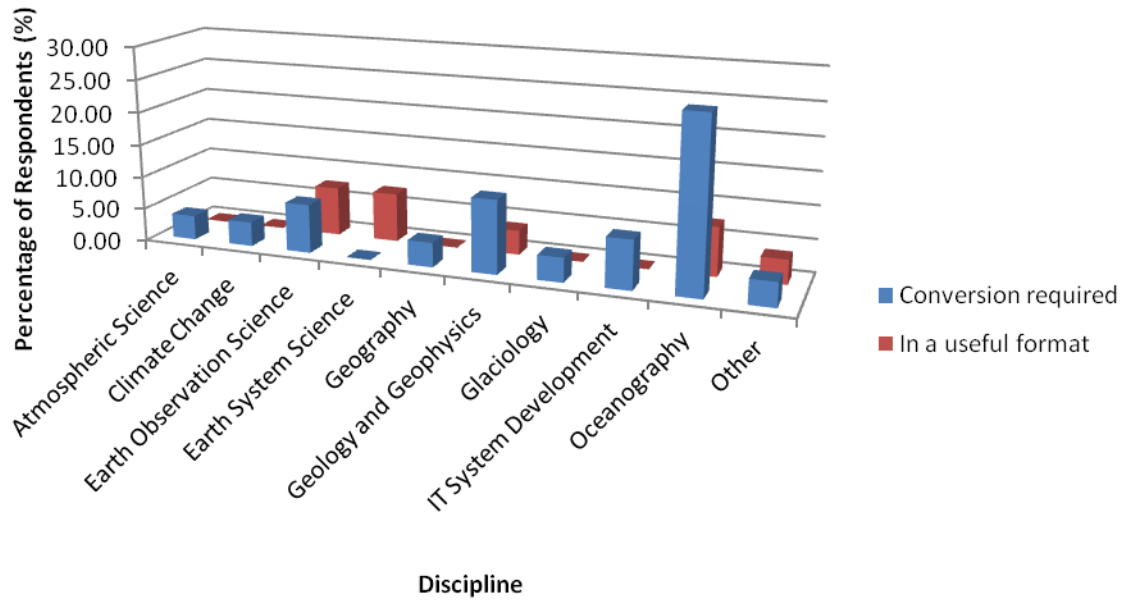


Figure 4.36. Archive users in North America - whether data is retrieved in a useful format *(based on 244 responses)*



4.4.2 Results from the direct user consultation ("one-to-one" interviews) - discovery of and access to archived data

There is a tendency towards the use of open source tools in order to develop systems for accessing and processing data across the majority of the disciplines surveyed, including organisations that use proprietary commercial technologies for the archive itself. Again technologies such as PHP and Python are frequently used to develop applications. Other technologies cited include "MDweb" (an open source cataloguing tool) for developing data referencing and validation systems. This tool is compliant with the ISO metadata standards 19115, and 19119.

There is an indication across the organisations surveyed of strong use of ISO metadata standards such as ISIO 19115 and ISO19119. The desire to promote interoperability between disparate data sets is one of the drivers for this compliance particularly within the in-situ data community. Another driver is the desire to promote access to individual data sets by facilitating metadata harvesting. Harvesting methods include use of OpenSearch 1.1 and the Open Archives Protocol for Metadata Harvesting (OAI-PMH).

The development of data catalogues often using open source and web services technologies is an important gateway for data access to a number of archives (e.g. the data catalogue developed by ISPRA and the NERC Data Catalogue). The CF (climate and forecasting) metadata standards are also frequently applied to earth observation and atmospheric science data.

The Meta Data Objects for Linking Environmental Sciences (MOLES) system developed by the British Atmospheric Data Centre (BADC) provides a means of semantically linking datasets between different institutions so that the meaning of the links is evident. MOLES also provides a method of elucidating the relationships between activities, data, instruments, algorithms etc in a semantically consistent and standards compliant manner. This methodology is in the public domain and is currently in use at a number of environmental data institutions in the UK including The British Atmospheric and Oceanographic Data Centres, the Plymouth Marine Laboratory, and the NERC Earth Observation Data Centre.

The Earth Observation Product Heterogeneous EO Missions Accessibility metadata schema (EOP-HMA) also provides a method of recording metadata for earth observation data

In line with the indications from the initial on-line survey (described in section 4), the second phase of user consultation showed that the earth observation domain commonly makes use of web services to access the archive through web based query forms, and offline ordering which also sometimes combined with an FTP download facility.. Specific web services technologies used include the WCS, WMS, WFS standards.

The use of the GeoTIFF and NetCDF formats to deliver data to the user is common practice in the earth observation domain. NetCDF is also commonly used across the various other earth science disciplines. In addition to NetCDF and GeoTIFF, other data formats for earth observation data include NASA AMES format. Other common data delivery formats used outside the earth observation discipline include data in text based tabular form including BADC format .csv, ESRI shape files, GML (geographic mark-up

language), and XML. Other image formats used include .jpeg, .png, and .tiff. There is a tendency for most data delivery to utilise a fairly restricted set of data formats including those aforementioned.

4.4.3 Conclusions - discovery of and access to archived data

4.4.3.1 Summary

For respondents in Europe the trend for archive service providers is very similar to that for dataset producers with accessing data through their own catalogue being the predominant method for finding data, with using federated systems still being significant but of lower relative importance. Comparing this with the trend for archive users, searching using a multi archive search is used more frequently by this group, with the number of respondents in Europe (and in fact in Asia) split approximately equally between these methods of finding data. This trend may possibly be an indication that archive service providers and dataset producers are more aware of the existence of their own catalogues, whereas potential archive users have a less vested interest in using a particular dedicated catalogue system. It is also likely that this trend reflects the need by users to access data from multiple domains, and this requirement is in fact cited by a number of archive users when asked about improvements in usability (below).

The trend towards use of the organisations own catalogue system is generally maintained across most of the scientific disciplines represented by the archive service provider and archive producer categories, with no major differences between discipline being apparent, other than possibly more significant use of federated catalogues by dataset producers in the geology and geophysics domain. Also the trend for archive users to make more use of federated search is apparent across most disciplines represented.

The use of various metadata standards has been explored in the surveys within Tasks 15.2 and 15.3. ISO metadata standards for earth science data, particularly ISIO 19115 and ISO 19119, are in common use across the earth science community, particularly in the geology and geophysics and geo-informatics disciplines. A number of organisations have based data catalogues on these standards. The data catalogues are often built using open source technologies such as Python or PHP. Methods used for harvesting metadata include OpenSearch 1.1 and the Open Archives Protocol for Metadata Harvesting (OAI-PMH).

The Earth Observation Product Heterogeneous EO Missions Accessibility metadata schema (EOP-HMA) provides another metadata schema for earth observation data, and the Meta Data Objects for Linking Environmental Sciences (MOLES) is another emerging standard providing a method of elucidating the relationships between activities, data, instruments, algorithms etc in a semantically consistent and standards compliant manner. Other relevant metadata standards are described in section 6 above.

The use of web services is one of the most common technologies used to access the archive, although using web based forms to select data with an FTP download, or sometimes off-line ordering for larger earth observation and sometimes geology and geophysics datasets, is also frequently used. Overall off-line ordering systems are used by less than 20% of the archive service providers surveyed in the initial on-line survey.

For both archive service providers and dataset producers, the original files and derived products from the data tend to be important formats for delivery of data. Delivering a partial representation of the

data tends to be more important in the data for archive service providers, than in the case of producers of major datasets. Again there is not a significant difference between different disciplines in these trends.

The independent search results in particular indicate a wide variety of portals of different types provided to access different types of data. Of particular note at the current time is the impact of large cyber infrastructure initiatives such as EarthCube and GEOSS in providing data access services.

4.4.3.2 Recommendations

● The existence of widely adopted metadata standards as shown by the surveys conducted both under Task 15.2 and 15.3, and the use of discovery tools based on open standards may provide a more constrained technology platform to interface with in the development of harmonised policies for data access than the quite variable archive system architectures.

● The survey results indicate a general mixture of federated and single catalogues for data discovery, although the use of federated catalogues is more common among the archive users compared to the service providers or dataset producers.

● The surveys also explored the ease of locating and accessing data and there were a number of areas identified for improvement by respondents including:

- Requirement for increased use of multi-archive search tools to make things easier for users
- Whilst generally good levels of discovery level metadata are available, there were indications of a requirement for information about the different processing levels applied to earth observation data, and also the need to be able to retrieve data from a particular sensor across multiple archives.
- Another area where there is a perceived gap in provision is in the usability of various web portals available for discovering and accessing data. For example, some earth observation science portals require some understanding of EO data in order to be able to select the required search criteria, and this therefore makes access to this data difficult for researchers from other disciplines.

● The use of web services to access data also raises issues of user authentication. This is an area which was not covered in detail by the surveys but may require more consideration in work package (WP33).

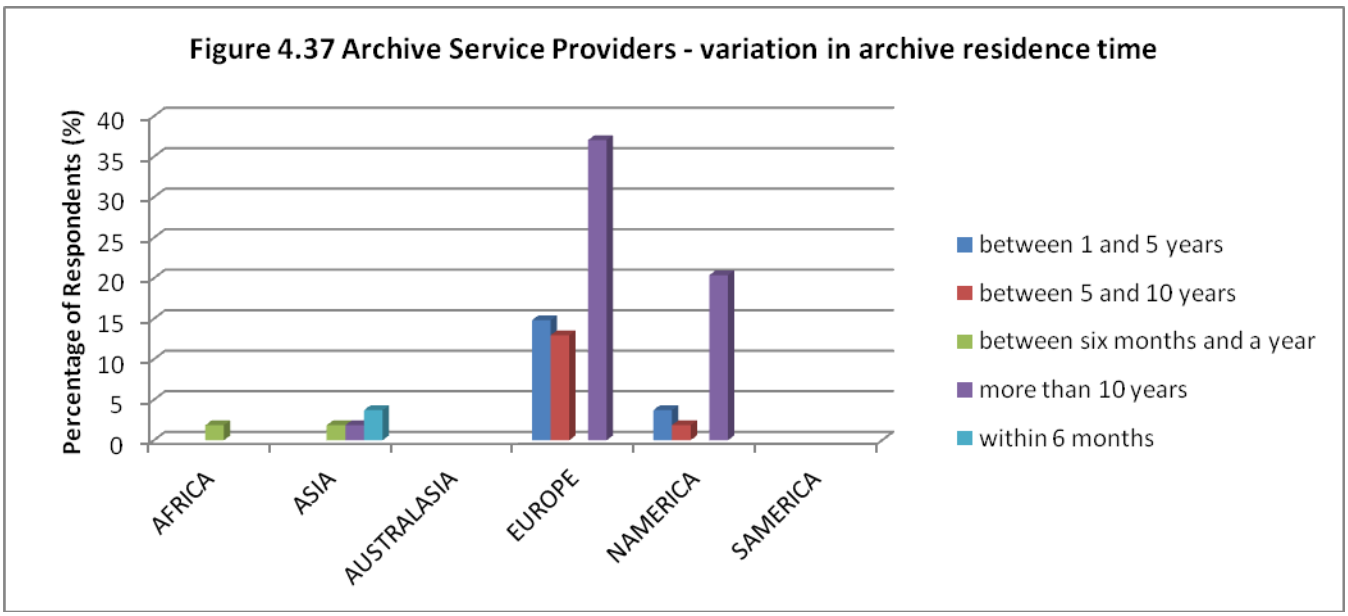
4.5 Preservation Issues

4.5.1 Results from the on-line survey

4.5.1.1 Archive Service Providers

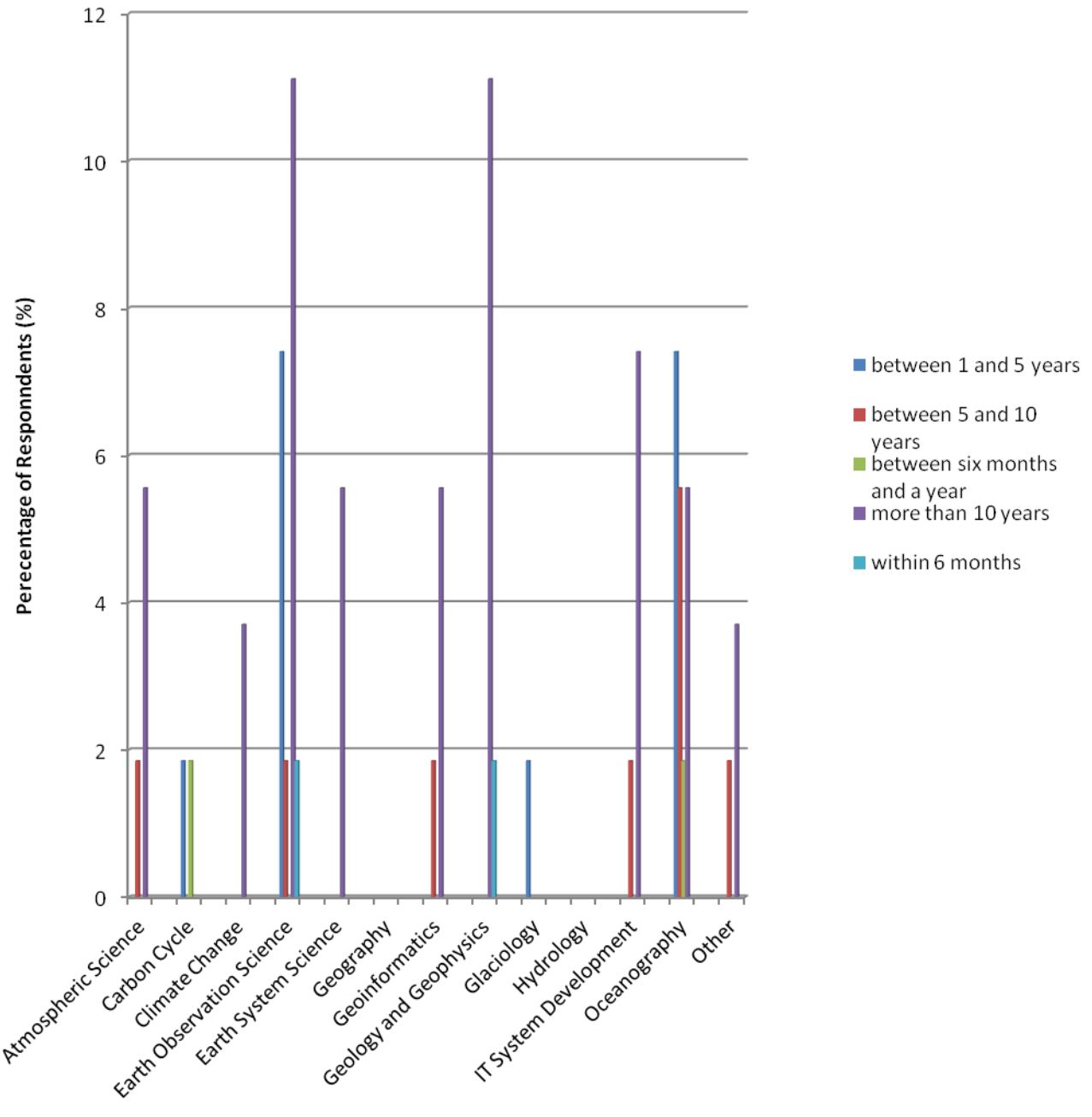
The period for which archive service providers expect their data to remain in its present archive before being migrated to a new archive system is shown in Figure 4.37. The data indicates that the majority of respondents in Europe and North America expect the data to remain in its present archive for over ten

years. However a significant proportion expect the data to remain in its present archive for less than 5 years (c.15%) and between five and ten years (c.13%), and so there is clearly a requirement for tools to support migration of data to a new archive after five and ten year intervals. For archive service providers in Europe and North America there is no indication of needing to migrate to a new archive at less than a five year interval. However the small number of respondents in Africa and Asia do indicate some requirement to migrate to a new archive after one year or less.



Comparing the time data resides in its current archive against discipline (Figure 4.38) indicates that for the three disciplines for which most data is available (Earth Observation, Geology and Geophysics and Oceanography), Earth Observation Science and Geology and Geophysics are characterised by data residing in its present archive mostly for more than ten years. For Earth Observation science the proportion of respondents indicating that data will remain between one and five years in its current archive is also significant, whilst in the Oceanography domain there is a more mixed pattern of residence time in the existing archive.

Figure 4.38. Archive Service Providers - variation of time in archive by discipline



4.5.1.2 Producers of major datasets

Again most of the respondents in this category are from Europe, with c.45% of respondents indicating that the data will remain in the archive for between five and ten years with c.25% indicating over 10 years in the same archive before migration is likely. The trend for the various disciplines also bears this out with Earth Observation Science and to some extent Atmospheric Science showing a strong

tendency to a 5 to 10 year period in the archive before migration, as well as archives which are anticipated to be in place for more than ten years (Figure 4.39).

Figure 4.39. Dataset Producers - variation in archive residence time *(based on 34 responses)*

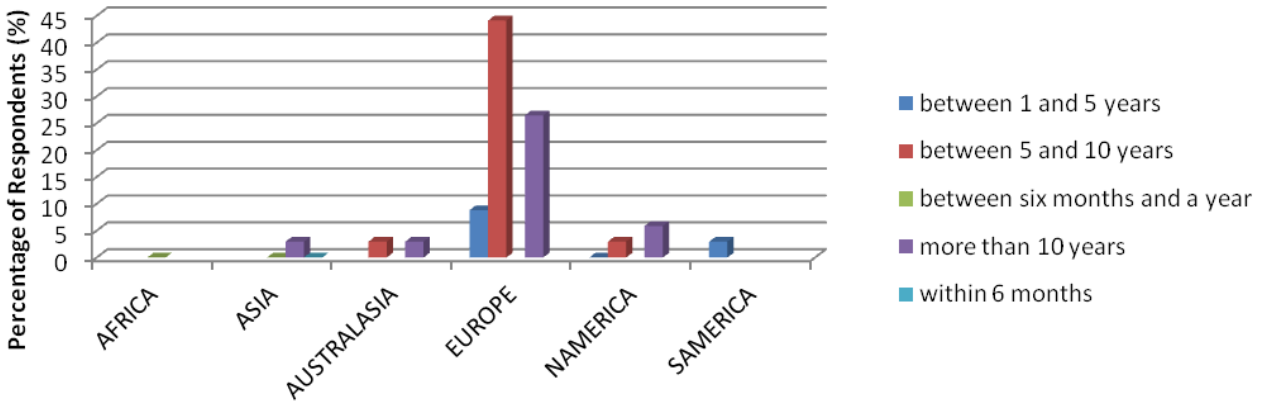
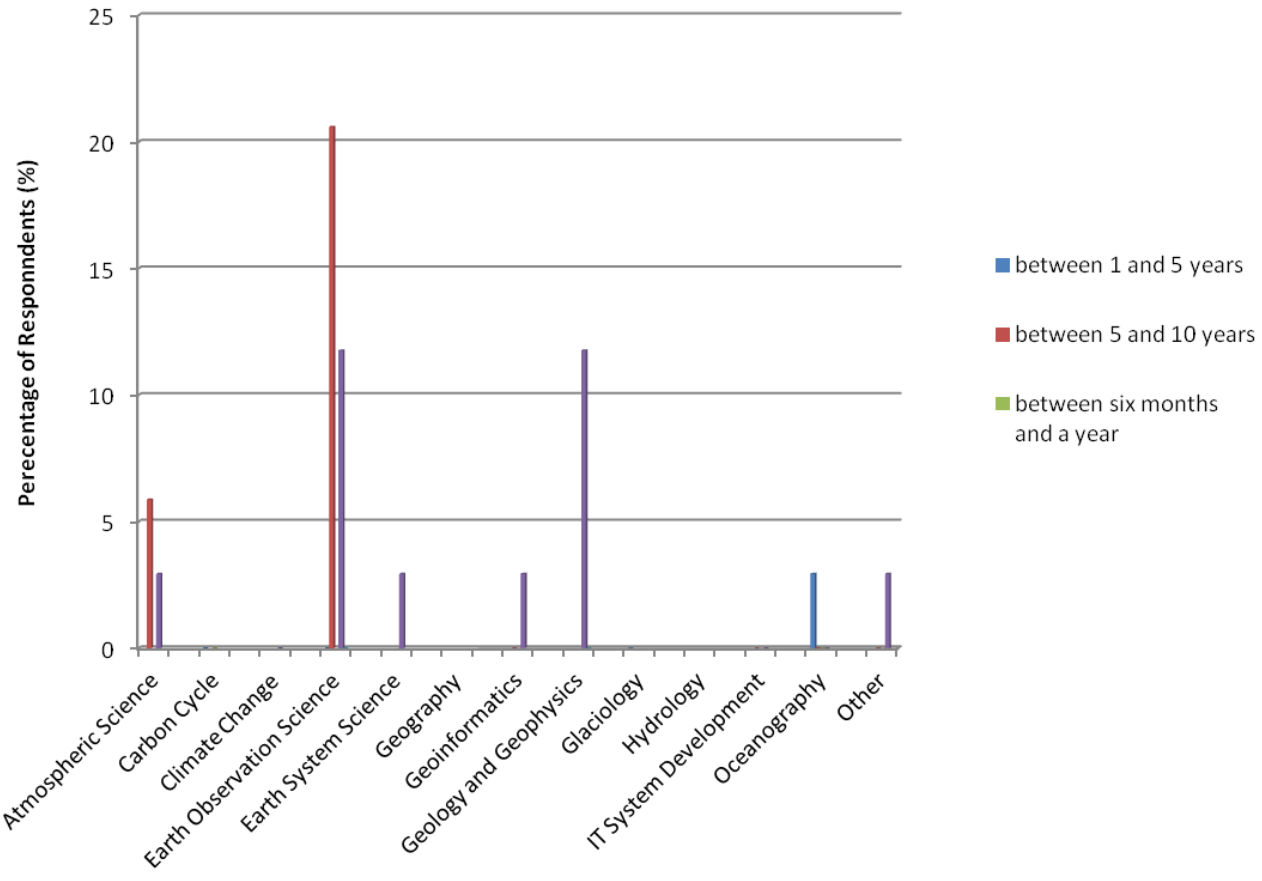


Figure 4.40 Dataset Producers - variation of time in archive by discipline *(based on 34 responses)*



4.5.2 Results from the direct user consultation ("one-to-one" interviews) - preservation issues

There is a tendency for the existence of more well developed data preservation policies and guidelines in the EO community. Where preservation policies are in place they usually follow the OAIS model and LTDP guidelines. Sometimes within other scientific disciplines a particular system may have developed from a smaller project system into a corporate or community resource, and there is a need to create a data preservation policy as the system grows. Also in some instances where a system or archive brings together data from a number of disparate disciplines the development of a coherent harmonised data preservation policy is sometimes hindered by the requirement to address a number of different data types and formats.

The tendency towards either a well developed OAIS and LTDP compliant policy, or the lack of a formal preservation policy could be regarded as an advantageous situation in which to promote SCIDIP-ES to users without a formal strategy in place. However, between these two extremes some organisations have their own locally defined data preservation policy (e.g. the NERC Data Policy), and such locally defined policies may embody a number of the LTDP principles.

The BADC follow the EISCAT (European incoherent scatter radar) guidelines for data preservation which relate to radar data on changes in magnetic flux and ionisation in the atmosphere.

Organisations that do not have a well defined and documented data preservation policy also tend to have a fairly pragmatic approach to preservation issues e.g. for example they may not expect preservation of data to be possible for more than 5 or 7 years, before that data has to be converted to another format or the archive migrated to another system.

The direct user consultation has shed some further light on the procedures used for archive migration in that it is fairly clear that the use of formal migration procedures is more common within the earth observation science domain. However in other earth science disciplines there is certainly recognition of the need to migrate data as software and other technologies develop. Where open source archiving or data access tools are in use there is evidence that organisations are actively upgrading their open source tools, as new versions of the underlying open source software become available.

Organisations who do not claim to have a formal data preservation policy nevertheless are aware of the need for preservation and have objectives to this end. Typical data preservation objectives include:

- Ability to share and exchange data in the long term
- Ability to share and exchange metadata and data products
- Facility to minimise the impact of software hardware upgrades on the operation of the archive and on accessibility/usability of the data it contains

4.5.3 Conclusions - preservation issues

4.5.3.1 Summary

In terms of general trends, archive service providers generally expect data to remain in the archive for at least ten years before being migrated to a new archive. However there are strong indications from the surveys that a number of archive service providers also expect data to only remain in the same archive for between 5 and 10 years, and for the dataset producer category there is a greater tendency for a five to ten year residence time in the archive.

