CEOS
COMMITTEE ON EARTH OBSERVATION SATELLITES

Committee on Earth Observation Satellites

Working Group on Information Systems and Services

Interoperability Handbook

Issue 1.1

February 2008

Table of Contents

Document Change Record

Change No.	Date	Changes from previous versions
1	October 1, 2007	Draft sent to WGISS 24 for approval
2	February 1, 2008	Inclusion of a paragraph devoted to the PMH protocol Inclusion of a comment about the use of the Z39.50 protocol by the IDN

Glossary

Acronyms

CCSDS	Consultative Committee for Space Data Systems
CEOS	Committee on Earth Observation Satellites
CORBA	Common Object Request Broker Architecture
CSDGM	FGDC Content Standard for Digital Geographic Metadata
DEM	Digital Elevation Model
DIF	CEOS Directory Interchange Format
DODS	Distributed Oceanographic Data System
DTED	Digital Terrain Elevation Data
EJB	Enterprise Java Beans
EOS	NASA's Earth Observing System
FGDC	Federal Geographic Data Committee
FTP	File Transfer Protocol
geoTIFF	TIFF format for georeferenced data
GIF	Graphics Interchange Format
GML	Geography Markup Language
GOS	Geospatial One-Stop
GrADS	Grid Analysis and Display System
HDF	Hierarchical Data Format
HSM	Hierarchical Storage Management (HSM)
HTTP	HyperText Transport Protocol
HTTPS	HTTP Secured
IDN	CEOS International Directory Network
IDL	Interface Description Language (or Interface Definition Language)
IP	Internet Protocol
ISO	International Organization for Standardization
MIME	Multipurpose Internet Mail Extensions
NSDI	National Spatial Data Infrastructure
OAI	Open Archives Initiative
OAIS	Reference Model for an “Open Archival Information System”
OGC	Open Geospatial Consortium
OGSA	Open Grid Service Architecture
OpenDAP	Open-Source Project for a Network Data Access Protocol

PAIMAS	Producer-Archive Interface Methodology Abstract Standard
PMH	Protocol for Metadata Harvesting (OAI-PMH)
PNG	Portable Network Graphics
SFTP	Secure File Transfer Protocol
SQL	Structured Query Language
SSH	Secure SHell
SVG	Scalable Vector Graphics
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Internet Protocol
TLS	Transport Layer Security
TIFF	Tagged Image File Format
URL	Uniform Resource Locator
WCS	Web Coverage Service
WFS	Web Feature Service
WGISS	CEOS Working Group on Information Systems and Services
WMO	World Meteorological Organization
WMS	Web Map Service
WSDL	Web Service Description Language
WSRF	Web Service Resource Framework
XML	eXtensible Markup Language

Foreword

The need for cooperation among the space agencies providing space borne data and the users of these data has been increasing over the past twenty five years as the result of two factors :