There is no indication of needing to migrate archives at less than a one year interval in Europe, though there are a small number of archive service providers in Asia and Africa who may need to migrate. Overall there are indications that there is a requirement for tools and services to support migration of data to a new archive on potentially a fairly frequent basis (less than ten years in some cases).

Additional data from the direct user consultation exercise supports this trend and indicates that both archive service providers and dataset producers have a fairly pragmatic expectation of how long data can be retained within one archive, and also how long they might expect to need to retain it in total. Therefore, whilst a number of stakeholders would like to preserve their data “Indefinitely”, in reality they probably expect to retain it for ten or possibly 20 years at a maximum.

Well defined data preservation policies are most common within the earth observation community, and these frequently follow the OAIS model and LTDP guidelines. A number of quite high profile earth science organisations do not have a formal documented data preservation policy. However in these cases they are very clear about their objectives in data preservation (section 5.4 above). In other cases an organisation may have a data management strategy which includes guidelines on data preservation.

The survey results for tasks 15.2 and 15.3 indicate a strong awareness within earth science organisations of data preservation issues. Even where a strongly developed data preservation policy is not in place it is clear that a number of organisations are adopting steps such as the inclusion of as much documentation about the data as possible within the archive, and conversion to more software independent formats (e.g. NetCDF) to increase the long term usability of their data.

4.5.3.2 Recommendations

● The use of a number of common software tools across a wide number of disciplines has the implication that if we can handle the data produced by these tools, then we will be able to address some of the needs of a wide variety of earth science disciplines.

● The general lack of well developed visualisation or data interpretation tools integrated within the archiving environment presents an opportunity to develop appropriate tools, for example a visualisation toolkit.

● There is a clear interest in greater interoperability between different data sets, for example researchers wanting to use datasets outside their own discipline, which could be further explored within the analysis to be undertaken in WP33.

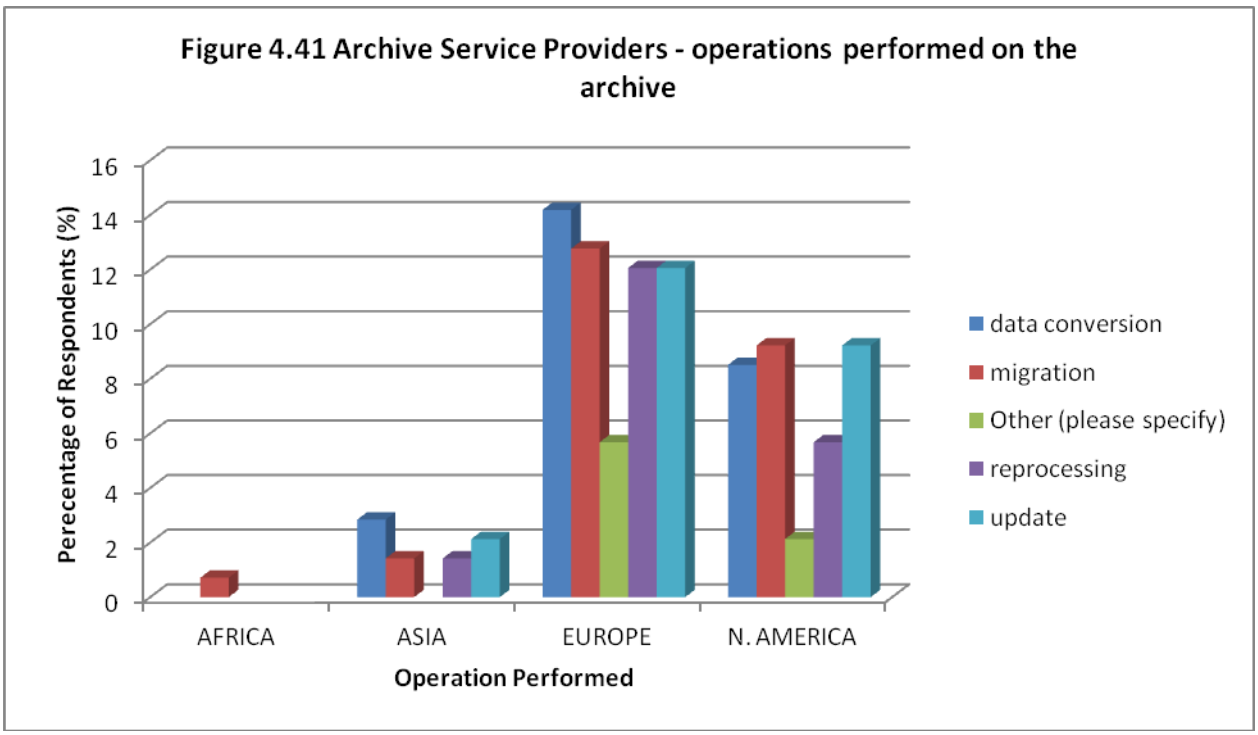
4.6 Processing, Knowledge Extraction, and Management

4.6.1 Results from the on-line survey

4.6.1.1 Archive Service Providers

Operations performed on the archive

The relative importance of data conversion, migration, pre-processing and update operations is fairly similar between Europe and North America with all of these operations being important (Figure 4.41)



Methods of analysis and exploitation of data

In terms of how users analyse and exploit the data, the on-line survey sought to understand what software was used, and how this was integrated with the archiving environment. The trend (Figure 4.42) indicates that within most disciplines a mixture of open source and proprietary software is used with no major differences apparent in the trends for the three disciplines having the most data (earth observation science, geology and geophysics and oceanography).

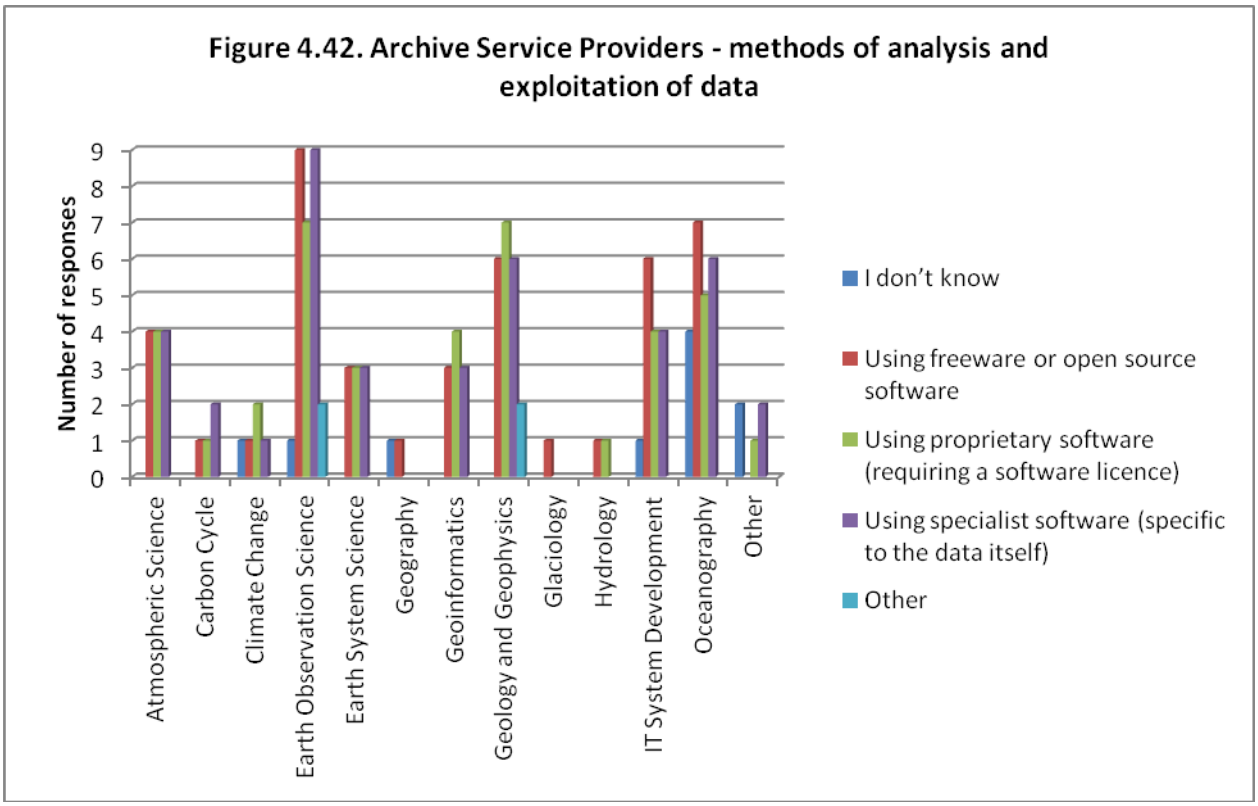


Figure 4.42. Archive Service Providers - methods of analysis and exploitation of data

Table 4.04. Software Used

Software Name
CDAT
ESRI products
MB_system
INTViewer
Fledermaus
CARIS
HIPS
Excel
Open Office Spreadsheet
Basic ERS & Envisat (A)ATSR
Meris Toolbox (BEAM)

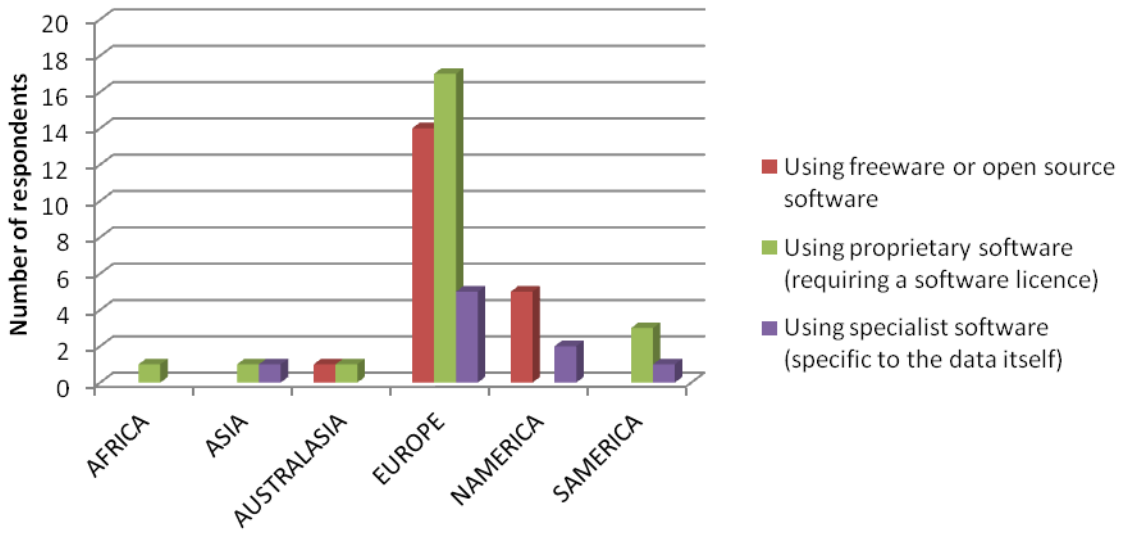
Matlab
IDL
C programming language
Fortran Programming Language
Ferrets
CDO
HDF Viewer
Oracle
ENVI
PCI geomatica
ERDAS
Caraibes
SOCET
PSOKINV(geodetic inversion)
ESA software – NEST,BEAM
Seismic Processing Tools (SeismicUnix, Geocluster, ProMAX)

4.6.1.2 Producers of Major Datasets

Methods of analysis and exploitation of data

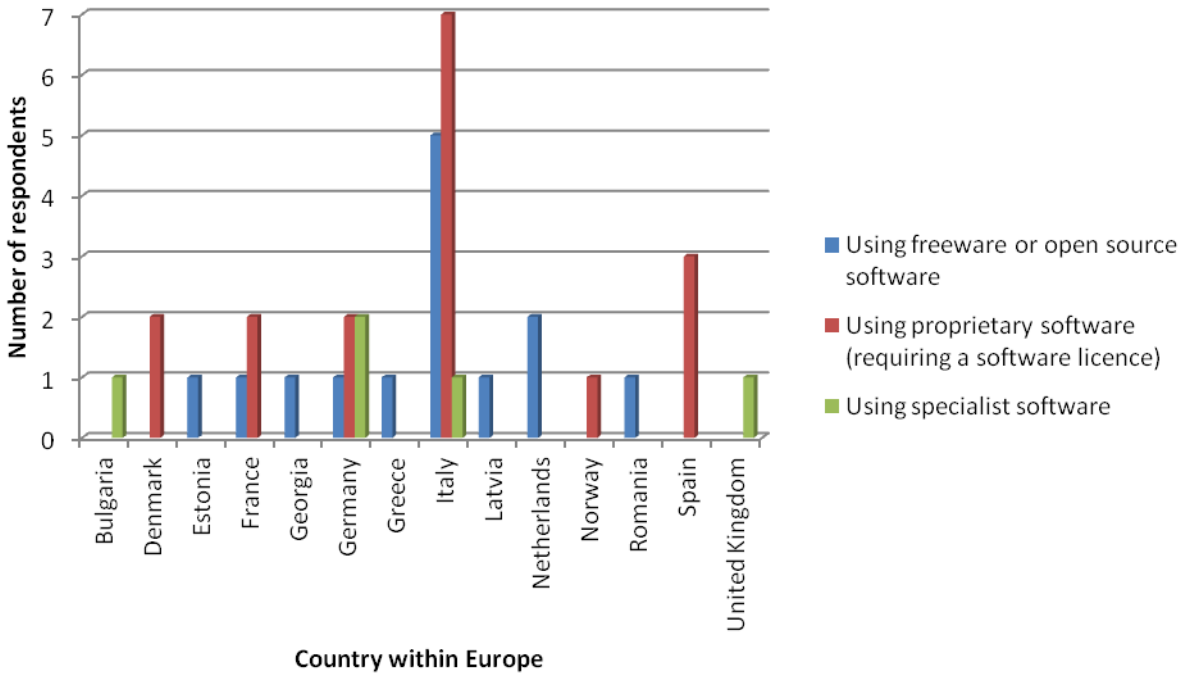
Trends in the type of software used by the "dataset producer" category to interpret the data are shown in Figures 4.43 and 4.44. Again most of the data is for Europe with some for North America. In Europe using both open source software and proprietary software are important (14 and 17 responses respectively). Table 4.04. above indicates some of the proprietary software which is commonly used.

Figure 4.43. Dataset Producers - methods of exploitation and analysis of data *(based on 52 responses)*



Breaking this down to the country level also shows a mixture of proprietary and open source software being used in most countries where there is a useful amount of data.

Figure 4.44. Dataset Producers in Europe: methods of exploitation by country and discipline *(based on 36 responses)*



This trend is also apparent in the Earth Observation Science discipline (Figure 4.45) where most respondents are using proprietary software, although open source software is also an important element. Taking into account just the data for Europe shows a similar trend (Figure 4.46)

Figure 4.45. Dataset Producers - methods of exploitation and analysis *(based on 52 responses)*

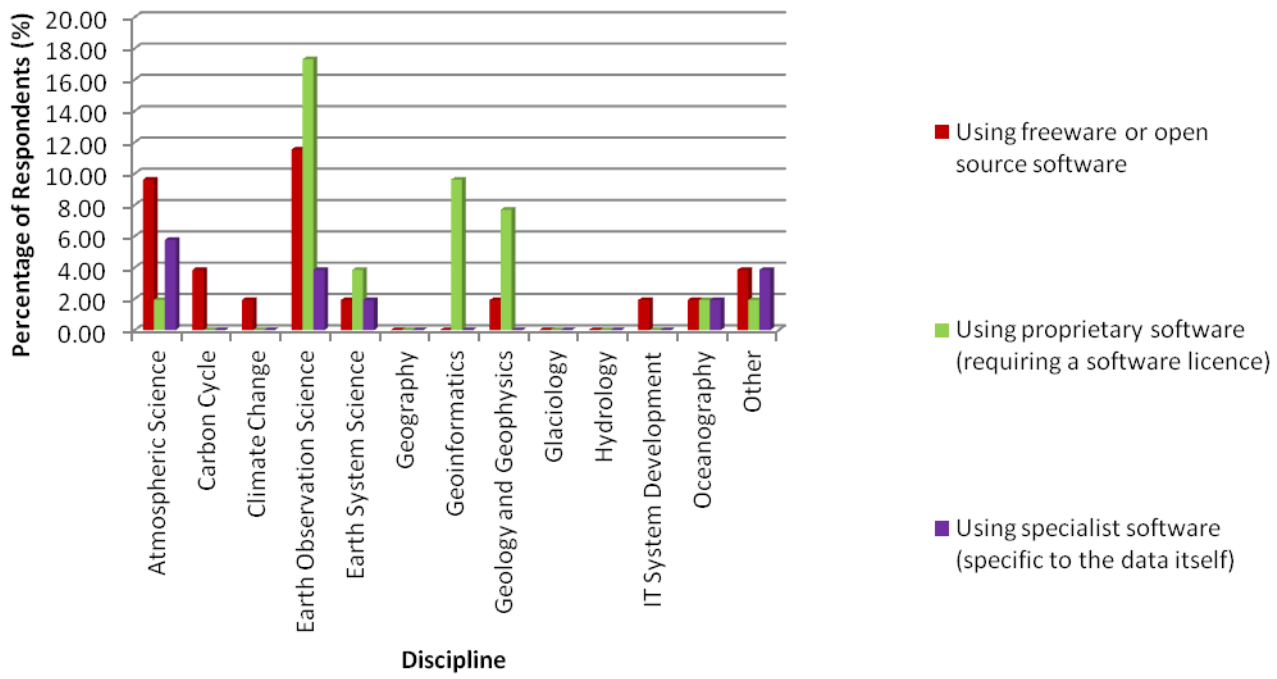
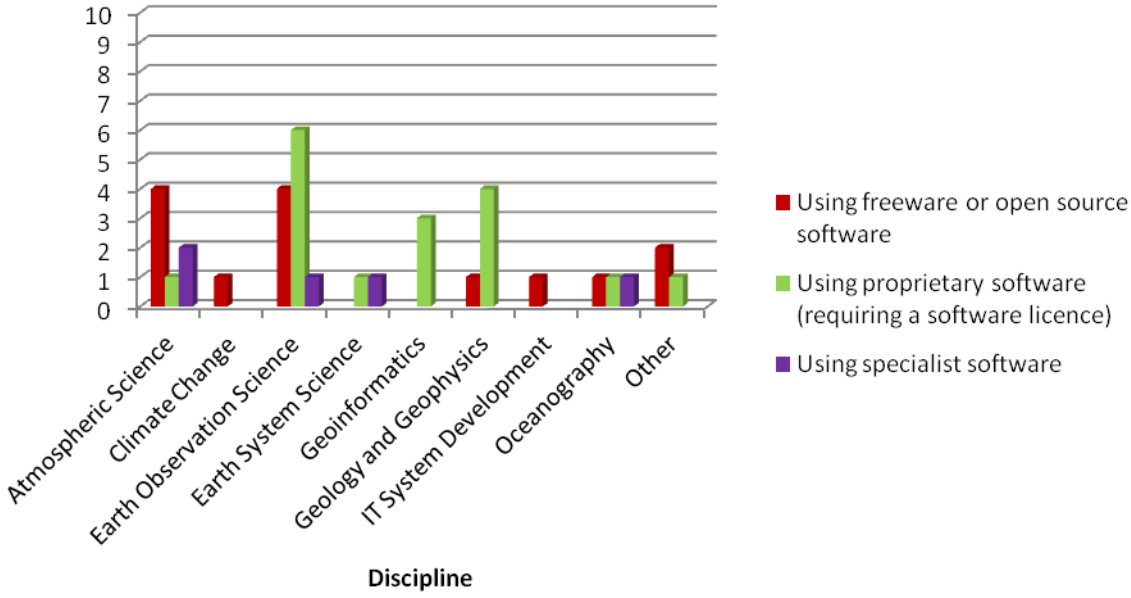


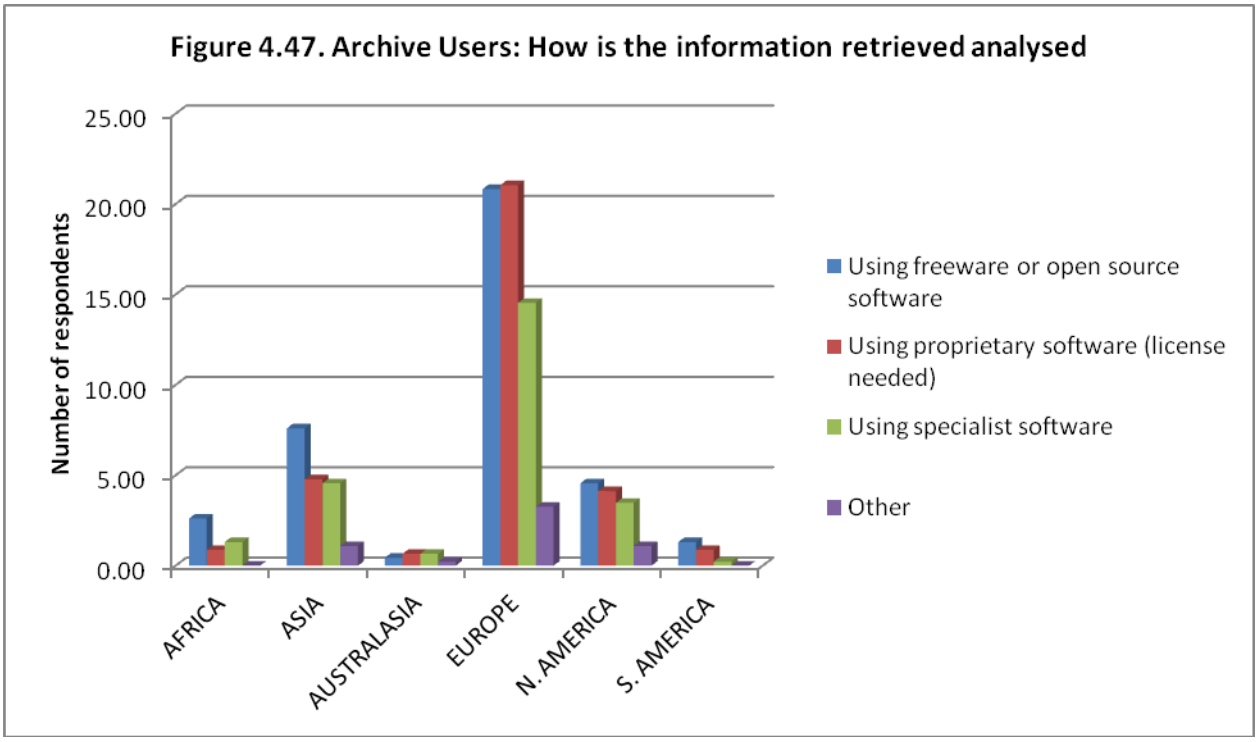
Figure 4.46. Dataset Producers in Europe - methods of exploitation and analysis *(based on 36 responses)*



4.6.1.3 End Users of Archived Data

Software used to analyse the information retrieved

Considering the methods used to analyse the information retrieved (Figure 4.47), the trend suggests that for Europe, North America and Asia where most respondents are located, there is generally an equal distribution between the use of open source, proprietary and specialist software. Comparing this by discipline for Europe (Figure 4.48) also indicates a spread of mechanisms used with open source and proprietary software dominating the Earth Observation, and Geology and Geophysics disciplines.



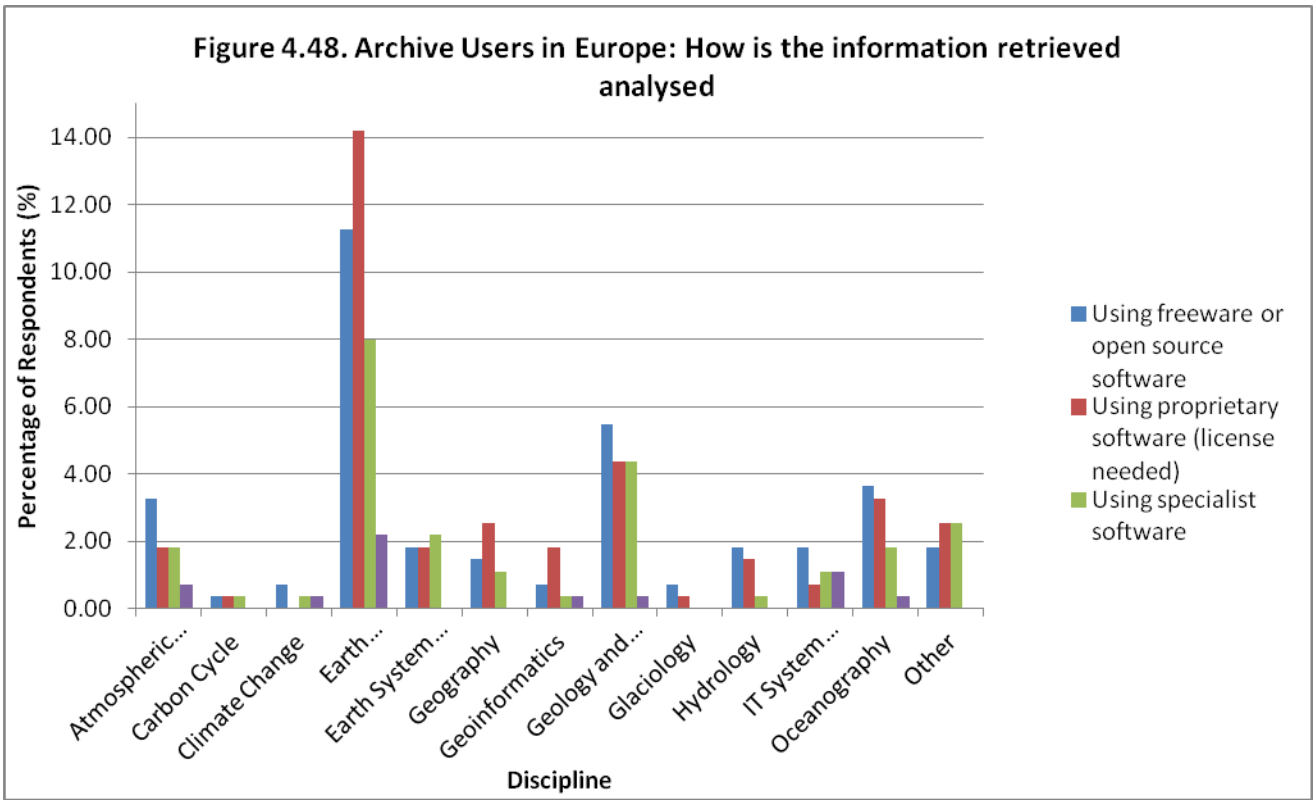


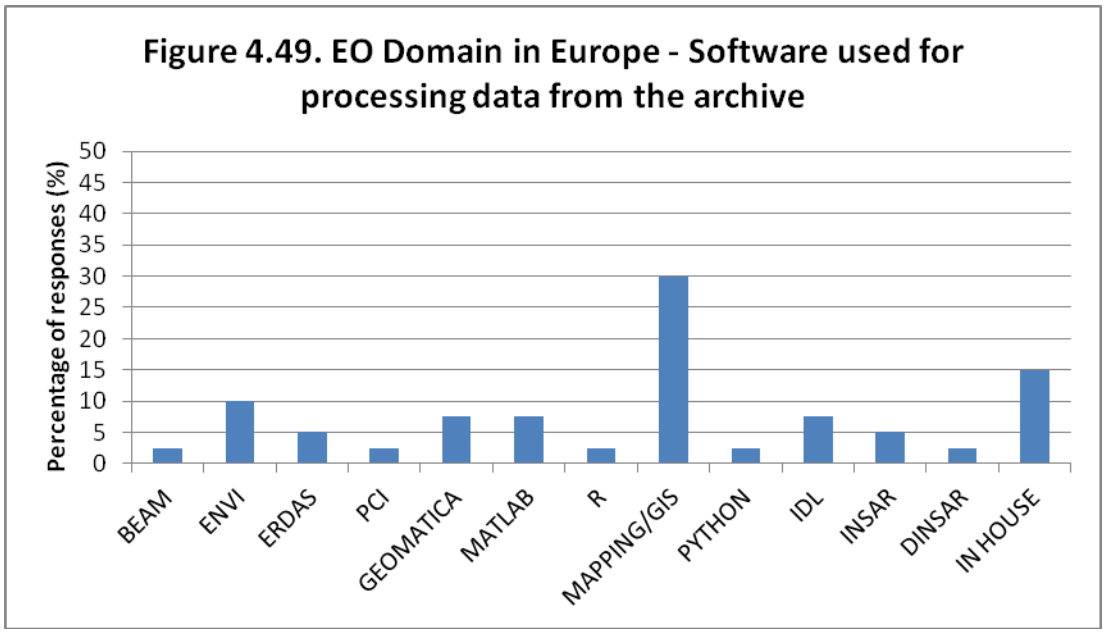
Table 4.05. Software and technologies used for processing archived data

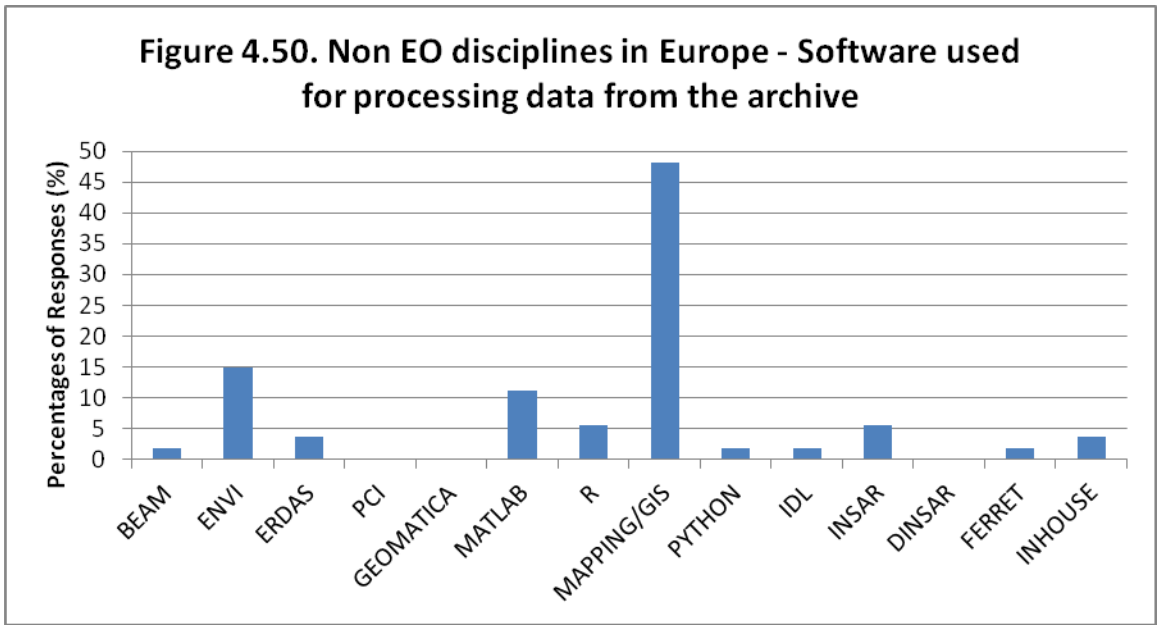
Software/Technology	**Discipline**
OTC image processing software	EO
BEAM	EO, Oceanography
BEST	EO
ESA NEST ARRAY	EO
ENVI	EO, Oceanography, Geology & Geophysics, Geoinformatics, Earth Systems science
ERDAS Imagine	Geoinformatics, Geology & Geophysics
PCI	EO, Geology & Geophysics
Geomatica	EO, Geology & Geophysics
GAMMA	EO
ROI-PAC	EO, Geology & Geophysics, Earth Systems Science

Software/Technology	Discipline
DORIS	EO, Geology & Geophysics
StaMPS	EO, Geology & Geophysics
MATLAB	EO, Oceanography
PYTHON	EO
IDL	EO, Oceanography, Earth Systems Science
INSAR	EO, Geology & Geophysics
DINSAR	EO
LULC	EO
PostGIS	EO
QGIS	EO
gvSIG	EO
ArcGIS	EO
SARSCAPE	EO
R	EO,Oceanography
Google Maps/Google Earth	Oceanography
IDRISI	
FORTRAN	Earth Systems Science

Table 4.05 indicates that many software packages commonly used in the EO domain, are also used across a number of other disciplines (e.g. ENVI, MATLAB, IDL, ArcGIS), this has an implication that if we develop tools and services to support processing of earth observation data, these should also be very relevant to other earth science disciplines.

Figures 4.49 and 4.50 show that similar software tools are used by both the earth observation and the predominant in-site data disciplines (atmospheric science, geology/geophysics, and oceanography), although mapping technologies show a slightly lower relative proportion in the EO discipline, due to some of the other technologies being used slightly more within the EO domain than outside it . There also seems to be a greater tendency to create custom/in-house software solutions in the earth observation domain than in other disciplines.





How is the users software integrated with the archive environment

Following from this, users were asked whether the software they use for their day to day work was integrated with the archiving environment. Here there is a greater degree of integration occurring in Europe than in North America, particularly in Earth Observation science (Figures 4.51 and 4.52). Although most users responding in Europe and North America indicate that the software they use for day to day work is not integrated with the archiving environment.

Figure 4.51. Archive Users in Europe - Is users software integrated with the archiving environment *(based on 244 responses)*

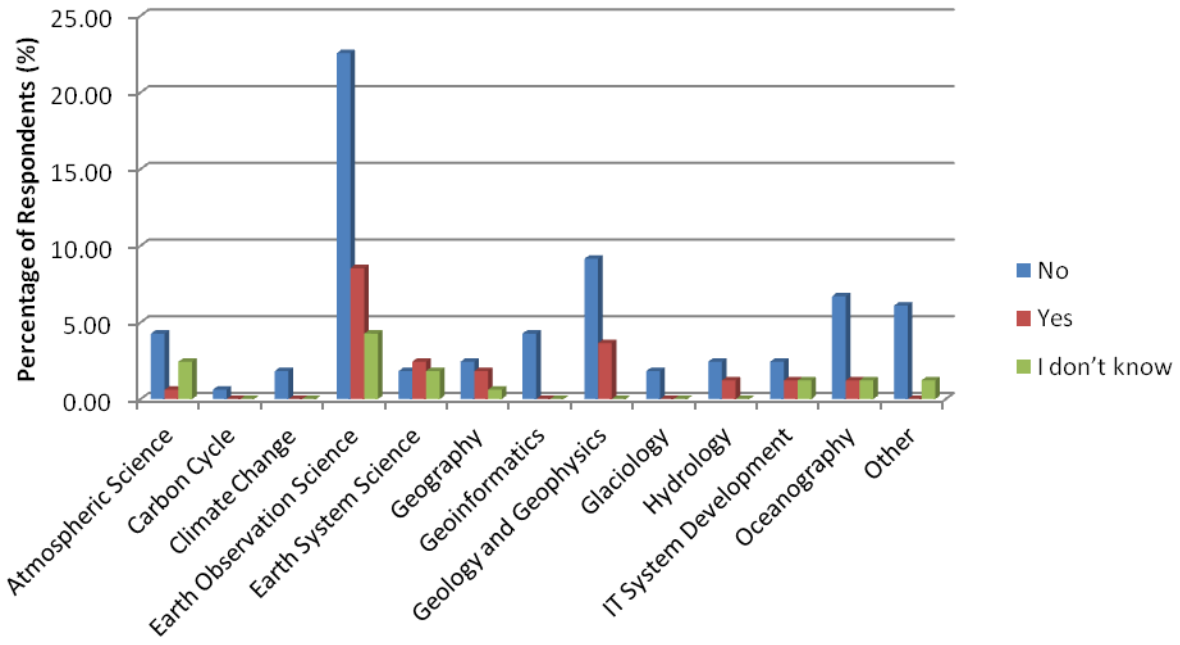
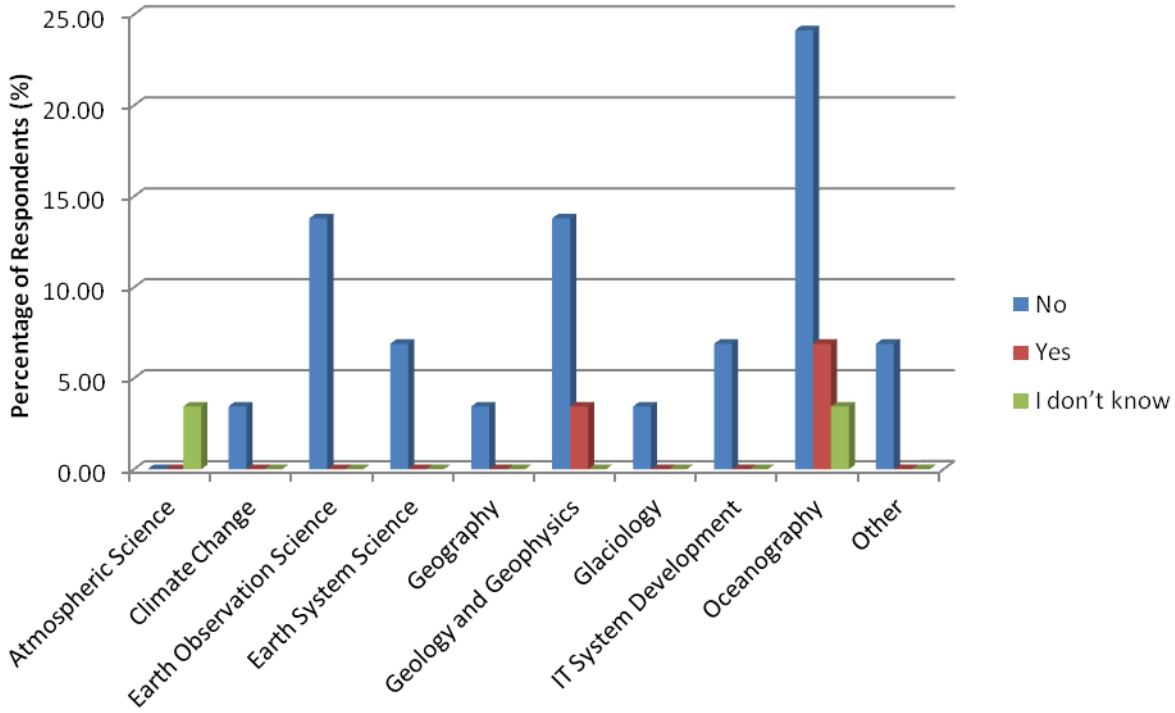


Figure 4.52. Archive Users in North America - Is users software integrated with the archiving environment *(based on 244 responses)*



Mechanisms to check the completeness of the data retrieved from the archive

Following through the workflow of data processing and analysis and submission to the archive, users indicate that there is a general lack of mechanisms to check the consistency and completeness of the data they are retrieving from the archive (Figure 4.53). Comparing the trend within the different disciplines (Figure 4.54) suggests that only in the earth observation and oceanography domains is there some use of such mechanisms.

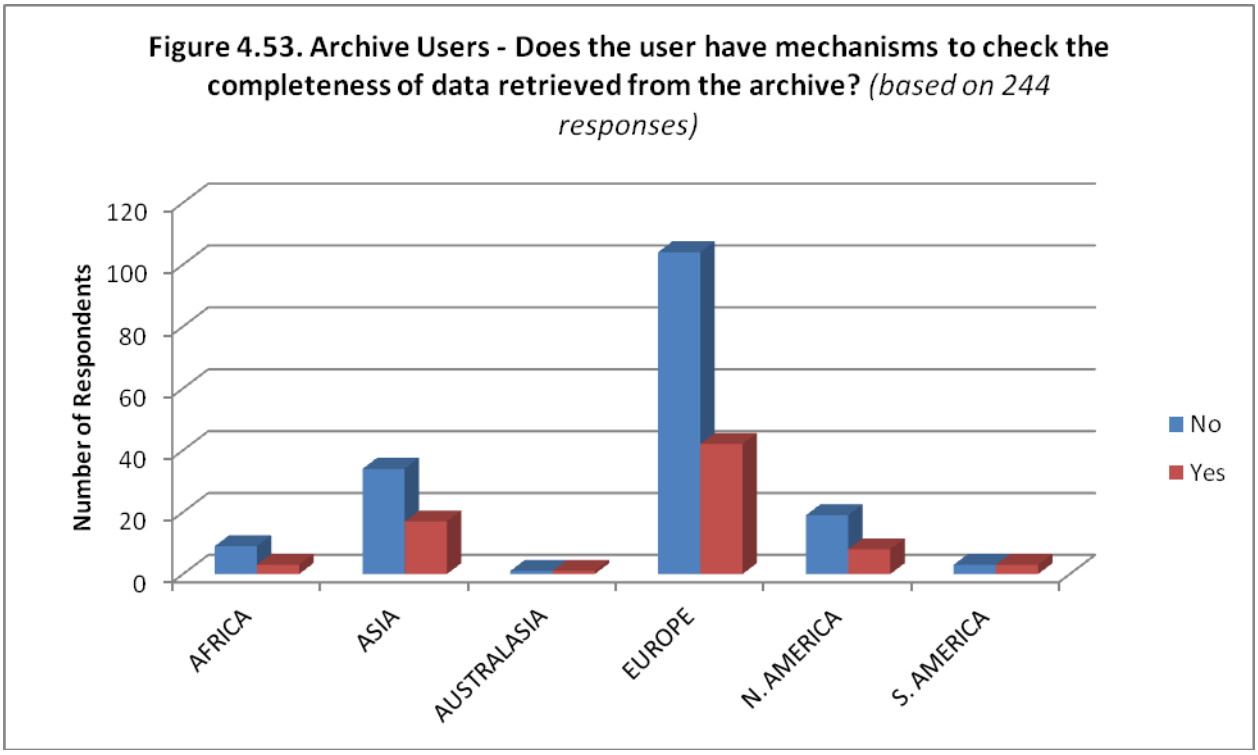
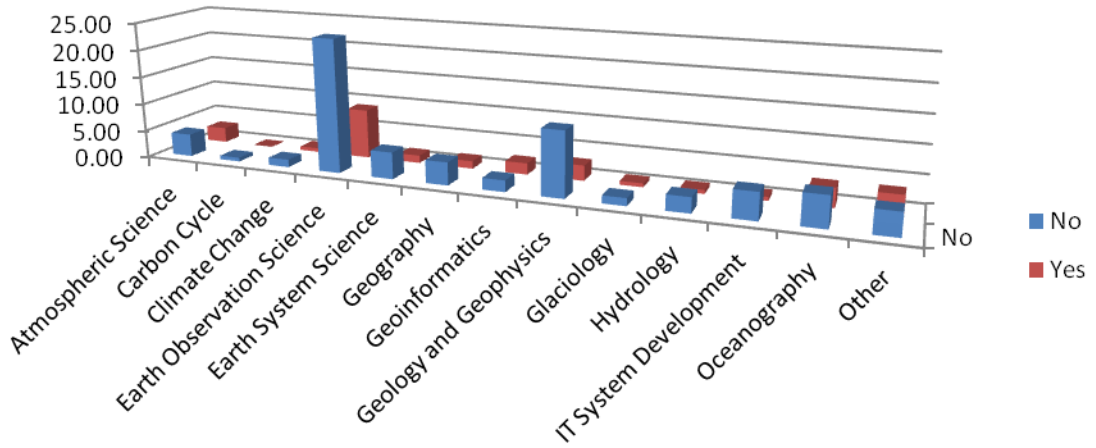


Figure 4.53. Archive Users - Does the user have mechanisms to check the completeness of data retrieved from the archive? *(based on 244 responses)*

Figure 4.54. Archive Users in Europe - Does the user have mechanisms to check the completeness of data retrieved from the archive *(based on 244 responses)*



Interoperability considerations

Respondents were asked whether there was data produced by other disciplines that they would like to have access to. Table 4.06. shows the data that people were interested in using from other disciplines.

Table 4.06 Interest in using data from other disciplines

Respondents Discipline	**Type of data required**	**Source Discipline for required data**
EARTH OBSERVATION	Agronomy and soils data	GEOGRAPHY, GEOLOGY & GEOPHYSICS
	Climate change Data	ATMOSPHERIC SCIENCE
	Ground reference data for disaster areas	GEOGRAPHY
	Meteorological data from local weather stations (digital tabular data)	ATMOSPHERIC SCIENCE
	Population: local statistical and census data (attributes to vector layers)	GEOGRAPHY/ECONOMICS
	Cartography: address data (vector)	GEOGRAPHY

	Hydrology	HYDROLOGY
IT SYSTEM DEVELOPMENT	ASTER GDEM data provided freely over FTP	EARTH OBSERVATION
	Many data on marine environment are not accessible on line and/or they are spread in a number of different databases	OCEANOGRAPHY
ATMOSPHERIC SCIENCE	Biological data	BIOLOGY
	Temperature and Pressure and wind speed and direction from lots of other sites	ATMOSPHERIC SCIENCE, EARTH OBSERVATION
CLIMATE CHANGE	High resolution census data, land use and land cover, elevation models, historical imagery datasets	GEOGRAPHY
GEOLOGY & GEOPHYSICS	Climate change data	EARTH OBSERVATION, ATMOSPHERIC SCIENCE
	Geochemical, biological, thermal, seismological data	GEOLOGY & GEOPHYSICS, BIOLOGY
	MERIS RR reprocessed dataset with MER_RR__2P product type	EARTH OBSERVATION
	INSAR data	
	LIDAR DATA	
	Meteorological data (pressure, temperature, water vapour)	ATMOSPHERIC SCIENCE
	Oceanography: raw SAR data with as much coverage as possible	OCEANOGRAPHY
	glaciology: raw SAR data other global SAR data of geophysical interest	GLACIOLOGY
	Seismology; much of the collected data in various countries are out of reach	GEOLOGY & GEOPHYSICS
	Water vapour maps (e.g. from reanalysis of meteorological data, GPS, radio sonde and satellite imagery), ionosphere TEC maps	ATMOSPHERIC SCIENCE, EARTH OBSERVATION
GEOGRAPHY	Extensive meteorological station networks are available in most countries but access to data is often restricted.	ATMOSPHERIC SCIENCE

OCEANOGRAPHY	Time series data for coastal winds at given locations in NETCDF format	ATMOSPHERIC SCIENCE
	Ice masks for ocean areas with seasonal ice cover, ocean photosynthetically active radiation (PAR) with direct and diffuse components	EARTH OBSERVATION
	SAR, ENVISAR, GRAVIMETRY	
	Southern Ocean ENVISAT-ASAR VV-VH data , High wind speed>15m/s	
	Surface wind in inland seas	ATMOSPHERIC SCIENCE
	SAR data	EARTH OBSERVATION

It is clear that Earth observation data is of interest to a wide variety of other disciplines particularly geology and geophysics and oceanography. Earth observation specialists are also interested in using data from other disciplines particularly from atmospheric science and geography. The geology and geophysics community is interested in using a wide variety of types of data from other disciplines, particularly data from earth observation, atmospheric science and oceanography. There is a strong interest in the oceanography community in using various types of earth observation and atmospheric science data.

In addition to interest from one discipline in using data from another, there is also a strong indication that there is interest in better access to certain types of data within a given discipline, e.g. the geology and geophysics community would like to access seismology data from different countries more easily. There is also an indication of a need for easier access to certain types of data, for example from IT systems specialists in being able to access oceanographic data in a more centralised way.

Common archiving formats

Users were also asked to indicate if there were any common data formats (e.g. images, tabular data etc.) that they tended to archive most frequently. Understanding any patterns in the use of these common formats may assist in developing tools and services which can, for example work, with a particular data format from a number of different disciplines, and assist the development of interoperability between data types and also domains.

Table 4.07 shows the most common data formats reported by respondents. For example, formats such as GEOTIFF and HDF for numerical data, and NetCDF for gridded data are in widespread use across a variety of disciplines, therefore being able to provide tools and services which support archiving these formats will facilitate increased interoperability between data from these disciplines. The survey showed that storing data in tabular

formats such as database files and spreadsheets is particularly common in geology and geophysics, climate change, and oceanography

Table 4.07.Preferred data formats for archiving

Data Format Type	Data Format	Disciplines
Images	GeoTiff	EARTH OBSERVATION EARTH SYSTEM SCIENCE GEOGRAPHY GEOLOGY & GEOPHYSICS OCEANOGRAPHY
	ERDAS Imagine .img files	EARTH OBSERVATION
	Multispectral processed images	EARTH OBSERVATION
	Raster and vector images	EARTH OBSERVATION GEOLOGY & GEOPHYSICS
	BITMAPS	OCEANOGRAPHY
ASCII formats	CSV	EARTH OBSERVATION and many others
	ASCII formats using other separators	OCEANOGRAPHY
	Text data within geodatabases	EARTH OBSERVATION
	XML/GML	EARTH OBSERVATION
	ASCII format grid	GEOLOGY & GEOPHYSICS
	Various text files	OCEANOGRAPHY
Numerical Data	HDF, HDF5	EARTH OBSERVATION ATMOSPHERIC SCIENCE EARTH SYSTEM SCIENCE GEOGRAPHY HYDROLOGY OCEANOGRAPHY

GIS	Shape files	EARTH OBSERVATION, CLIMATE CHANGE GEOLOGY & GEOPHYSICS
	Thematic maps in vector format	EARTH OBSERVATION
Proprietary formats	ENVI .HDR files	EARTH OBSERVATION GEOINFORMATICS
	PCI .pix files	EARTH OBSERVATION
Binary formats	NetCDF	EARTH OBSERVATION CLIMATE CHANGE HYDROLOGY OCEANOGRAPHY
ENVISAT N1	ENVISAT N1	EARTH OBSERVATION
Standard Archive Format (SAFE)	SAFE	EARTH OBSERVATION
Compressed file formats	.zip and .tgz	CLIMATE CHANGE
	Mr SID compressions	
Data in tabular form	tables	CLIMATE CHANGE GEOLOGY & GEOPHYSICS
	Database file formats	GEOLOGY & GEOPHYSICS
	Spreadsheet formats	OCEANOGRAPHY

4.6.2 Results from the direct user consultation ("one-to-one" interviews) - processing, knowledge extraction and management

In most cases the data from the direct user consultation exercise supports the conclusions of the initial on-line survey (section 4) which indicates that for the most part the software used to analyse the data retrieved is not integrated with the archiving environment. There is also an indication that such integration may not always be a high priority requirement, providing good facilities to discover and access the data are in place, since within most domains there is well established interpretation software available.

The facility to undertake visualisation or other manipulation of the data within the archiving environment is quite variable, with a number of systems (e.g. MOIST) providing the facility to plot data on-line, whilst other systems (for example the CEDA archive) provide more advanced facilities for data visualisation. From the respondents interviewed there is a greater tendency for some sort of knowledge extraction facility (e.g. interrogation of the data via a GIS interface, or basic statistics within disciplines outside the earth observation domain.

The surveys indicate a general lack of available tools to check the completeness of what is being retrieved from the archive. A number of archive service providers have some sort of “check sum” tool which compares the number of files, and their content between receipt and download, but there would seem to be a gap in the provision of user friendly tools of this type.

A consistent feature of the one-to-one interview responses was that archive service providers and dataset producers frequently undertake a certain degree of processing to convert raw data into a format (e. NetCDF) which can be both used more easily by the scientists, and also potentially preserved more easily for future use.

4.6.3 Conclusions - processing, knowledge extraction and management

4.6.3.1 Summary

The overall trend is for the software used to analyse data to be quite separate from the software used to access and discover data in the archive. Basic tools, for example for plotting data, are sometimes provided within the software used to access the archive e.g. as used in the MOIST system. Occasionally more complex visualisation tools are provided (e.g. in the case of the CEDA archive) but this is less common.

Processing of the data is undertaken using a range of software tools. One interesting trend emerging from the initial on-line survey is that a number of the software tools used within the earth observation domain are also used across a number of other disciplines (e.g. ENVI, MATLAB, IDL, and ArcGIS).

Overall open source software is used to analyse retrieved data as much as proprietary software in most disciplines, particularly, atmospheric science, earth observation science, and oceanography. However, taking the results just for Europe suggests that proprietary software is used more than open source in a number of disciplines, particularly, earth observation science, geo-informatics, and geoscience.

When considering the level of integration of software used by respondents for working with the data and the associated archiving environment, the trend is for there to be little integration between the two for those respondents from North America. However there appears to be a greater degree of integration within Europe.

The data from the survey indicates that there appears to be a general absence of mechanisms for checking the completeness of the data retrieved from an archive. However in Earth Observation science and also to a lesser extent in the oceanography domain there is some indication of such mechanisms being present. This indicates a gap in the services and toolkits currently available and the requirement to develop additional functionality as part of the SciDIP-ES project.

The results of the initial on-line survey indicate that there is tendency towards a fairly constrained number of file formats being used by different disciplines. As anticipated various image file formats are particularly important, as are tabular and other database format data. NetCDF is a very widely used format across many earth science disciplines (further details are provided in Table 12, in section 5 above).

An important conclusion from the surveys is that there is considerable interest from researchers in using data from outside their own discipline. Those working in earth observation and in geology and geophysics in particular are interested in using data from a number of other disciplines. There is also interest in researchers within a given discipline being able to access other types of data that they may not normally use, or which comes from other geographic areas

4.6.3.2 Recommendations

● There is therefore an implication that tools and services are required to support migration between archives at between five and ten year intervals.

● The web search activities have highlighted a number of very relevant technologies to support data archiving however very few of these available technologies have actually been recorded in the online survey responses or in the one-to-one interviews during the direct user consultation. It would therefore be useful to gain a greater understanding of why a number of the available open source frameworks and technologies are apparently not in wider use, and therefore what further gaps in provision this might indicate.

● Where an archive or system contains data from a number of different disciplines it is evident that the development of a coherent harmonised data preservation policy is sometimes hindered by the requirement to address a number of different types of data.

● The survey results suggest that a number of earth science organisations do not have a formally defined data preservation policy, though such organisations are often implementing elements of such a policy, and this is an advantageous situation in which to propose harmonised policies.

4.7 Metadata and semantics

In this section the term *semantic model,* or just *model,* will be used as a generalization of the terms *metadata*, *metadata standards, schemas, ontologies*, etc. By aggregating the responses to the survey 44 distinct models have been identified. Figure 4.55 shows the most popular semantic models and the number of responses in the survey that referenced each of them. A number of other models were identified by a single respondent and these have been omitted from the figure for reasons of clarity, however, all of the models identified in the survey are shown below.

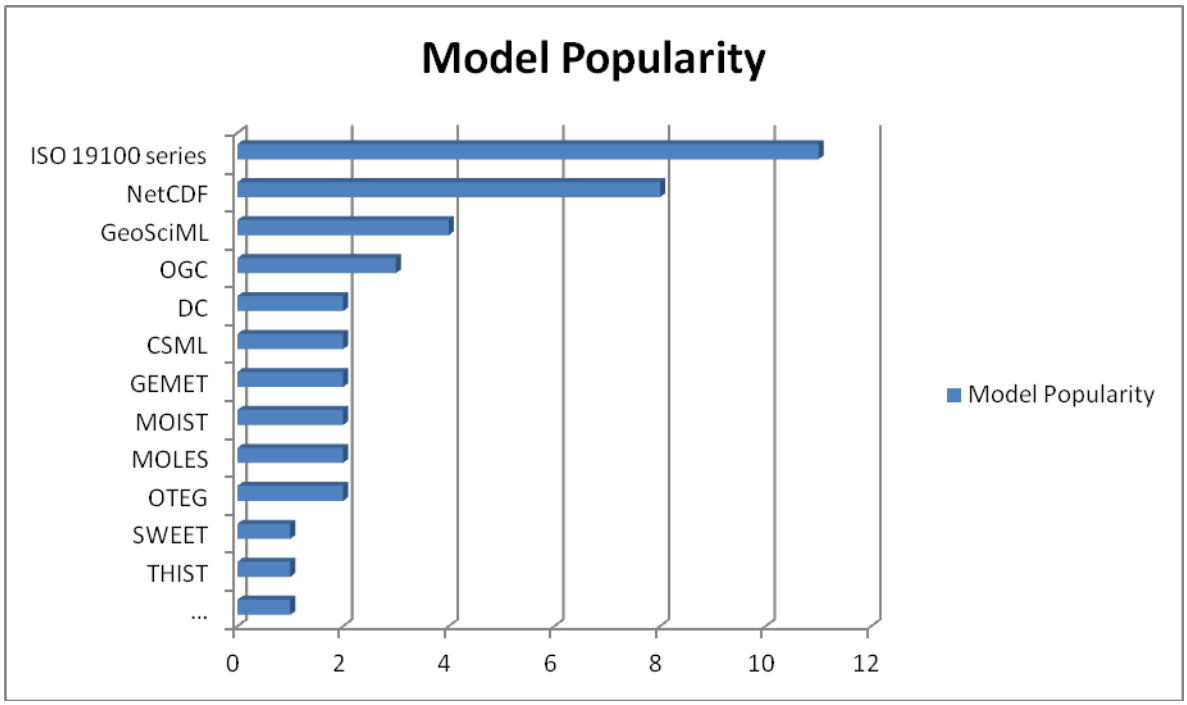


Figure 4.55 Popularity of Semantic Models

Figure 4.55 shows the number of semantic models identified by each respondent. These responses have then been grouped according to their organisation. This chart does not omit duplicates where the same model has been identified by more than one participant from different organisations (duplicates from respondents at the same institution have been removed). The intention has only been to summarize how many responses were received from each organization (and an estimation of the number models that are used in an organization).

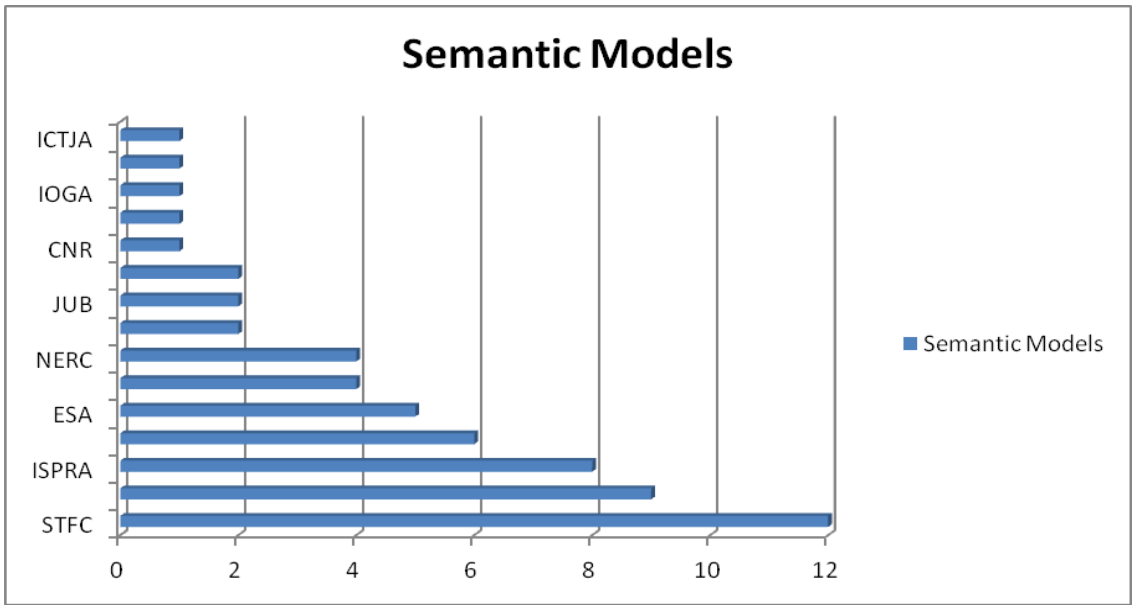


Figure 4.56 Semantic models reported per organization

Figure 4.57 (left) shows the availability of these models in various **formats** (XML, UML, RDF/S, OWL, other). Most of these models (approximately 60%) exist in XML format. Figure 4.57 (right) shows the results regarding their **usage.** It follows that most of them are used for querying and exchanging data. Only about 20% are used for natively storing data.

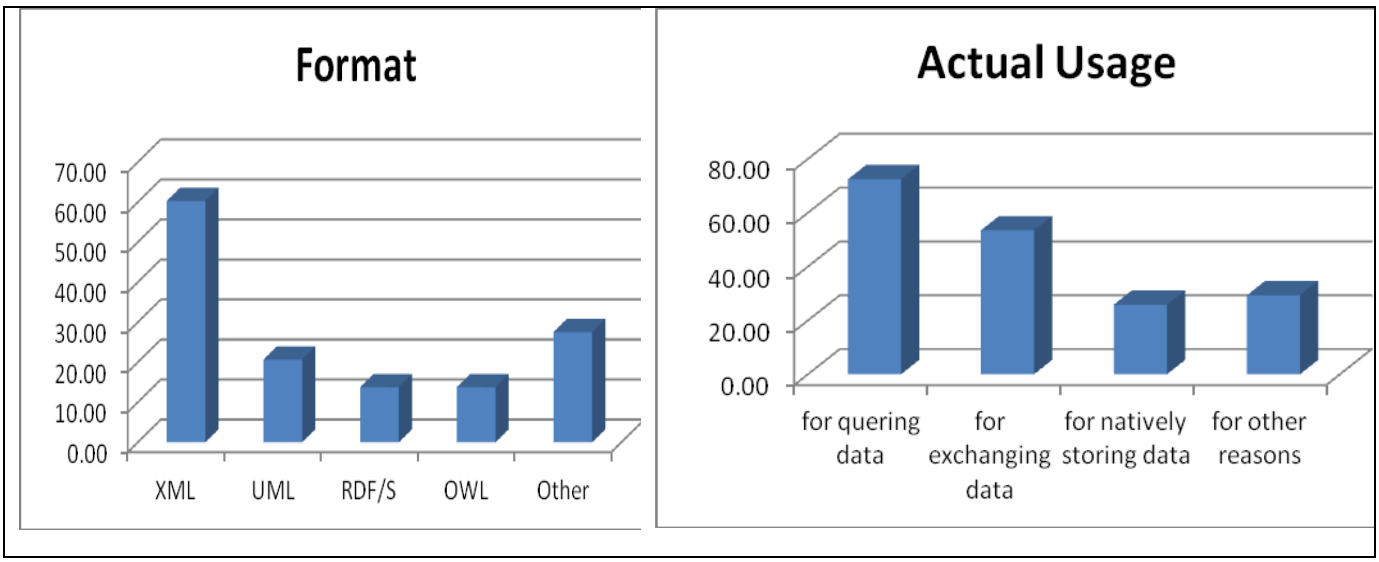


Figure 4.57 Format and Actual Usage Information

4.7.1 State of the Art

This section uses the analysis of the survey responses described in the earlier sections to provide an overview of the current state of the art in metadata, semantics and ontologies currently in use in the earth sciences. In particular we perform a relatively high level analysis of the survey results in order to identify which categories of metadata schemas, semantics and ontologies are most widely used and for what purpose. (A deeper analysis of the survey results will be carried out as part of task T33.2).

4.7.2 Existing Semantic Models

The survey shows that each community uses semantic models that fit for their own needs. For example, a volcanologist may use a specific ontology, designed for that purpose, for modelling the observations and measurement of several volcanoes.

The following table summarizes the semantic models that are currently in use by the earth science communities. The table below is an aggregation of the responses received in the survey. The first column gives the official name that is used to identify the model; the second one is a brief description of the model while the last column describes the formats in which these models are available. The semantics models are sorted alphabetically (based on the identified semantic model name).

Name	Description	Format			
		XML	RDF/ OWL	UML	Other
ABC	A simple core ontology incorporating time, place and events		✓		
Air Quality Metadata	Simple metadata used to describe several atmospheric datasets.				✓
Arc Marine Metadata	A data model used to describe the structure and semantics of marine information.		✓		
CF (Conventions, Standard Names, NetCDF)	A controlled vocabulary for Climate and Forecast metadata, described through the NetCDF API	✓			✓
CGI Vocabularies	A set of GeoScience vocabularies and ontologies designed to facilitate information exchange		✓		
CIDOC CRM	Provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation		✓		
Comité national Télédétection	Simple metadata about available spatial images for Madagascar				✓
Common Data Index	A fine-grained index to individual data measurements. (Designed upon the ISO 19115 standard)	✓			
CSML	Climate Science Modelling Language is a data model for encoding climate, atmospheric and oceanographic data	✓			
CUAHSI HIS	A Hydrologic Ontology that was designed to support the discovery of time-series data collected at a fixed point, including physical, chemical and biological measurements		✓		
Dublin Core	Dublin Core is a set of vocabulary terms that can be exploited for generally describing resources, and can be used for multiple purposes	✓	✓		
DublinCore	A Dublin Core lightweight profile used for		✓		

Name	Description	Format			
		XML	RDF/ OWL	UML	Other
Lite4G	Geospatial data				
Directory Interchange Format (DIF)	A descriptive and standardized format for exchanging information about scientific data sets	✓			
DLR Ontology	An ontology (in owl) used to describe geospatial data		✓		
DIMAP	A format designed to describe geographic data. Initially it was specially designed for mage data, however it can also handle vector data	✓			
DOLCE	The Descriptive Ontology for Linguistic and Cognitive Engineering is a general upper level ontology		✓		
EOP-HMA	Standard for the ESA Heterogeneous Missions Accessibility	✓			
EarthResourceML	An exchange format for mineral resources information	✓		✓	
GCMD	Global Change Master Directory is a directory of Earth Science datasets and related services & tools.				✓
GEMET	The General Multilingual Environmental Thesaurus aims to define a general terminology for the field of environment	✓	✓		
GeoSciML Vocabularies	Vocabularies of the Geoscience Mark-up Language	✓	✓		
GML 3.1.1	Application schema for Earth Observation products	✓			
GML Coverage	Application schema for modelling coverages (digital geospatial information representing space/time varying phenomena)	✓			
INSPIRE datasets	ISO 19115 and ISO 19119 compliant metadata	✓		✓	
ISO 19100 series	A series of standards that support data management, acquiring, processing,	✓		✓	

Name	Description	Format			
		XML	RDF/ OWL	UML	Other
	analyzing, accessing, presenting and transferring data between different users/systems, for geographic information				
ISO 19115	Defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and the temporal schema, spatial reference, and distribution of digital geographic data.	✓		✓	
ISO 19115-part 2	Extends the existing geographic metadata standard by defining the schema required for describing imagery and gridded data	✓		✓	
ISO 19119	Identifies and defines the architecture patterns for service interfaces used for geographic information	✓		✓	
ISO 19139	Defines the Geographic Metadata XML encoding (gmd). It provides XML schemas that enhances interoperability by providing a common specification for describing, validating and exchanging metadata about geographic datasets, dataset series, etc	✓		✓	
ISO 19156	Defines a conceptual schema for observations and for features involved in sampling when making observations	✓		✓	
IDEC Thesaurus	A thesaurus for geospatial concepts	✓			
MOIST	Multidisciplinary Oceanic Information SysTem aims at hosting multidisciplinary data and metadata	✓			
MOLES	The Metadata Objects for Linking Environmental Sciences models has been developed to encode the relationships between the tools used to obtain data, the activities which organized their use, and the dataset produced	✓		✓	
MyOCean Catalogue	A catalogue of several oceanographic products				✓

Name	Description	Format			
		XML	RDF/ OWL	UML	Other
NOA Ontology	Used in TELEIOS project		✓		
Observations and Measurements Standard	Observations and Measurements is an OGC standard, which defines XML schemas for observations, and for features involved in sampling when making observations	✓			
OpenSearch	Collection of simple formats for sharing search results	✓			
OTEG Ontology	Ontology for describing Earth observation datasets		✓		
POLDER/ PARASOL schema	Schema for archiving data in the POLDER/PARASOL missions	✓			
PREMIS Ontology	The PREMIS Data Dictionary for Preservation Metadata is a digital preservation standard based on the OAIS reference model		✓		
SensorML	SensorML is an OGC standard. It provides standard models and an XML encoding for describing sensors and measurement processes.	✓		✓	
SimpleLithology	SimpleLithology is a CGI vocabulary that is used for describing several Lithology concepts		✓		
SWEET	A set of ontologies that can be used for providing a common semantic framework for representing Earth Science data, information and knowledge		✓		
THIST	Italian Thesaurus of Earth Resources has been resulted from integrating the terminological database of APAT database and the thesaurus published by CNR since 1997				✓

The following table describes the fields of applicability for the semantic models described earlier. The categories are derived from the earth science disciplines.

Name	Top Level	Earth Science Field				
		Atmosphere	Biosphere	Geology / GeoInformatics	Cryosphere	Oceonagraphy
ABC	✓					
Air Quality Metadata		✓				
Arc Marine Ontology						✓
CF (Conventions, Standard Names, NetCDF)		✓				
CGI Vocabularies		✓		✓		
CIDOC CRM	✓					
Comité national Télédétection			✓	✓		
Common Data Index		✓		✓		✓
CSML		✓				✓
CUAHSI HIS						✓
Dublin Core	✓					
DublinCore Lite4G				✓		
Directory Interchange Format (DIF)						
DLR Ontology				✓		
DIMAP				✓		
DOLCE	✓					
EOP-HMA		✓		✓		✓
EarthResourceML				✓		
GCMD		✓	✓	✓	✓	✓
GEMET	✓					
GeoSciML Vocabularies				✓		

GML 3.1.1				✓		
GML Coverage				✓		
INSPIRE datasets				✓		
ISO 19100 series				✓		
ISO 19115				✓		
ISO 19115-part 2				✓		
ISO 19119				✓		
ISO 19139				✓		
ISO 19156				✓		
IDEC Thesaurus				✓		
MOIST						✓
MOLES	✓					
MyOCean Catalogue						✓
NOA Ontology						
Observations and Measurements Standard				✓		
OpenSearch	✓					
OTEG Ontology				✓		
POLDER/ PARASOL schema				✓		
PREMIS Ontology	✓					
SensorML				✓		
SimpleLithology				✓		
SWEET		✓		✓	✓	✓
THIST				✓		

4.7.3 The Main Semantic models within SCIDIP-ES

This section lists those semantic models that are mainly used by the partners of SCIDIP-ES. The following table shows the models that are in use by each SCIDIP-ES partner and if that model was reported either through the questionnaire or by some other method. As these models appear to be the most important further information for each is provided in Annex E.

Semantic Model	Used by SCIDIP-ES partner	Is a Questionnaire response	Provided by SCIDIP-ES partner
ISO 19100	CNES, DLR, GIM, ISPRA, NERC, STFC	✓	✓
CF Conventions	INGV, STFC	✓	✓
CGI Vocabularies and Ontologies	ISPRA	✓	✓
OGC	CNES, DLR, ESA, JUB, STFC	✓	
CSML	STFC	✓	✓
GEMET	GIM, ISPRA	✓	✓
MOIST	INGV	✓	
MOLES	STFC	✓	✓
OTEG	ESA	✓	✓
SWEET		✓	
THIST	ISPRA	✓	
CIDOC CRM	ESA, FORTH		✓
Dublin Core	ESA, INGV	✓	✓
VOID			✓
SKOS	NERC, UTV	✓	✓
ABC		✓	

4.7.4 Conclusions – metadata, semantics and ontologies

4.7.4.1 Summary

The results of the survey can be summarised as follows; the information gathering activities yielded 44 distinct models, which we categorized according to their main role and broad domain (i.e. top level, Atmosphere, Biosphere, Geology/Geoinformatics, Cryoshere, Oceanography). As regards their purpose, their main use is for querying and exchanging data (only 20% of these models are used for natively storing data) and 60% of these models are available in XML. Most of these deal with metadata for geology/geoinformatics which is capturing information about the solid earth and the study of the capture, qualification, classification, storage, processing and production of spatial information of earth.

4.7.4.2 Recommendations

There are a range of key metadata, semantic and ontology models, their purpose and format has been documented in section 4 of this report, and provides a good basis for a more in depth analysis of the survey data as part of WP33.

Annex A. References

[1] Carl Lagoze and Jane Hunter, The ABC Ontology and Model, DC-2001: International Conference on Dublin Core and Metadata Applications, October 2011, Tokyo, Japan.

[2] Brian Eaton, Jonathan Gregory, Bob Drach, Karl Taylor, Steve Hankin, NetCDF Climate and Forecast (CF) metadata conventions, Version 1.1, January 2008.

[3] Andrew Woolf, Bryan Lawrence, Roy Lowry, Kerstin Kleese van Dam, Ray Cramer, Marta Gutierrez, Siva Kondapalli, Sue Latham, Kevin O'Neill, Ag Stephens, Climate Science Modelling Language: Standards-based Markup for Metocean Data, In procs of the 85th meeting of American Meteorological Society, 2005

[4] Alliance Permanent Access to the Records of Science in Europe Network deliverable on WP24-Provenance and Authenticity, D24.1 Report on Authenticity and Plan for Interoperable Authenticity Evaluation System.

[5] Alliance Permanent Access to the Records of Science in Europe Network deliverable on WP22-Identifiers and citability, Survey on Persistent Identifiers.

[6] Maria Theodoridou, Yannis Tzitzikas, Martin Doerr, Yannis Marketakis, Valantis Melessanakis, Modeling and Querying Provenance by extending CIDOC CRM, Distributed and Parallel Databases, 27(2), Springer, 2010.

[7] Robert G. Raskin, Michael J. Pan, Knowledge Representation in the Semantic Web for Earth and Environmental Terminology (SWEET), Computers & Geosciences 31 (9), p. 1119-1125, November 2005.

[8] Spiros Ventouras, Bryan Lawrence and Simon Cox, The MOLES-v3 Information Model, EGU General Assembly Vienna, 2010

[9] International organization for standardization: The CIDOC conceptual reference model (2006). Ref. No. ISO 21127:2006. http://cidoc.ics.forth.gr/

[10] R. Yang, X. Deng, M. Kafatos, C. Wang, X.S. Wang. An XML-based distributed metadata server (DIMES) supporting Earth science metadata. In Proceedings of the 13th International Conference on Scientific and Statistical Database Management (SSDBM'2001), Virginia, USA, July 2001.

[11] R.G. Raskin, M.J. Pan. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Computer and Geosciences Journal, vol. 31 (9), pp. 1119-1125, Elsevier, November 2005.

[12] R. Ramachandran, S. Graves, H. Conover, K. Moe. Earth Science Markup Language (ESML):: a solution for scientific data-application interoperability problem. Computers & Geosciences Journal, vol. 30 (1), pp. 117-124, Elsevier, February 2004.

[13] S.W. Houlding. XML—An Opportunity for <Meaningful> Data Standards in Geosciences. Computers & Geosciences Journal, vol. 27 (7), pp. 839-849, Elsevier, August 2001.

[14] V. Parekh, J.P. Gwo, T. Finin. Ontology Based Semantic Metadata for Geoscience Data. In Proceedings of the International Conference on Information Knowledge Engineering (IKE'2004), Las Vegas, USA, June 2004.

[15] A.Y. Chen, A. Donnellan, D. McLeod, G. Fox, J. Parker, J. Rundle, L. Grant, M. Pierce, M. Gould, S. Chung, S. Gao. Interoperability and Semantics for Heterogeneous Earthquake Science Data. In Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Florida, October 2003.

[16] M.B. Bluementhal, J. del Corral, M. Bell, E. Grover-Kopec. An Ontological Approach to Geoscience Dataset Cataloging. In AAAI 2006 Fall Symposium on Semantic Web for Collaborative Knowledge Acquisition. Washington DC, October 2006.

[17] K. Asch, B. Brodaric, J.L. Laxton, F. Robida. An International Initiative for Data Harmonization in Geology. In Proceedings of the 10th EC GI & GIS Workshop, Warsaw, Poland, June 2004.

[18] J. Helly, H. Staudigel, A. Koppers. Scalable Models of Data Sharing in Earth Sciences. Geochemistry Geophysics Geosystems Journal, vol. 4(1), 2003.

[19] F. Reitsma, J. Albrecht. Modeling With the Semantic Web in the Geosciences. Intelligent Systems vol. 20(2), pp. 86-88, IEEE, 2005.

[20] L. Bernard, U. Einspanier, S. Haubrock, S. Hubner, E. Klien, W. Kuhn, E. Lessing, M. Lutz, U. Visser. Ontology-Based Discovery and Retrieval of Geographic Information in Spatial Data Infrastructures. Geotechnologien Science Report, vol. 4, pp. 15-29, Citeseer, 2004.

[21] K. Lin, B. Ludasher. A System for Semantic Integration of Geologic Maps via Ontologies. In Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Florida, October 2003.

Annex B. Figures and Tables

B.1. List of Figures

B.2. List of Tables

Annex C. Results from the independent search activity

Table A below gives an overview of the resources available for the various topics identified above that are currently in use within the earth science domain.

Table B provides information on other relevant initiatives that may provide useful resources for the purposes of the SciDIP-ES project but which fall outside the earth science domain.

Table A.

Resource Name	*Domain*	*URL*	*Topic*
Long term preservation of earth observation space data European LTDP common guidelines	Long-term data preservation	http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_DraftV2.pdf	Data preservation policies
EUMETSAT	Meteorology	http://earth.esa.int/gscb/ltdp/presentations/12.ltdp_approach_eumetsat.pdf	Data preservation policies
Implementation Plan for the National Geological and Geophysical Data Preservation Program	Geology	http://datapreservation.usgs.gov/docs/2006DataPreservation.pdf	Data preservation policies

Resource Name	*Domain*	*URL*	*Topic*
Toward Implementation of the Global Earth Observation System of Systems **Data Sharing Principles**	Non-domain specific	http://www.spacelaw.olemiss.edu/jsl/pdfs/articles/jsl-35-i-foreword.pdf	Data preservation policies
MATLAB	Cross – domain including earth science	http://www.mathworks.co.uk/products/matlab/	Data preservation technologies
R	Cross – domain including earth science	http://www.r-project.org/	Data preservation technologies
Ruby (Ferret)	Cross – domain including earth science	http://www.ruby-lang.org/en/	Data preservation technologies
Perl	Cross – domain including earth science		Data preservation technologies
Python	Cross – domain including earth science	http://www.python.org/	Data preservation technologies
ENVI	Earth observation	http://www.ittvis.com	Data preservation technologies

Resource Name	Domain	URL	Topic
Erdas Imagine	Cross – domain including earth science	http://www.erdas.com/	Data preservation technologies
EAST	Cross – domain including earth science	http://public.ccsds.org/publications/archive/644x0b2.pdf	Data preservation technologies
DFDL	Cross – domain including earth science	http://www.ogf.org/Public_Comment_Docs/Documents/2010-03/draft-gwdrp-dfdl-core-v1.0.pdf	Data preservation technologies
DRB	Cross – domain including earth science	GAEL-P243-DOC-001-01-01_DRB_API_handbook.pdf	Data preservation technologies
GeoNetwork OpenSource	Cross – domain including earth science	http://geonetwork-opensource.org	Data preservation technologies
National Geospatial Digital Archive (NGDA)	Geospatial data	http://www.ngda.org/	Data preservation technologies
JASMINE and CEMS	Cross domain including earth science	http://www.stfc.ac.uk/e-Science/news+and+events/38663.aspx	Data preservation technologies
ULISSE project	Earth Observation	http://www.ulisse-space.eu/	Data preservation, Data discovery
Kalideos remote sensing	Earth Observation	http://kalideos.cnes.fr/spip.php?article12	Data preservation,

Resource Name	Domain	URL	Topic
reference databases			Data discovery
NOESIS	Earth science	http://dev.gomrc.org/GoMRC-Noesis/	Data discovery
NERC Data Grid	Earth science	http://ndg.badc.rl.ac.uk/	Data discovery
Global Change Master Directory (GCMD)	Earth systems	http://gcmd.nasa.gov	Data discovery
Geo-Seas project	Marine geoscience	http://www.geo-seas.eu	Data discovery, Data access
SeaDataNet	Oceanography	http://seadatanet.org	Data discovery, Data access
MyOcean	Oceanography	http://www.myocean.eu/	Data discovery, Data access
ICARE	Atmosphere	http://www.icare.univ-lille1.fr/	Data discovery, Data access
SATMOS	Meteorology	http://www.satmos.meteo.fr/cgi-bin/ht.pl?page=index3.html&lang=en	Data discovery, Data access
Archimer	Oceanography	http://archimer.ifremer.fr/default.jsp?la=en	Data discovery, Data access

Resource Name	Domain	URL	Topic
SITools2	Scientific data	http://sitools2.sourceforge.net/	Data discovery, Data access
EOWEB-NG	Earth Observation	https://centaurus.caf.dlr.de:8443/eoweb-ng/template/default/welcome/entryPage.vm	Data Discovery, Data Access, Metadata
EOLI	Earth Observation	http://earth.esa.int/EOLi/EOLi.html	Data Discovery, Data Access, Metadata
EOC Geoservice	Earth Observation	https://geoservice.dlr.de/catalogue	Data Discovery, Data Access, Metadata
INSPIRE Geoportal	Spatial Data (various categories)	http://inspire-geoportal.ec.europa.eu/discovery/	Data Discovery
GEO Portal	Spatial Data (various categories)	http://www.geoportal.org/web/guest/geo_home	Data Discovery
CEOCAT CCRS Earth Observation	Earth Observation	http://ceocat.ccrs.nrcan.gc.ca/	Data discovery
USGS National Map Seamless Server	Earth Observation	http://seamless.usgs.gov/	Data discovery, Data access

Resource Name	Domain	URL	Topic
Euro GEOSS	Earth Science	http://www.eurogeoss.eu/about/default.aspx	Data Access
NASA Geocover	Earth observation	http://earthexplorer.usgs.gov/	Data discovery, Data access
USGS Earth Explorer	Spatial Data (various categories)	http://earthexplorer.usgs.gov/	Data discovery, Data access
JRC Community image data portal	Earth Observation	http://cidportal.jrc.ec.europa.eu/imagearchive/main/	Data discovery, Data access
EUMETSAT Product Navigator	Earth Observation	http://www.eumetsat.int/Home/Main/Access_to_Data/ProductNavigator/index.htm	Data discovery, Data access
Digital Globe Image Finder	Earth Observation	http://browse.digitalglobe.com/imagefinder/main.jsp?	Data discovery, Data access
ISDC Information Systems and Data Center	Geology and Geophysics	http://isdc.gfz-potsdam.de/	Data discovery, Data access
GENESI-DEC	Earth Observation	http://www.genesi-dec.eu/	Data discover, Data access, Data mangement
Heterogeneous Missions	Earth Observation,	http://earth.esa.int/hma	Data discover, Data access,

Resource Name	Domain	URL	Topic
Accessibility (HMA)	Geoinformation		Data mangement
USGS - Hazards Data Distribution System (HDDS)	Earth Observation, Geoscience, Environmental Science	http://hdds.usgs.gov/hdds/	Data discover, Data access, Data mangement
Geoland2	Earth observation/Geoinformatics	http://www.gmes-geoland.info/	Data access
Sextant	Oceanography	http://www.ifremer.fr/sextant/en/web/guest/catalogue	Data access
Cross-Border SDI (Spatial Data Infrastructure) Project	Geo-informatics	http://www.thecarbonproject.com/Projects/crossborder.php	Data access
JERICO (Joint European Research Infrastructure network for Coastal Observatories)	Biosphere/ Oceanography/ Hydrology	http://www.jerico-fp7.eu/	Data access
MERCURY	Earth science	http://mercury.ornl.gov/ http://www.semantic-web-journal.net/sites/default/files/swj245.pdf	Data access
Earth System	Earth systems	http://www.earthsystemgrid.org	Data access

Resource Name	Domain	URL	Topic
Grid (ESGF)			
GEOBRAIN	Earth Observation	http://geobrain.laits.gmu.edu/	Data access
ISO19110	Non-domain specific	http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39965	Metadata
ISO19115	Non-domain specific	http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44361 http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229	Metadata
ISO19119	Non-domain specific	http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39890	Metadata
GS-Soil Metadata profile	Soil science	https://www.sbg.ac.at/zgis/gssoil/webdocs/Deliv/GS%20SOIL_D3.4_best_practice_guideline_metadata.pdf	Metadata
INSPIRE Metadata Implementing Rule	Geoinfomatics	http://inspire.jrc.ec.europa.eu/index.cfm/pageid/101	Metadata
MMI Project (Marine Metadata Interoperability Project)	Oceanography	http://marinemetadata.org/ http://marinemetadata.org/references/marineprofile19115	Metadata
Meta-T Project	Oceanography	http://marinemetadata.org/community/teams/metat/	Metadata
INSPIRE data specifications	Geoinfomatics	http://inspire.jrc.ec.europa.eu/index.cfm/pageid/101	Metadata
NetCDF CF Conventions	Oceanography/Atmosphere/Earth	http://cf-pcmdi.llnl.gov/	Metadata

Resource Name	Domain	URL	Topic
	Systems		
SensorML	Geology & Geophysics	http://www.opengeospatial.org/standards/sensorml	Metadata
Earth Observation Metadata profile of the OGC Observations and Measurements	Earth Observation	https://portal.opengeospatial.org/modules/admin/license_agreement.php?suppressHeaders=0&access_license_id=3&target=http://portal.opengeospatial.org/files/%3fartifact_id=31065	Metadata
FGDC Metadata standard	Non-domain specific	http://www.fgdc.gov/metadata/geospatial-metadata-standards	Metadata
Dublin Core Metadata Element Set	Cross-domain	http://dublincore.org/documents/dces/	Metadata
Ecological Modelling Language (EML)	Ecology	http://knb.ecoinformatics.org/software/eml/	Metadata
GeoMS	Earth Observation	http://avdc.gsfc.nasa.gov/index.php?site=1178067684	Metadata
EOP-HMA	Earth observation	http://bp.schemas.opengis.net/06-080r2/hma/1.0/hma.xsd	Metadata
Darwin Core	Biodiversity/informatics	http://rs.tdwg.org/dwc/	Metadata

Resource Name	Domain	URL	Topic
PREMIS Data Dictionary for Preservation Metadata	Cross domain	http://www.loc.gov/standards/premis/index.html	Metadata
XML Formatted Data Unit – XFDU	Earth Observation	http://sindbad.gsfc.nasa.gov/xfdu/	Metadata, Data exchange formats
Encoded Archival Description	Cross domain	http://www.loc.gov/ead/	Metadata, Data Exchange Formats
Geography mark-up language (GML)	Geography and earth sciences	http://www.opengeospatial.org/standards/gml	Metadata, Data Exchange Formats
GeoSciML	Earth sciences	www.geosciml.org/	Metadata, Data Exchange Formats
GEneral Multilingual Environmental Thesaurus – GEMET	Earth Observation	http://www.eionet.europa.eu/gemet	Metadata, Semantics
ESA OTEG	Earth Observation	http://gmesdata.esa.int/OTE/navigateInfoDomain	Metadata, Semantics
Geonames	Geography	http://www.geonames.org/ontology/documentation.html	Metadata, Semantics
Earth Science	Earth Science	http://esml.itsc.uah.edu/	Semantics

Resource Name	Domain	URL	Topic
Semantic Markup Language (ESML)			
Semantic Web for Earth and Environmental Terminology (SWEET)	Earth Systems	http://sweet.jpl.nasa.gov	Semantics
Earth Sciences Semantics Portal (ESSP)	Earth Science	http://d1sweb.dataone.utk.edu/	Semantics
EarthCube	Non-domain specific	http://earthcube.ning.com/ http://www.nsf.gov/geo/earthcube/index.jsp	Semantics, Metadata,Data Management
Rasdaman	Earth Science	http://www.rasdaman.org	Data Management
EarthServer	Earth Science	http://www.earthserver.eu/	Data Management
GEON	Earth Science	http://geongrid.org/	Data management
DBCP (Data Buoy Cooperation Panel)	Atmosphere/Oceanography	http://www.jcommops.org/dbcp/	Data Management
DBCP /Argos active off-line assessment of	Atmosphere/Oceanography	http://www.jcommops.org/doc/satcom/argos/Argos-GTS-sub-system-ref-guide.pdf	Data Management

Resource Name	Domain	URL	Topic
data quality			
CNPDI (Canadian Network for Polar Data Infrastructure)	Glaciology	http://cnpdi.ca/	Data Management
IPYDIS (International Polar Year Data and Information Service)	Glaciology/Atmosphere	http://ipydis.org/index.html	Data Management
MOIST	Geology	http://moist.rm.ingv.it	Data management
ESWUA	Atmosphere	http://www.ewua.ingv.it	Data management
NEPTUNE	Geology	http://www.neptunecanada.ca/	Data management
Earth System Modeling Framework (ESMF)	Earth systems	http://www.earthsystemmodeling.org/	Data management
INSPIRE Transformation Network Service	Geoinfomatics	http://inspire.jrc.ec.europa.eu/index.cfm/pageid/5	Data Processing
INSPIRE	Geoinfomatics	http://inspire.jrc.ec.europa.eu/index.cfm/pageid/5	Data

Resource Name	Domain	URL	Topic
Invoke Service			processing
Web Processing Service (WPS)	Geoinformatics	http://www.opengeospatial.org/standards/wps	Data processing
Open Virtualization Format (OVF)	Non-domain specific	http://www.dmtf.org/standards/ovf	Data processing
ESDI-HUMBOLT	Geoinformatics	http://www.esdi-humboldt.eu	Data processing
Earth System Modeling Framework (ESMF)	Earth Systems/ Oceanography/ Hydrology/Atmosphere	http://www.earthsystemmodeling.org	Data processing
Open Modelling Interface (OpenMI)	Non-domain specific	http://www.openmi.org	Data processing
SEAMLESS Association	Biosphere/Earth Systems	http://www.seamlessassociation.org/	Data processing
Kepler Project	Non-domain specific	https://kepler-project.org/	Data processing
Business Processes Execution Language (BPEL)	Non-domain specific	http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-bpel/ws-bpel.pdf	Data processing
KEEP FP7	Non-domain	http://www.keep-project.eu/ezpub2/index.php	Data

Resource Name	Domain	URL	Topic
project	specific		processing
CORINE	Land Cover	http://www.eea.europa.eu/publications/COR0-landcover	Knowledge Extraction
KEO (KNOWLEDGE-CENTRED EARTH OBSERVATION SYSTEM)	Earth Observation	http://keo-karisma.esrin.esa.int/keo-home/Welcome.html	Knowledge extraction
MARS (Monitoring Agricultural ResourceS)	Biosphere/Global vegetation	http://www.marsop.info	Knowledge extraction
PRAIS portal : Performance review and assessment of implementation system	Desertification	http://www.unccd-prais.com/	Knowledge extraction
SSE Service Support environment	Earth Observation	http://services.eoportal.org/	Knowledge extraction
Scape (scalable preservation environments)	Domain independent	http://www.scape-project.eu/about/project	Other relevant initiatives (Europe)
EUDAT	Scientific data – not specific to earth sciences	http://www.eudat.eu/	Other relevant initiatives (Europe)

The domain names used in this table are taken from the domain definition diagram

Table B.

Resource Name	Domain	URL	Topic
Archives de France	Non-domain specific	http://www.archivesdefrance.culture.gouv.fr/gerer/records-management-et-collecte/	Data Management
CINES		http://www.cines.fr/?lang=en	Data Management
Challenges of the Digital Era for film heritage institutions	Art (Cinematics)	http://ec.europa.eu/information_society/newsroom/cf/itemdetail.cfm?item_id=7765	Data Preservation
BEST		http://logiciels.cnes.fr/BEST/EN/best.htm	Data processing
ENSURE	Health, finace and non-science areas	http://ensure-fp7-plone.fe.up.pt/site	Other relevant initiatives (Europe)
OpenAIRE (Open Access Infrastructure for Research in Europe)	Non-domain specific	http://www.openaire.eu/	Other relevant initiatives (Europe)
GUIDELINES FOR THE PRESERVATION OF DIGITAL	Non-domain specific	http://unesdoc.unesco.org/images/0013/001300/130071e.pdf	Digital preservation policies

Resource Name	Domain	URL	Topic
HERITAGE			
Data Preservation at LEP (CERN)	Physics	http://www.dphep.org/sites/site_dphep/content/e20/e36625/e46703/dplta-lep.pdf	Digital preservation policies
Harnessing the power of digital data for science and society	Non-domain specific	http://www.nitrd.gov/About/Harnessing_Power_Web.pdf	Digital preservation policies
Principles and Good Practice for Preserving Data	Non-domain specific	http://www.ihsn.org/home/index.php?q=focus/principles-and-good-practice-preserving-data	Digital preservation policies
Digital Preservation policies (JISC)	Non-domain specific	www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf	Digital preservation policies
Collection development policy for the national geospatial digital archive (NGDA)	Non-domain specific	http://www.ngda.org/docs/NGDA_Collection_Development_Policy.pdf	Digital preservation policies
Facing Off with Digital Preservation Policy	Non-domain specific	http://blogs.loc.gov/digitalpreservation/2011/07/facing-off-with-digital-preservation-policy/	Digital preservation policies
OCLC Digital	Non-domain	http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf	Digital

Resource Name	Domain	URL	Topic
Archive Preservation Policy	specific		preservation policies
UK archive Preservation Policy	Non-domain specific	http://www.data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf	Digital preservation policies
Libraries as Distributors of Geospatial Data: Data Management Policies as Tools for Managing Partnerships	Non-domain specific	http://muse.jhu.edu/journals/lib/summary/v055/55.2steinhart.html	Digital preservation policies
DeepArc	Cross – domain	http://deeparc.sourceforge.net/	Data preservation technologies
Xing	Cross – domain	http://sourceforge.net/projects/xinq/	Data preservation technologies
KEEP Emulation Framework	Cross – domain	http://emuframework.sourceforge.net/	Data preservation technologies
Dioscuri	Cross – domain	http://dioscuri.sourceforge.net/`	Data preservation technologies

Resource Name	Domain	URL	Topic
PLATO	Cross – domain	http://www.ifs.tuwien.ac.at/dp/plato/intro.html	Data preservation technologies
DRAMBORA	Cross – domain	http://www.repositoryaudit.eu/	Data preservation technologies
Data Asset Framework	Cross domain	http://www.data-audit.eu/	Data preservation technologies
DMP online	Cross domain	http://www.dcc.ac.uk/dmponline	Data preservation technologies
D-SPACE	Cross domain	http://www.dcc.ac.uk/dmponline	Data preservation technologies
E-Prints	Cross domain	http://www.eprints.org/software/	Data preservation technologies
Hierarchical storage management	Cross domain	http://www.stfc.ac.uk/e-Science/services/atlas-petabyte-storage/22459.aspx	Data preservation technologies
Clotho	Biology	http://www.clothocad.org	Data processing
CAA Forum	Non-domain specific	http://www.cca-forum.org/index.html	Data processing
Open DA	Non-domain specific	http://www.openda.org	Data processing

Annex D. Questionnaires used for tasks 15.2 and 15.3

E.1 Initial on-line survey questions

Question	Category
Name:	User information
Organisation:	User information
Email address:	User information
Position in organisation:	User information
Country:	User information
Select your own primary activity from:	User information
If you are involved in more than one activity please also select this below:	User information
1.01 Select main discipline areas for which you are providing data:	Digital archive service provider
Other disciplines	Digital archive service provider
1.02 Select main job role:	Digital archive service provider
1.03 What do users of your archive receive:	Digital archive service provider
1.04 How do your users analyse and exploit the data:	Digital archive service provider
1.04a If possible please name the software used:	Digital archive service provider
1.05 What type of storage system (software and hardware) do you use:	Digital archive service provider
1.05a Which system do you use?	Digital archive service provider
1.06 What are the critical operations you need to perform on your archive? (Please select all that apply)	Digital archive service provider
1.07 How long do you expect the data to be preserved in its present repository?	Digital archive service provider

Question	Category
1.08 How do you provide/prepare/package the data to/for the users (re-process it, combine it with other data to generate maps, indexes, etc.)?	Digital archive service provider
1.09 What additional software do you need to perform the above activities? (reprocessing etc).	Digital archive service provider
1.10 How do users find data:	Digital archive service provider
1.11 How are data made accessible to users:	Digital archive service provider
1.12 What access constraints are imposed on the data:	Digital archive service provider
1.13 When providing data to users, is any other supporting information (e.g. documentation) provided with the data	Digital archive service provider
1.14 In what form and how is this supporting information made available?	Digital archive service provider
2.01 Select main discipline areas for which you are providing data:	Creator/producer of major data sets
Other disciplines	Creator/producer of major data sets
2.02 Select main job role:	Creator/producer of major data sets
2.03 Do you use a centralised structured archive system to store and preserve your organisations data?	Creator/producer of major data sets
2.04 Is this system managed in-house or provided by another organisation?	Creator/producer of major data sets
2.05 Are you able to indicate the name of the organisation providing this service for you?	Creator/producer of major data sets
2.06 What is the approximate time scale between submitting data to the archive and retrieving it for use:	Creator/producer of major data sets
1.03 What do users of your archive receive:	Creator/producer of major data sets
1.04: How do your users analyse and exploit the data:	Creator/producer of major data sets
1.04a: If possible please name the software used:	Creator/producer of major data sets
1.05 What type of storage system do you use?	Creator/producer of major

Question	Category
	data sets
1.05a Which system do you use?	Creator/producer of major data sets
1.06 What are the critical operations you need to perform on your archive? (Please select all that apply)	Creator/producer of major data sets
1.07 How long do you expect the data to be preserved in its present repository?	Creator/producer of major data sets
1.08 How do you provide/prepare/package the data to/for the users (re-process it, combine it with other data to generate maps, indexes, etc.)?	Creator/producer of major data sets
1.09 What additional software do you need to perform the above activities? (reprocessing etc).	Creator/producer of major data sets
1.10 How do users find data:	Creator/producer of major data sets
1.11 How are data made accessible to users:	Creator/producer of major data sets
1.12 What access constraints are imposed on the data:	Creator/producer of major data sets
1.13 When providing data to users, is any other supporting information (e.g. documentation) provided with the data?	Creator/producer of major data sets
1.14 In what form and how is this supporting information made available?	Creator/producer of major data sets
2.07 How do you store your digital materials/products?	Creator/producer of major data sets
2.08 How do you allow discovery of data in your store e.g.	Creator/producer of major data sets
2.11 Are you satisfied with the way users are able to find and use your data at present?	Creator/producer of major data sets
2.11a What would you improve?	Creator/producer of major data sets
3.01 Select main discipline areas for which you are providing data:	End user of archived data
Other disciplines	End user of archived data
3.02 Select main job role:	End user of archived data
3.03 What approach do you generally follow to find the data you require?	End user of archived data
3.04 How easy is it to locate the data you require?	End user of archived data

Question	Category
3.05 How can ease of locating the data in the archive be improved?	End user of archived data
3.06 How easy is it to Access the data you require?	End user of archived data
3.07 How can access to the archive be improved?	End user of archived data
3.08 What searching/browsing functionality is provided:	End user of archived data
3.09 How do you analyse the information retrieved:	End user of archived data
3.10 Is the software which you use day to day for your work integrated with the software used to access the archive of preserved data:	End user of archived data
3.11 Is other supporting information (e.g. documentation on using the data) provided as standard?	End user of archived data
3.12 In what form is this supporting information provided?	End user of archived data
3.13 Would you need to have access to additional information about the data when discovering it?	End user of archived data
3.14 Do you have any mechanism to check the consistency and completeness of data you are retrieving from the archive?	End user of archived data
3.15 When you receive data from the archive is it in a useful format to use or is conversion required?	End user of archived data
3.16 What is the approximate time scale between submitting data to the archive and retrieving it for use:	End user of archived data
3.17 How do you use the data for your specific activities (re-process it, combine it with other data to generate maps, indexes, etc.). And what additional software do you need to perform that.	End user of archived data
3.18 How easy is it to submit new data to the archive provider:	End user of archived data
3.19 Please provide a comment on how the submission procedure could be improved	End user of archived data
3.20 Are there any datasets produced by other disciplines which you don't currently have access to, but would like to make use? If so please indicate the discipline area, and the type/format of data of interest.	End user of archived data
3.21 In broad terms is there any predominant format to the data you normally archive (e.g. commonly images, often data in tabular form etc)? If so please provide a brief description of the formats commonly used.	End user of archived data
3.22 On retrieving data from the archive, do you receive back sufficient information for that data to be useful, if not what is missing (e.g. metadata, visualisation tools?).	End user of archived data
4.01 Would you be interested in more information about archive systems and digital preservation practices?	Other interest in SCIDIP-ES
4.02 Would you like to be part of a community dealing with digital data preservation and to receive updates on the SCIDIP-ES project?	Other interest in SCIDIP-ES
4.03 Would you like to receive the SCIDIP-ES newsletter?	Other interest in SCIDIP-ES
4.04 Do you have any further comments relevant to this survey you wish to make?	Other interest in SCIDIP-ES

E2. Questions to guide direct user consultation (one-to-one interviews) in Task 15.2

Questions for archive service providers

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
System Architecture/Infrastructure				
1.01 Please describe (describe further) the commercial/open source/other system you are using for data archiving e.g. is this a dedicated archive system for data preservation, or does this form part of your day-to-day corporate database systems.	1.05,1.05a	Whether archiving system is commercial/open source (or combination of these). Clarify response from on-line questionnaire where relevant Name of system (if not already provided) How archiving for preservation relates to other data systems		
1.02 Please describe any standards (internal or external to your organisation) to which the archiving software used		Whether such standards exist. Specify standards used		

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
conforms.				
1.03 If this is an open source/customised system please provide the name of the open source product or technology used, and some information on the technologies involved (e.g. Java, C++ etc)		Name of open source tools/technologies (If not already provided) Underlying languages/databases (if known) How do the open source/customised components relate to each other, and to other parts of the system architecture		
1.04 If an Open Source archiving system is being used how did you come to adopt this – e.g. on the basis of the standards used, because you are linked to a project which produced this, because this system is relevant to your domain etc.		Reason for choice of open source Reason for choosing a specific system		

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
1.05 What is the average number of concurrent users of the archive system		Number of concurrent users		
1.06 Where the organisations activities fall into more than one of the activity areas (e.g. an archive service provider who is also a producer of major datasets) how does this affect the system architecture (if at all)				
Data Discovery				
1.07 Describe (further) the tools and services available to you (or provided by you) for discovering and retrieving the data you require/provide e.g. including tools for browsing and searching. Where these are commercially available products obtain the names, and software supplier	1.10	Names and functionality of each of the tools and services used. Where commercial – software supplier details, software platform, operating system, and if possible version numbers		

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
details, and if possible version numbers				
1.08 Where the tools etc used have been customised from a commercial/open source product, to what extent has customisation been undertaken, and for what reason		Reason for customisation Extent of customisation – tools or services, or both		
1.09 Aim to understand where these tools and services sit in terms of the system architecture e.g. is the archive store local or remote to the user, are web services, or ftp etc used for access and retrieval. Are the tools web based or desktop enabled (if known)		Broad description of system architecture (as far as possible via interview) Are web services used, if so what protocols and standards are used		
1.10 Explore issues mentioned in the on-line survey concerned with use of formal metadata standards versus adhoc systems		Which formal metadata standard(s) are used		

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
Data Access				
1.11 When data is retrieved from the archive in what format is this (expand on on-line questionnaire answer 1.03 as required) e.g. just the original data in its original format, something created by conversion from the original data. Also does the information retrieved include information on how the data was generated, or how it is intended to be used? Where such additional information is present, what tools are available to help you find and access this information	1.03,1.13,1.14	Format of data retrieved from the archive, and is this the same as was submitted (elaborate on on-line response) Is additional information supplied about the data What format is the supporting information in		
1.12 What are the mechanisms used for accessing data e.g. via web services, off-line delivery, ftp etc. If via web services what	1.11	What are mechanisms, e.g. ftp ,web services. If web services which standards and protocols used		

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
standards are currently used (if known)		(if not already captured above)		
Preservation				
1.13 Explore comments made on how long data will be stored in its present repository (Qu. 1.07 in on-line questionnaire)	1.07	What migration processes are in place What tools are available to assist migration		
1.14 Expand on comments made about how data is prepared and packaged (Qu. 1.08 in on-line questionnaire), and what software is required	1.08,1.09			
1.15 What systems/procedures do you have in place to support migrating your archived data from its current store to a new archive system		Systems available (or absence of)		
1.16 Other gaps in provision – e.g. if they have a tool to check consistency and completeness of what is		Describe tool used How does it work		

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
retrieved from the archive, how does this work. If no tool is currently available, how would you want such a tool to work		If no tool currently available, how would you want such a tool to work		
Processing				
1.17 How does the user analyse the information retrieved from the archive, and whether the analysis software is integrated with the software used to access the archive (if applicable to archive service provider)	1.04	Determine software and methods use (elaborate on users response to the on-line survey) How integrated is the processing software with the software used to access the archive		
1.18 Elaborate on the critical operations performed on the archive e.g. what migration, reprocessing, data conversion etc	1.06	Describe critical operations performed		
Knowledge Extraction				
1.19 What visualisation/analysis tools (or other knowledge		List tools available Which data is supported		

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
extraction tools) are provided or available – are these specific to certain data types, how could these be extended/ improved?		What are gaps		
1.20 Aim to understand how and to what extent the knowledge extraction and visualisation tools are integrated with the archiving environment		Are knowledge extraction tools integrated with archive environment How does this integration work		
Preservation Policies and Guidelines				
1.21 Aim to understand where they are in terms of adopting digital preservation policies and methods etc		Current preservation initiatives in which involved Any understanding of OAIS and similar models		
1.22 Does your organisation have a data preservation policy/strategy, if so is this clearly documented (would it		What is policy/strategy for data preservation – is it documented		

Question/Area of Focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
be possible for the SCIDIP-ES project to obtain a copy for research purposes).				
1.23 What are they key objectives of this policy – what are the main data types, preservation requirements etc.		What are the main points, and priorities what types of data covered		
1.24 If no formal policy exists what are your objectives in data preservation		Record Objectives		
1.25 Do you have a data/information management strategy		Main points, and priorities for preservation		
1.26 Finally, follow up on any comments made in the on-line survey, if not already covered in the questions above				

Questions for creators/producers of major datasets

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
Data Discovery				
2.01 Describe (further) the tools and services available to you (or provided by you) for discovering and retrieving the data you require/provide e.g. including tools for browsing and searching. Where these are commercially available products obtain the names, and software supplier details, and if possible version numbers	1.10,2.08,2.11	Names and functionality of each of the tools and services used. Where commercial – software supplier details, software platform, operating system, and if possible version numbers		
2.02 Is this a commercial or open source solution (if known)		Record commercial or open source, or other (e.g. customised)		
2.03 Where the tools etc used have been customised from a commercial/open source product, to what extent has customisation been undertaken, and for what reason		Reason for customisation Extent of customisation – tools or services, or both		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
2.04 Aim to understand where these tools and services sit in terms of the system architecture e.g. is the archive store local or remote to the user, are web services, or ftp etc used for access and retrieval. Are the tools web based or desktop enabled (if known)		Broad description of system architecture (as far as possible via interview) Are web services used, if so what protocols and standards are used		
2.05 Can the user locate all the archived data they need using one search interface or do you need to use several? –please describe further how searching for multiple types of data works		How many search interfaces are used to locate the required data Where more than one interface is used, which tool/interface is used for which type of data What problems does using multiple search interfaces create Explore interest in		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
		harmonisation of data catalogues		
2.06 Explore issues mentioned in the on-line survey concerned with use of formal metadata standards versus adhoc systems		Which formal metadata standard(s) are used		
Data Access				
2.07 When data is retrieved from the archive in what format is this e.g. just the original data in its original format, something created by conversion from the original data. Also does the information retrieved include information on how the data was generated, or how it is intended to be used? Where such additional information is present, what tools are available to help you find and access this information	1.03,1.13,1.14	Format of data retrieved from the archive, and is this the same as was submitted (elaborate on on-line response) Is additional information supplied about the data What format is the supporting information in		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
2.08 What are the mechanisms used for accessing data e.g. via web services, off-line delivery, ftp etc. If via web services what standards are currently used (if known)	1.11	What are mechanisms, e.g. ftp ,web services. If web services which standards and protocols used (if not already captured above)		
2.09 Where respondent has indicated that accessing/locating data was not very easy, explore this potential gap in provision, what areas could be improved e.g. what would make this easier – requirements for additional software tools? Where they have indicated that accessing or locating the data is fairly easy, then what particular features of the system procedures (e.g. what tools and services)		What could be improved Identify gaps in provision Where access is easy, what tools/services facilitate this		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
make it easy?				
Preservation				
2.10 Explore comments made on how long data will be stored in its present repository	1.07	What migration processes are in place What tools are available to assist migration		
2.11Expand on comments made about how data is prepared and packaged (Qu. 1.08 in on-line questionnaire), and what software is required	1.08,1.09			
2.12 Over what period of time do you normally need to preserve your data	2.06	Time period e.g. 50 years or whatever Elaborate on comments made in the on-line questionnaire (Qu's.2.06 and3.16) about the interval between submitting and retrieving data from the archive (Sometimes not clear		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
		Do different types of data require different periods of preservation		
2.13 Firm up and elaborate on responses to Qu. 2.07 about how data is stored, and common data formats used – how much of data is in these formats, what are next most important formats	2.07	Elaborate on response to Qu.3.21		
2.14 What systems/procedures do you have in place to support migrating your archived data from its current store to a new archive system – follow up on archive life cycle question in on-line survey		Systems available (or absence of)		
2.15 Other gaps in provision – e.g. if they have a tool to check consistency and completeness of what is retrieved from the archive, how does this work. If no		Describe tool used How does it work		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
tool is currently available, how would you want such a tool to work				
2.16 Expand on response to Qu. 3.22 in on-line survey - whether they receive sufficient information back from the archive for that to be useful, and if not what is missing				
Processing				
2.17 How does the user analyses the information retrieved from the archive, and whether the analysis software is integrated with the software used to access the archive	1.04	Determine software and methods use (elaborate on users response to the on-line survey) How integrated is the processing software with the software used to access the archive		
2.18 How far does the metadata and supporting information provided with the data support any		Describe how metadata and supporting info support processing		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
onward processing which you need to do				
2.19 How easy is it to submit new data to the archive		Easy/moderate/difficult How could submission procedure be improved (elaborate on response to Qu. 3.19 in on-line survey)		
2.20 Elaborate on the critical operations performed on the archive e.g. what migration, reprocessing, data conversion etc	1.06	Cover critical operations performed		
2.21 Briefly describe the processing workflow for data which you retrieve (i.e. a typical workflow for your discipline) what intermediate products are created, and how are these preserved (in order to determine common workflows which it would be good to support).		Brief and simple workflow description Follow up on online survey Qu.3.15 about whether data received from the archive in a useful format or requires conversion – elaborate on what conversion and using which software		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
		Also explore response to Qu.3.17 about how data from the archive is used, and what software is used		
2.22 Explore interest in using data from one discipline within another. If they did not complete the on-line survey explore their interest in this. What particular benefit do they see from such interoperability, and for what specific types of data		To what extent is respondent interested in interoperability between disciplines Between which disciplines (top 3) For which types/formats of data (capture priority order) What benefits do they hope to realise from greater interoperability		
Knowledge Extraction				
2.23 What visualisation/analysis tools (or other knowledge extraction tools) are provided or available – are		List tools available Which data is supported What are gaps		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
these specific to certain data types, how could these be extended/ improved?				
2.24 Aim to understand how and to what extent the knowledge extraction and visualisation tools are integrated with the archiving environment		Are knowledge extraction tools integrated with archive environment How does this integration work		
Preservation Policies and Guidelines				
2.25 Aim to understand where they are in terms of adopting digital preservation policies and methods etc		Current preservation initiatives in which involved Any understanding of OAIS and similar models		
2.26 Does your organisation have a data preservation policy/strategy, if so is this clearly documented (would it be possible for the SCIDIP-ES project to obtain a copy for		What is policy/strategy for data preservation – is it documented		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
research purposes).				
2.27 What are they key objectives of this policy – what are the main data types, preservation requirements etc.		What are the main points, and priorities what types of data covered		
2.28 If no formal policy exists what are your objectives in data preservation		Record Objectives		
2.29 Do you have a data/information management strategy		Main points, and priorities for preservation		

Questions for end users of archived data

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
Data Discovery				

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
3.01 Describe (further) the tools and services available to you (or provided by you) for discovering and retrieving the data you require/provide e.g. including tools for browsing and searching. Where these are commercially available products obtain the names, and software supplier details, and if possible version numbers	3.08	Names and functionality of each of the tools and services used. Where commercial – software supplier details, software platform, operating system, and if possible version numbers		
3.02 Is this a commercial or open source solution (if known)		Record commercial or open source, or other (e.g. customised)		
3.03 Where the tools etc used have been customised from a commercial/open source product, to what extent has customisation been undertaken, and for what reason		Reason for customisation Extent of customisation – tools or services, or both		
3.04 Aim to understand		Broad description of system		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
where these tools and services sit in terms of the system architecture e.g. is the archive store local or remote to the user, are web services, or ftp etc used for access and retrieval. Are the tools web based or desktop enabled (if known)		architecture (as far as possible via interview) Are web services used, if so what protocols and standards are used		
3.05 Can the user locate all the archived data they need using one search interface or do you need to use several? –please describe further how searching for multiple types of data works		How many search interfaces are used to locate the required data Where more than one interface is used, which tool/interface is used for which type of data What problems does using multiple search interfaces create Explore interest in harmonisation of data		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
		catalogues		
3.06 Explore issues mentioned in the on-line survey concerned with use of formal metadata standards versus adhoc systems		Which formal metadata standard(s) are used		
Data Access				
3.07 When data is retrieved from the archive in what format is this e.g. just the original data in its original format, something created by conversion from the original data. Also does the information retrieved include information on how the data was generated, or how it is intended to be used? Where such additional information is present, what tools are available to help you find and access this information		Format of data retrieved from the archive, and is this the same as was submitted (elaborate on on-line response) Is additional information supplied about the data What format is the supporting information in		
3.08 What are the		What are mechanisms, e.g.		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
mechanisms used for accessing data e.g. via web services, off-line delivery, ftp etc. If via web services what standards are currently used (if known)		ftp ,web services. If web services which standards and protocols used (if not already captured above)		
3.09 Where respondent has indicated that accessing/locating data was not very easy, explore this potential gap in provision, what areas could be improved e.g. what would make this easier – requirements for additional software tools? Where they have indicated that accessing or locating the data is fairly easy, then what particular features of the system procedures (e.g. what tools and services) make it easy?		What could be improved Identify gaps in provision Where access is easy, what tools/services facilitate this		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
Preservation				
3.10 How long data will be stored in its present repository		What migration processes are in place What tools are available to assist migration		
3.11 How is data prepared and packaged what software is required				
3.12 Over what period of time do you normally need to preserve your data	3.16	Time period e.g. 50 years or whatever Elaborate on comments made in the on-line questionnaire (Qu's.2.06 and3.16) about the interval between submitting and retrieving data from the archive (Sometimes not clear Do different types of data require different periods of preservation		
3.13 Firm up and elaborate	3.21	Elaborate on response to		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
on responses to Qu. 2.07 and 3.21 about how data is stored, and common data formats used – how much of data is in these formats, what are next most important formats		Qu.3.21		
3.14 What systems/procedures do you have in place to support migrating your archived data from its current store to a new archive system – follow up on archive life cycle question in on-line survey		Systems available (or absence of)		
3.15 Other gaps in provision – e.g. if they have a tool to check consistency and completeness of what is retrieved from the archive, how does this work. If no tool is currently available, how would you want such a tool to work	3.14	Describe tool used How does it work		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
3.16 Expand on response to Qu. 3.22 in on-line survey - whether they receive sufficient information back from the archive for that to be useful, and if not what is missing	**3.22**			
Processing				
3.17 Further explore responses to Qu. 3.09 and 3.10 in online survey about how the user analyses the information retrieved from the archive, and whether the analysis software is integrated with the software used to access the archive	3.09,3.10	Determine software and methods use (elaborate on users response to the on-line survey) How integrated is the processing software with the software used to access the archive		
3.18 How far does the metadata and supporting information provided with the data support any onward processing which you need to do		Describe how metadata and supporting info support processing		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
3.19 How easy is it to submit new data to the archive (also follow up on response made to online survey Qu.3.18, where applicable).	3.18	Easy/moderate/difficult How could submission procedure be improved (elaborate on response to Qu. 3.19 in on-line survey)		
3.20 Elaborate on the critical operations performed on the archive e.g. what migration, reprocessing, data conversion etc		Cover critical operations performed		
3.21 Briefly describe the processing workflow for data which you retrieve (i.e. a typical workflow for your discipline) what intermediate products are created, and how are these preserved (in order to determine common workflows which it would be good to support).	3.15,3.17	Brief and simple workflow description Follow up on online survey Qu.3.15 about whether data received from the archive in a useful format or requires conversion – elaborate on what conversion and using which software Also explore response to Qu.3.17 about how data from		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
		the archive is used, and what software is used		
3.22 Expand on any comments the respondent has made on interoperability of data e.g. interest in using data from one discipline within another. If they did not complete the on-line survey explore their interest in this. What particular benefit do they see from such interoperability, and for what specific types of data	3.20	To what extent is respondent interested in interoperability between disciplines Between which disciplines (top 3) For which types/formats of data (capture priority order) What benefits do they hope to realise from greater interoperability		
Knowledge Extraction				
3.23 What visualisation/analysis tools (or other knowledge extraction tools) are provided or available – are these specific to certain data types, how could these be extended/ improved?		List tools available Which data is supported What are gaps		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
3.24 Aim to understand how and to what extent the knowledge extraction and visualisation tools are integrated with the archiving environment		Are knowledge extraction tools integrated with archive environment How does this integration work		
Preservation Policies and Guidelines				
3.25 Aim to understand where they are in terms of adopting digital preservation policies and methods etc		Current preservation initiatives in which involved Any understanding of OAIS and similar models		
3.26 Does your organisation have a data preservation policy/strategy, if so is this clearly documented (would it be possible for the SCIDIP-ES project to obtain a copy for research purposes).		What is policy/strategy for data preservation – is it documented		
3.27 What are they key		What are the main points,		

Question/Area of focus	Related on-line survey question (where applicable)	Key Information to Capture	RECORD RESPONSES TO QUESTIONS	
objectives of this policy – what are the main data types, preservation requirements etc.		and priorities what types of data covered		
3.28 If no formal policy exists what are your objectives in data preservation		Record Objectives		
3.29 Do you have a data/information management strategy		Main points, and priorities for preservation		

E.3 SCIDIP-ES – Questionnaire on Semantics, Metadata and Ontologies (Task 15.3)

The SCIDIP-ES project will build on the experience of the European Space Agency's Earth Observation Long Term Data Preservation (LTDP) programme with the aim of setting-up a European framework for the long term preservation of Earth Science (ES) data through the definition of common preservation policies, the harmonization of metadata and semantics and the deployment of generic data preservation services.
The aim of this questionnaire is to develop an understanding in terms of Semantics, Metadata and Ontologies that are currently in use by the Earth Science Community. The responses of the questionnaire will be the basis for the definition of the most appropriate strategy to have harmonized semantics, metadata and ontologies.

1. Name

2. Email

3. Country

4. Organization

5. Organization Type
<Select>

6. Select your activity

- ☐ SCIDIP-ES Partner
- ☐ Developer of Metadata Schema/Ontology
- ☐ User of Metadata Schema/Ontology
- ☐ Software Engineer
- ☐ Other (please specify)
- ☐ SCIDIP-ES Partner
- ☐ Developer of Metadata Schema/Ontology
- ☐ User of Metadata Schema/Ontology
- ☐ Software Engineer
- ☐ Other (please specify)

7. Name and describe in brief the main (up to five) metadata schemas, ontologies or other semantics-related formats that you use. For each please provide its name and URI, a short description, its current use, and available formats (UML, XML, RDF/S, OWL, other).

1	**Name**	**URL/URI**
	Formats	☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other ☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other
	Usage	☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ Other ☐ Other
	Description	
2	**Name**	**URL/URI**
	Formats	☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other ☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other
	Usage	☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ Other ☐ Other
	Description	
3	**Name**	**URL/URI**
	Formats	☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other ☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other
	Usage	☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ for querying data ☐ for exchanging data ☐ for natively storing data

		☐ Other ☐ Other **Description**
6	**Name** **Formats** **Usage** **Description**	**URL/URI** ☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other ☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other ☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ Other ☐ Other
5	**Name** **Formats** **Usage** **Description**	**URL/URI** ☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other ☐ XML ☐ UML ☐ RDF/S ☐ OWL ☐ Other ☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ for querying data ☐ for exchanging data ☐ for natively storing data ☐ Other ☐ Other

8. Name those of the above for which your organization has datasets described according to them? (Provide links to these datasets, indicate whether they are published as Linked Data).

Annex E. Extra Material for the main Semantic Models

Here we provide more information about the semantic models described at Section 6. To aid the reader we include some indicative diagrams & figures.

ISO 19100 series

The ISO 19100[24] is a series of standards that supports data management, acquiring, processing, analyzing, accessing, presenting and transferring data between different users, systems and locations for geographic information (i.e. information concerning objects or phenomena that are directly or indirectly associated with a location relative to Earth). The Geospatial community mainly uses are standards related to "Metadata, Data Content and Definition". Among these we found that the ISO-19115 standard is the one that most users use.

ISO 19110:2005 defines the methodology for cataloguing feature types and specifies how the classification of feature types is organized into a feature catalogue and presented to the users of a set of geographic data. It also applies to the cataloguing of feature types that are represented in digital form. Its principles can be extended to the cataloguing of other forms of geographic data.

ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and the temporal schema, spatial reference, and distribution of digital geographic data. The standard defines

- Mandatory and conditional metadata sections, metadata entities, and metadata elements
- The minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data)
- Optional metadata elements – to allow for a more extensive standard description of geographic data, if required.
- A method for extending metadata to fit specialized needs.

ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and the temporal schema, spatial reference, and distribution of digital geographic data. The standard defines

- Mandatory and conditional metadata sections, metadata entities, and metadata elements
- The minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data)
- Optional metadata elements – to allow for a more extensive standard description of geographic data, if required.
- A method for extending metadata to fit specialized needs.

ISO 19115-2:2009 extends the existing geographic standard by defining the schema required for describing imagery and gridded data. It provides information about the properties of the measuring equipment used to acquire the data, the geometry of the measuring process employed by the equipment, and the production process used to digitize the raw data.

[24] http://www.isotc211.org/

ISO 19126:2009 specifies a schema for feature concept dictionaries to be established and managed as registers. It does not specify schemas for feature catalogues or for the management of feature catalogue as registers.
ISO 19131:2007 specifies requirements for the specification of geographic data products, based upon the concepts of other ISO 19100 International Standards. It describes the content and structure of data product specification and it also provides help in the creation of data product specifications, so that they are easily understood and fit for their intended purpose.
ISO 19139:2007 defines Geographic Metadata XML (gmd) encoding, an XML schema implementation derived from ISO 19115:2003. It provides XML schemas that enhances interoperability by providing a common specification for describing, validating and exchanging metadata about geographic datasets, dataset series, individual geographic features, feature attributes, feature types, software etc.
ISO 19141-1:2009 establishes the structure of a geographic information classification system, together with the mechanism for defining and registering the classifiers for such a system. It specifies the use of discrete coverages to represent the result of applying the classification system to a particular area and defines the technical structure of a register of classifiers in accordance with ISO 19135.
ISO 19156 defines a conceptual schema for observations and for features involved in sampling when making observations. These provide models for the exchange of information describing observation acts and their results, both within and between different scientific and technical communities.

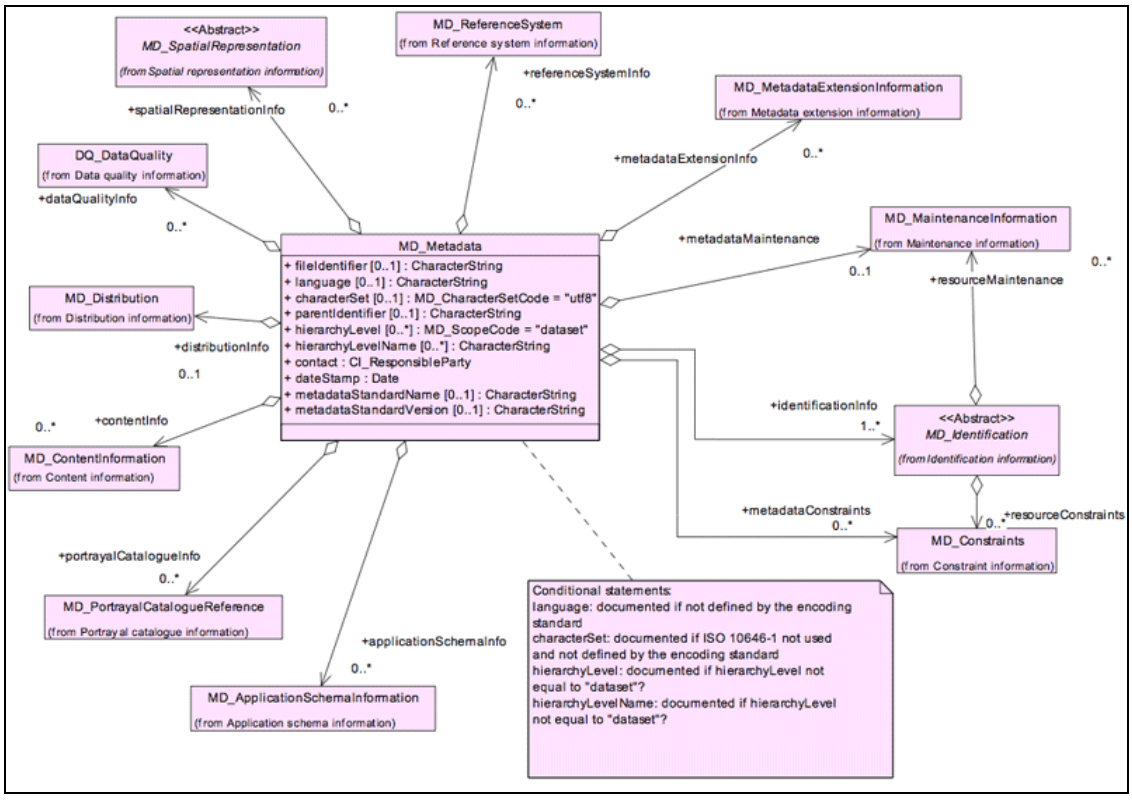


Figure F.1 The ISO 19115 metadata model ***(source: http://www.landmap.ac.uk)***

CF Metadata

The conventions of Climate and Forecast (CF) metadata[2] are designed to promote the processing and sharing of files created with the NetCDF API. Although CF is framed as a standard for data written in NetCDF, its ideas relate to metadata design in general, and hence can be contained in other formats such as XML. The NetCDF library is designed to read and write data that has been structured according to well-defined rules and is easily ported across various computer platforms. The purpose of CF conventions is to require conforming datasets to contain sufficient metadata that they are self-describing in the sense that each variable in the file has an associated description of what it represents, and that each value can be located in space (relative to earth-based coordinates) and time. The standard is intended for use with climate and forecast data, for atmosphere, surface and ocean, and was designed with model-generated data particularly in mind. CF conventions generalize and extend the COARDS[25] conventions.

The conventions define metadata which provide a description of what the data in each variable represents. This enables users from different sources to decide which quantities are comparable. For this purpose CF requires all variables to have units, unless they contain dimensionless numbers. Units are expressed as strings according to the Unidata udunits, which supports many possible units and varieties of syntax (i.e. percent, meter, metre, m, km, etc), although there are several non-SI units that are not directly supported. In addition variables should have associated with them a `standard_name`, which is used to identify the quantity and must be consistent with the unit, a `long_name` and ancillary variables, which are pointers to variables providing metadata about individual data values.

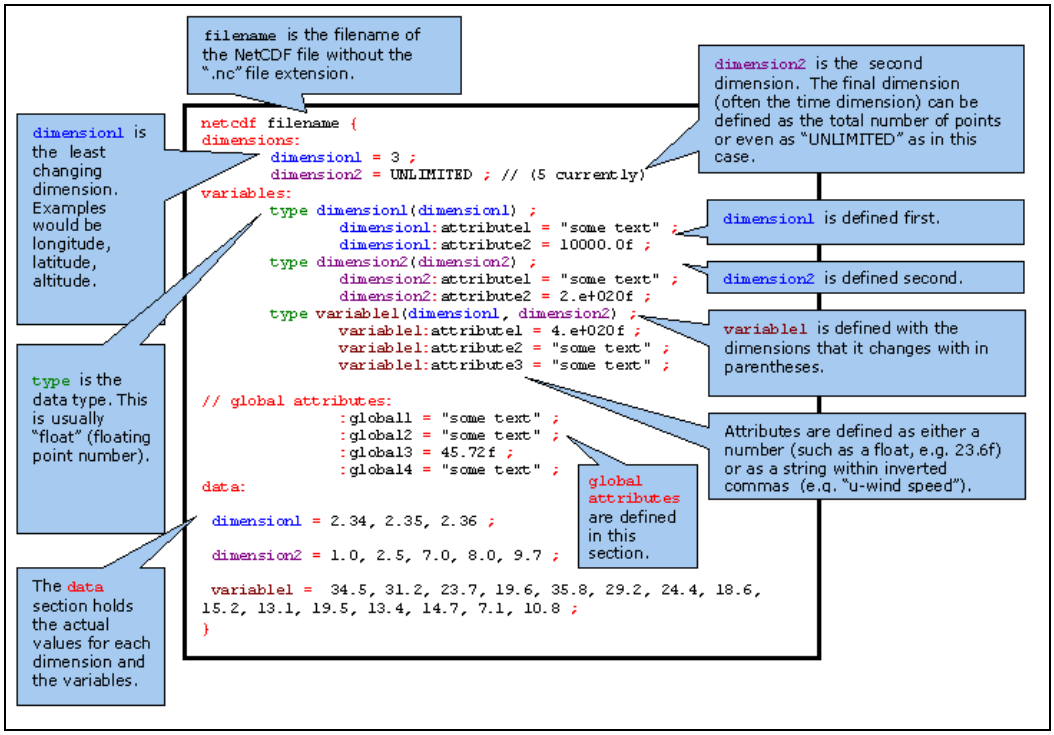


Figure F.2 The structure and syntax of a CDL (ASCII) equivalent to a NetCDF file

(source: http://badc.nerc.ac.uk/help/formats/netcdf/index_cf.html)

[25] http://ferret.wrc.noaa.gov/noaa_coop/coop_cdf_profile.html

CGI Vocabularies and Ontologies

The Commission for the Management and Application of Geoscience Information is a commission of the International Union of the Geological Sciences. Its mission is to enable the global exchange of knowledge about geosciences information and systems. In order to further the objectives of the CGI, various technical collaboration activities are being supported (carried out by formal Working Groups). The CGI Interoperability Working Group has the specific objective of developing a conceptual model of geoscientific information drawing on existing data models and implement an agreed subset of this model in an agreed schema language. The GeoSciML project (initiated in 2003) is part of this working group. It accommodates the short-term goal of representing geo-science information, associated with geologic maps and observations, as well as being extensible in the long-term to other geo-science data. Several vocabularies have already been produced for GeoSciML. A complete listing of these ontologies can be found at: https://www.seegrid.csiro.au/subversion/CGI_CDTGVocabulary/tags/SKOSVocabularies/. One such ontology is shown in the figure below.

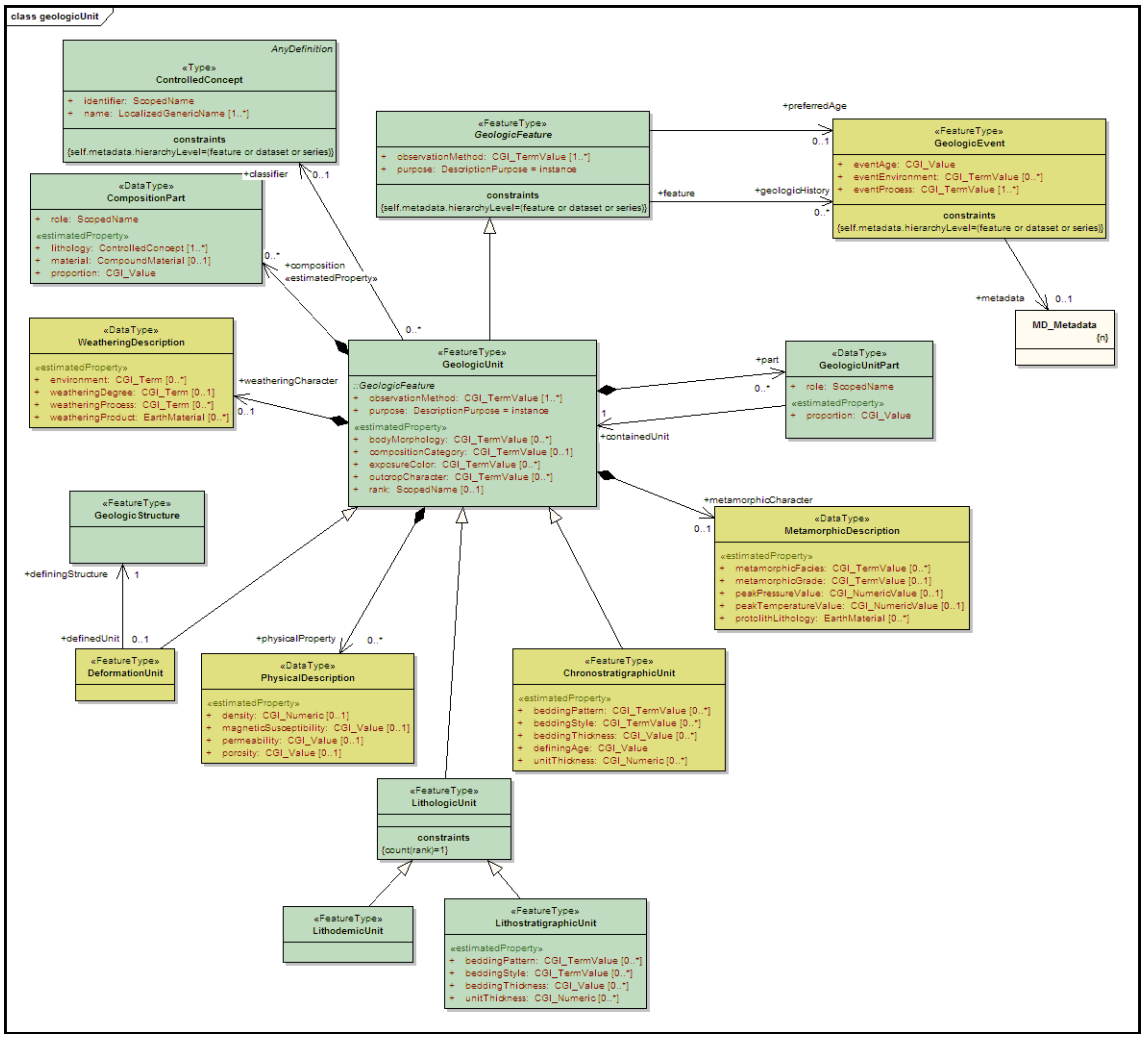


Figure F.3 Vocabulary of the GeologicUnit

(source: https://www.seegrid.csiro.au/wiki/CGIModel/GeoSciMLModel#Working_with_the_UML_model)

OGC

The Open Geospatial Consortium[26] is an international voluntary consensus standards organization. In the OGC more than 400 commercial, governmental, non-profit and research organizations worldwide collaborate in a consensus process encouraging development and implementation of open standards for geospatial content and services, GIS data processing and data sharing. OGC standards are technical documents that detail interfaces or encodings which are used by software developers to build open interfaces and encodings into their products and services.

Some of these OGC standards are:

- CSW: Catalogue Service for the Web which provides access to catalog information
- GML: Geographic Markup Language – an XML-format language for geographic information
- SensorML: Sensor Model Language – an XML-encoding for describing sensors and measurement processes
- WPS: Web Processing Service – provides the rules for standardizing how inputs and outputs are invoked in geospatial processing services
- Observations and Measurements: defines a conceptual schema encoding for observation and measurements
- OGC Reference Model – describes a framework for implementing interoperable solutions and applications for geospatial services, data and applications.
- and others.

[26] http://www.opengeospatial.org/

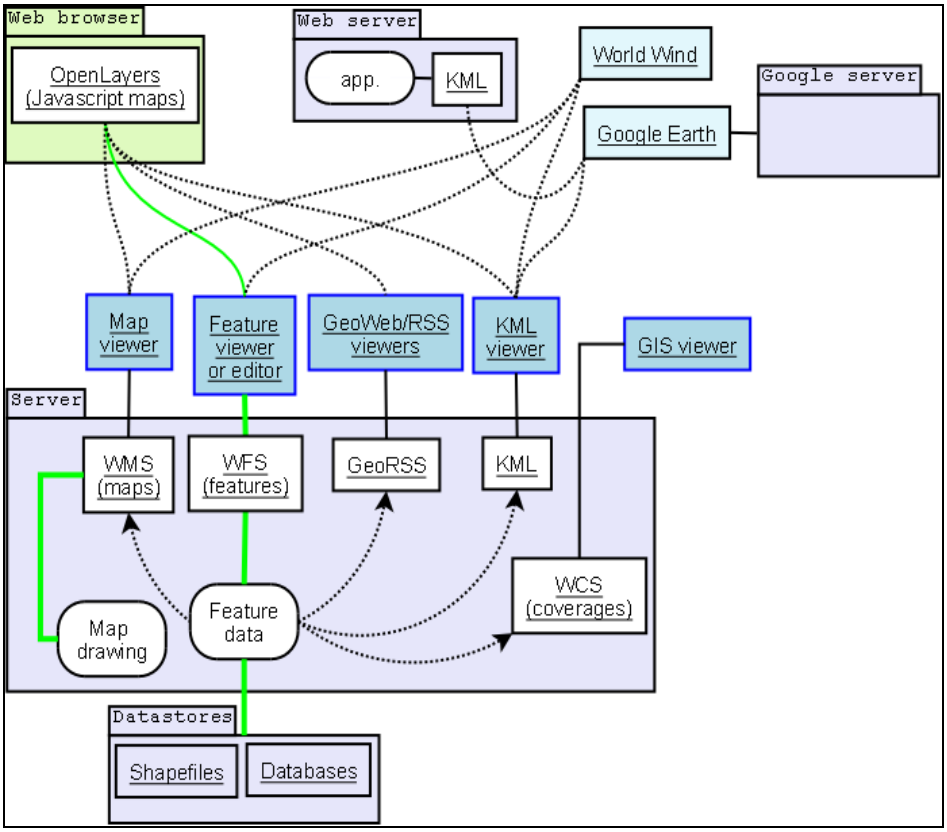


Figure F.4 How OGC Standards help GIS tools communicate

(source: http://en.wikipedia.org/wiki/Open_Geospatial_Consortium)

Climate Science Modelling Language

Climate Science Modelling Language (CSML)[3] is a data model for encoding climate, atmospheric and oceanographic data in terms of geometry-based observation classes such as Points, Profiles, Trajectories and Grids. The current version of CSML is V3.0, while earlier versions of CSML were developed as part of the NERC DataGrid[27] projects.
The following figures shows a classification of observation types using CSML.

[27] http://ndg.badc.rl.ac.uk/

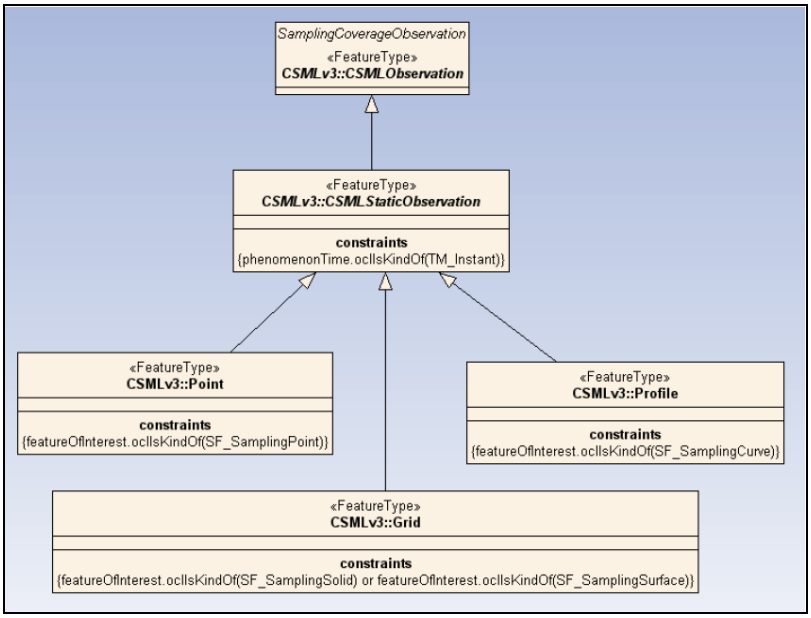


Figure F.5 A Classification of Observation types

(source http://external.opengis.org/twiki_public/pub/MetOceanDWG/MetOceanDWGBonn/CSMLV3_Lowe.pdf)

GEMET

The General Multilingual Environmental Thesaurus[28], is a compilation of several multilingual vocabulary and has been designed as a general thesaurus, aiming to define a core general terminology for the environment. The basic idea for the development of GEMET was to use the best of the presently available excellent multilingual thesauri, in order to save time, energy and funds. GEMET was conceived as a "general" thesaurus, aimed to define a common general language, a core of general terminology for the environment. Specific thesauri and descriptor systems (e.g. on Nature Conservation, on Wastes, on Energy, etc.) have been excluded from the first step of development of the thesaurus and have been taken into account only for their structure and upper level terminology. It has been compiled by merging the terms from several multilingual documents.

[28] http://www.eionet.europa.eu/gemet

Figure F.6 A Screenshot of the GEMET themes listing in English language

(source: http://home.badc.rl.ac.uk/lawrence/blog/2008)

MOIST

Multidisciplinary Oceanic Information SysTem[29] is a relational database, aiming at hosting multidisciplinary data and metadata about seas. It has been initiated within the ESONET NoE project and currently under development in the frame of the ESFRI European Research distributed Infrastructure EMSO - The European Multidisciplinary Seafloor and water column Observatory.
MOIST is aimed at hosting multidisciplinary data and metadata and is organised in two functional blocks. The core part is the harvesting engine that indexes data and keep track of the data source. It is the unique access point for EMSO data mining and retrieval. This central part is connected to the EMSO nodes, sited around European continental margin from the Arctic to the Atlantic, the Mediterranean and the Black Sea, which preserve their own data acquisition systems and databases.

MOLES

The Metadata Objects for Linking Environmental Sciences (**MOLES**)[8] models has been developed within the Natural Environment Research Council (NERC)[30] DataGrid project[31] to fill a missing part of the 'metadata spectrum'. It is a framework within which to encode the relationships between the tools used to obtain data, the activities which organized their use, and the dataset produced. MOLES is primarily of use to consumers of data especially in an interdisciplinary context, to allow them to establish details of provenance, and to compare and contrast such information without resource to discipline-specific metadata or private communications with the original investigators. MOLES is also of use to the custodians of data, providing an organizing paradigm for the data and metadata. The MOLES supports a number of first-class entities, which together provide linkage between key characteristics of the description of data.

[29] http://moist.rm.ingv.it/
[30] http://www.nerc.ac.uk/
[31] http://www.bodc.ac.uk/projects/uk/ndg/

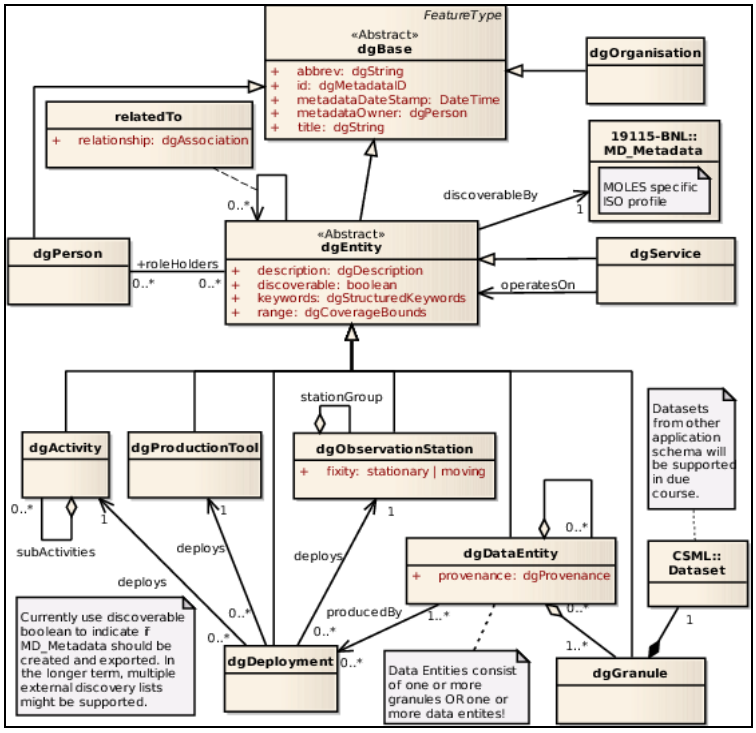


Figure F.7 MOLES basic concepts ***(source: http://home.badc.rl.ac.uk/lawrence/blog/2008)***

OTEG

ESA funded a project called "Open Access Ontology / Terminology for the GMES Space Component" to design openly available semantics including a multi-domain thesaurus and vocabulary and a GMES Space Component Data Access taxonomy. The system is based on a set of interrelated ontologies that users can browse. ESA OTEG Ontology (in particular version 0.8) is comprised from 63 classes. An online application for browsing the OTEG ontology is available from ESA OTEG Ontology http://gmesdata.esa.int/OTE/navigateInfoDomain. The following figure depicts the listing of OTEG themes using the OTEG themes navigator map.

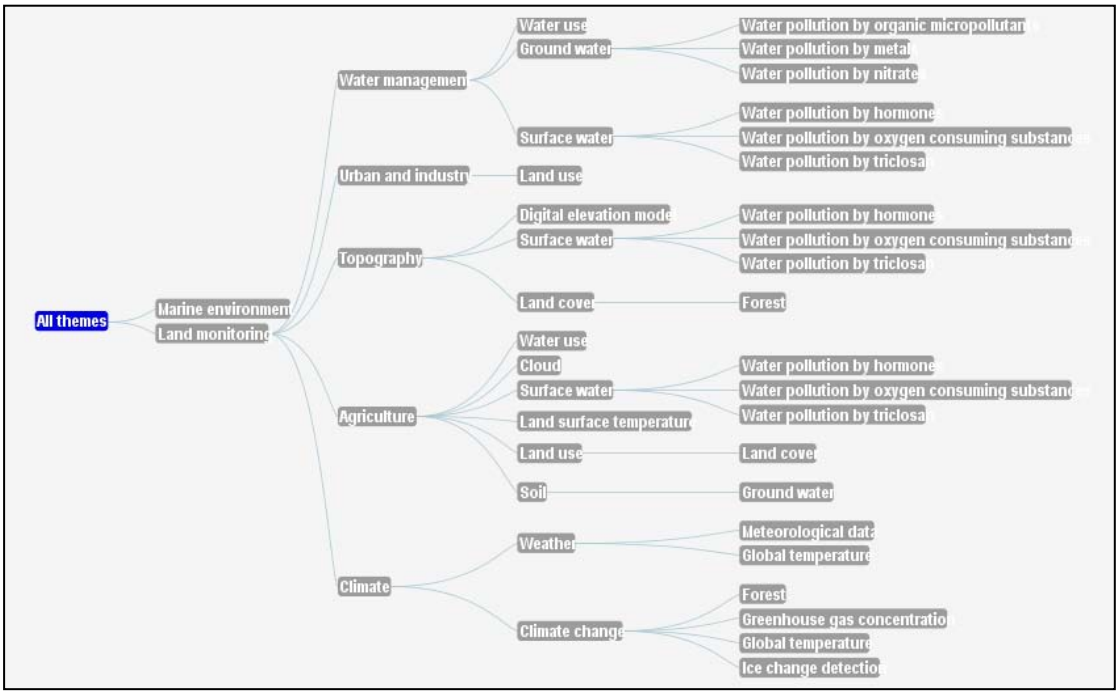


Figure F.8 Screenshot from the OTEG interactive map

(source: http://gmesdata.esa.int/OTE/navigateInfoDomain)

SWEET Ontologies

SWEET [7] which is an abbreviation for Semantic Web for Earth and Environmental Terminology defines a set of ontologies that provide a common semantic framework for representing Earth Science data, information and knowledge. The development of SWEET ontologies result in a common semantic framework that enables ontology-aware software tools to understand the meaning of concepts and terms in documents and web pages. SWEET2.2 is highly modular with more than 6000 concepts in 200 ontologies.

Briefly:

- SWEET Ontologies were created by NASA through the ESTO(Earth Science Technology Office)

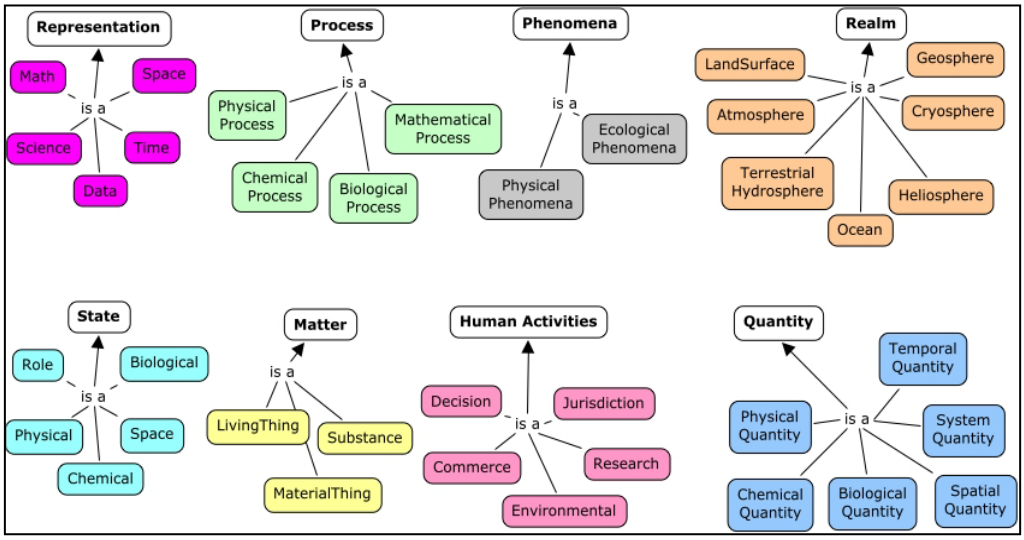


Figure F.9 The eight top-level concepts of SWEET

(source: http://sweet.jpl.nasa.gov/ontology/)

funding.

- They are publicly available and are written in the OWL ontology language.
- The main guiding principles for the development of the ontologies were: (a) scalability, (b) application independence, (c) natural language independence, (d) orthogonality, (e) community involvement.

CIDOC CRM

CIDOC Conceptual Reference Model (ISO 21127)[9] is a core ontology of 80 classes and 132 relations describing the underlying semantics of over a hundred database schemata and structures from all museum disciplines, archives and libraries. It provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. CIDOC CRM is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. CIDOC CRM is the result of long-term interdisciplinary work and agreement. It has been derived by integrating (in a bottom-up manner) hundreds of metadata schemas and is stable (almost no change the last 10 years). We could say that the basic design principles are (a) empirical bottom-up knowledge engineering and (b) object-oriented modeling. As regards the latter, CIDOC CRM has a rich structure of "intermediate" classes and relations, which apart from being very useful for building query services (enabling queries at various levels of abstraction and granularity), it makes its extension to other domains easier and reduces the risk of over-generalization/specialization. In essence, it is a generic model for recording the "what has happened" in human scale.
CIDOC-CRM has been used from several SCIDIP-ES partners (FORTH, ESA) as a core conceptual schema for various purposes. In particular ESA has been used a extension of it (the CRM Digital ontology [6]) in the past for modeling the provenance of digital information.

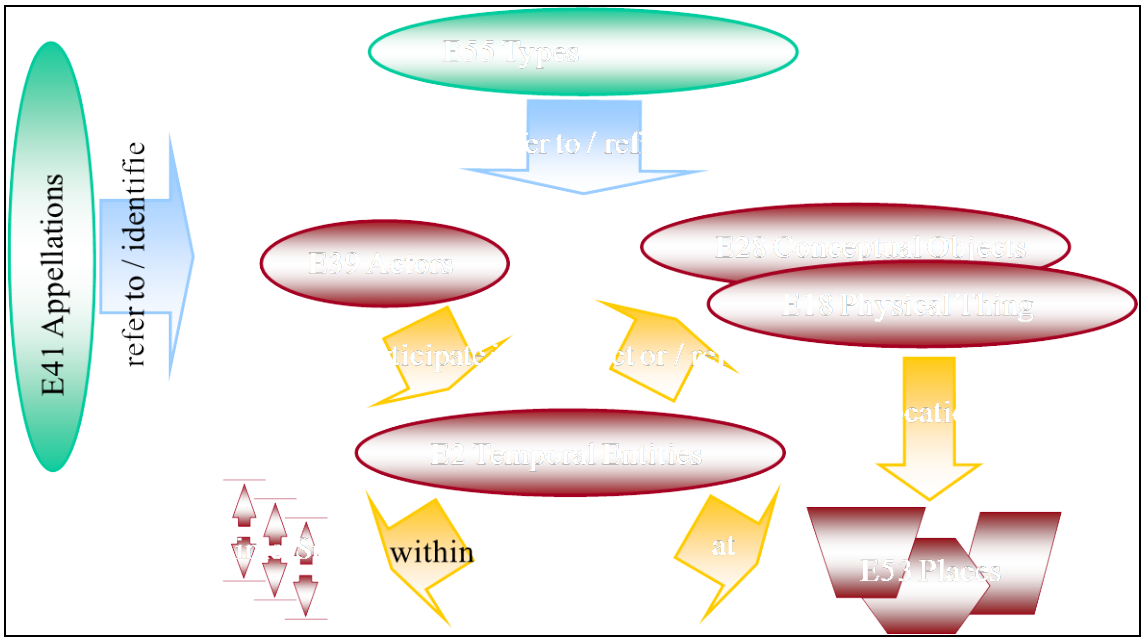


Figure F.10 The main concepts of CIDOC CRM

Dublin Core

Dublin Core[32] metadata are a set of vocabulary terms which can be used to describe resources for the purposes of discovery. Dublin Core is actually born as a set of properties describing metadata about a published resource. Typically, the Dublin Core vocabulary has been adopted and extended by many other domain vocabularies (e.g. the Bibliographic Ontology, or BIBO[33]) to directly refer to the domain objects represented in RDF (thus for describing data, e.g. the dct:date of a book is the date of publication of a book) though, as its properties can be connected to any rdfs:Resource, Dublin Core is also used by specific metadata vocabularies for describing RDF vocabularies and RDF Datasets, thus to provide overall information about data collections as a whole.

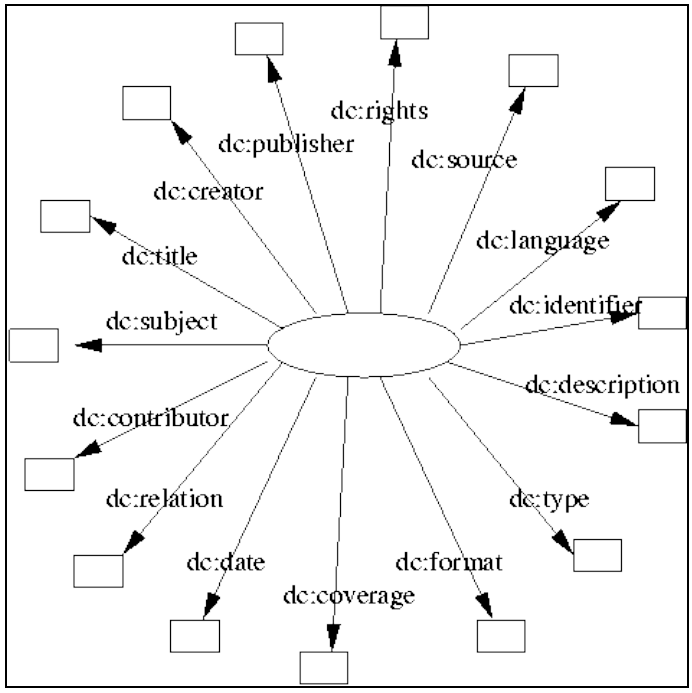


Figure F.11 Dublin Core Elements *(source: http://dublincore.org/documents/dcq-rdf-xml/)*

VOID

The Vocabulary of Interlinked Datasets[34] (**VoiD**) is a vocabulary exclusively focused on providing metadata for describing datasets as a whole. It provides terms and patterns for describing RDF datasets, and is intended as a bridge between the publishers and users of RDF data. VoiD descriptions can be used in many situations, ranging from data discovery to cataloging and archiving of datasets, but most importantly it helps users find the right data for their tasks.

[32] http://dublincore.org/
[33] http://bibliontology.com/
[34] http://www.w3.org/TR/void/

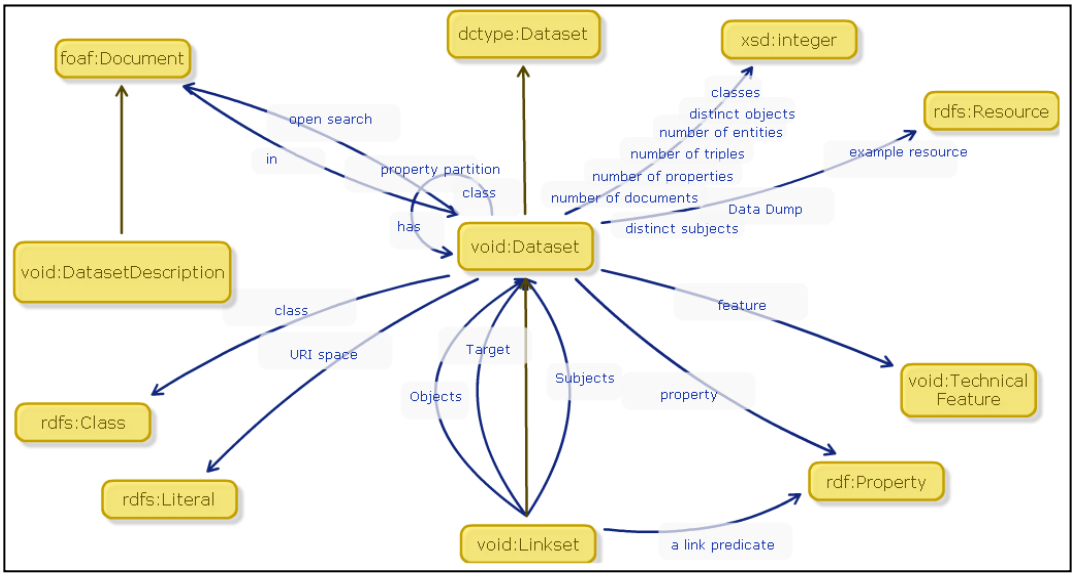


Figure F.12 Overview of VoID RDF schema *(source: http://vocab.deri.ie/void)*

SKOS

Another set of properties which are defined to provide metadata about resources is represented by the range of notation and documentation properties in the **SKOS** (Simple Knowledge Organization Systems) vocabulary[35]. The documentation properties are used to both provide information about the object of the domain which is being described (so these are said to be thought for scheme users), but also about the resource describing the concepts, and not the concepts themselves (so these are said to be thought for the Thesaurus Manager/Knowledge Engineer).

[35] http://www.w3.org/TR/skos-reference/

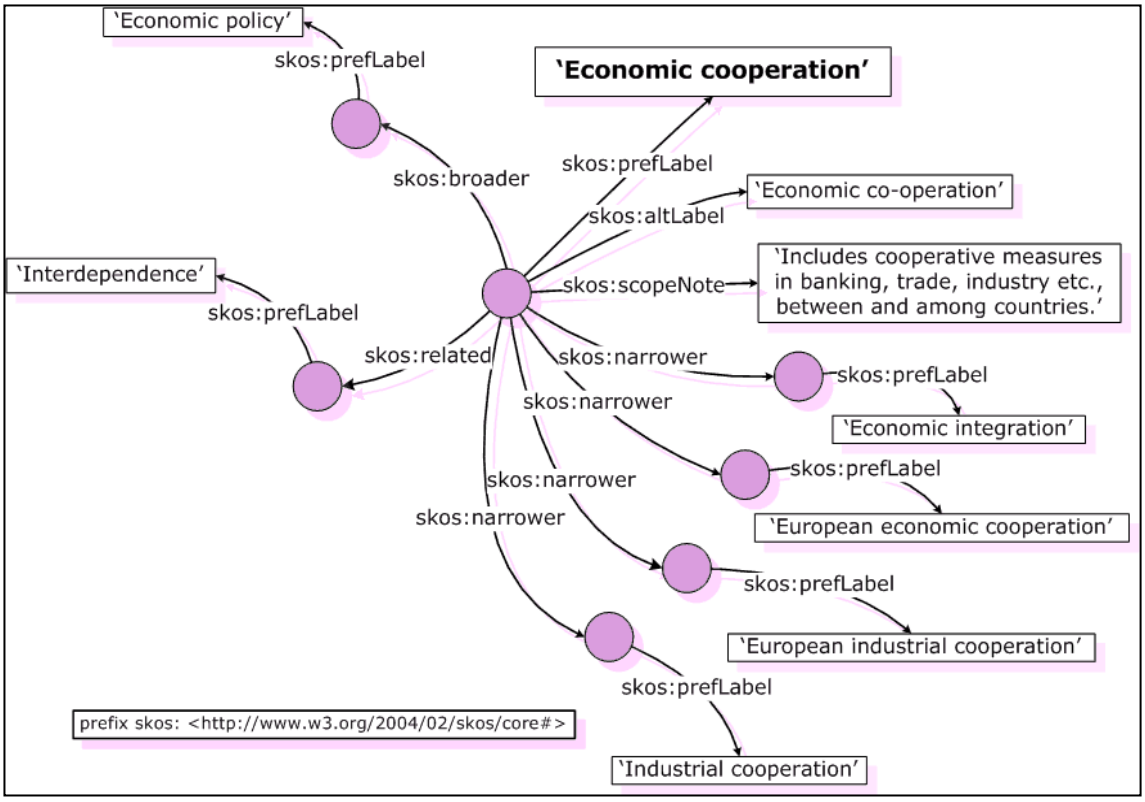


Figure F.13 RDF Graph describing 'Economic cooperation' using SKOS concepts *(source: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/)*

ABC Ontology Model

ABC ontology [1] is considered to be a conceptual model incorporating time, place and events, as well as information more traditionally encoded in metadata. This model has been developed within the Harmony international digital library project. The initial and ongoing goal of work on the ABC model is threefold:

- To provide a conceptual basis for understanding and analyzing existing metadata vocabularies and instances
- To give guidance to communities beginning to examine and develop descriptive vocabularies
- To develop a conceptual basis for automated mapping amongst metadata vocabularies

In particular the ABC vocabulary model is intended a metadata vocabulary per se, but as a basic model and ontology that provides a notional basis for developing domain, role, or community specific vocabularies. To this end it incorporates a number of basic entities and relationships covering the notions of time and objects modifications, agency, places, concepts and tangible objects. The primitive class in the ABC ontology is the *Entity* class. In general it consists of 12 classes and 18 properties and is expressed in RDF format. Its usage lies mainly in the context of the Harmony project, for modelling multimedia content and enhancing the interoperability and exchange of such information within digital libraries and archives.

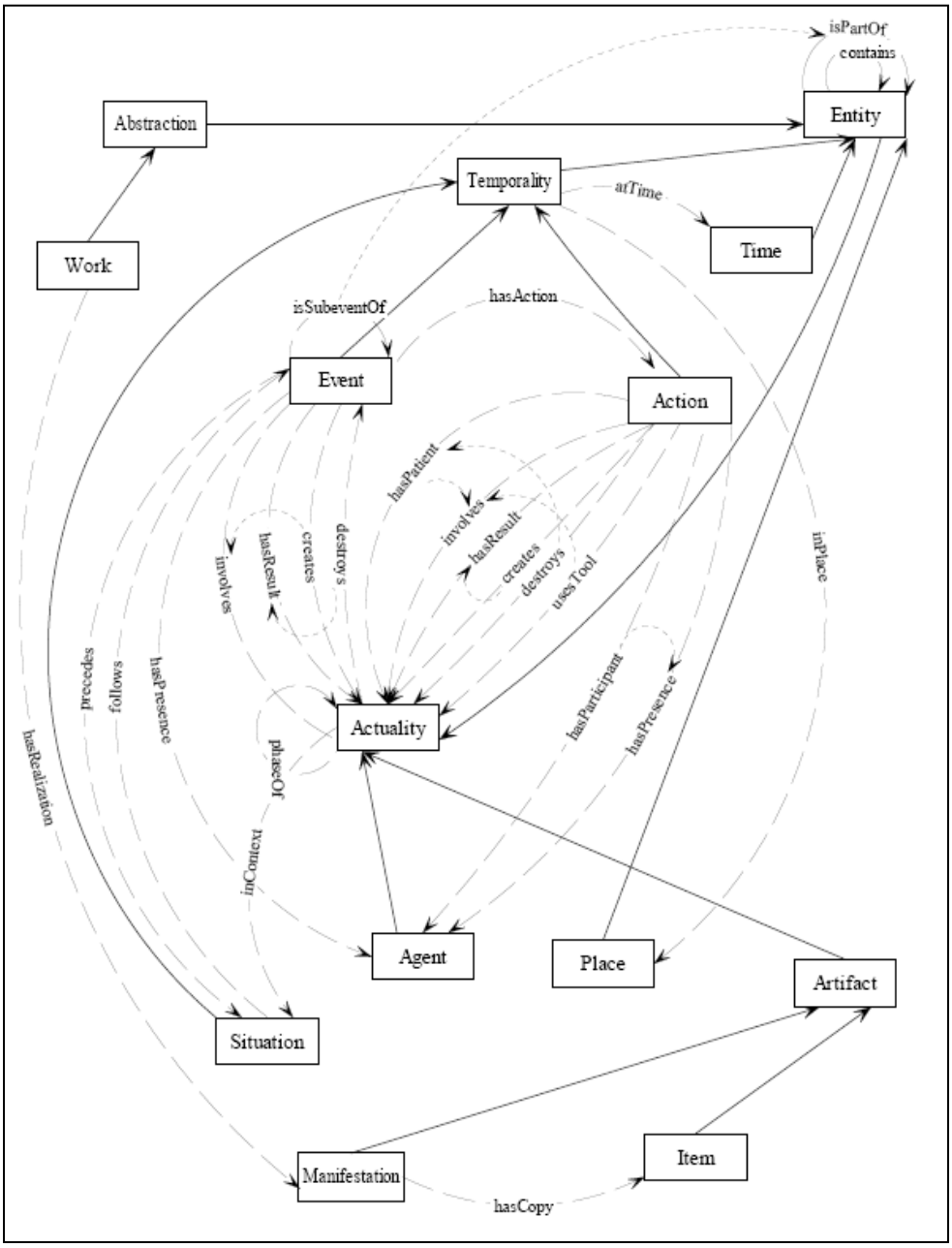


Figure F.14 The ABC ontology (source: The ABC ontology and Model[1])

Global Change Master Directory

The Global Change Master Directory[36] is a directory of Earth Science datasets and related services and tools, many of which are targeted for the use, analysis, and display of the data. The directory is part of NASA's Earth Observing System Data and Information System (EOSDIS). It is actually an entire data portal, as well as a set of descriptions and broader content metadata files using the Directory Interchange Format (DIF). In particular the data are organized using various orthogonal category schemes that can be browsed. Users may perform searches through the Directory's website using controlled keywords, free-text searches, map/date searches or any combination of these. Users may also search or refine a search by data centre, location, instrument, platform, project, or temporal/spatial resolution. The following figure depicts the top level categorization of the datasets in GCMD.

Figure F.15 GCMD top level categories *(source: http://gcmd.nasa.gov/)*

[36] http://gcmd.nasa.gov/index.html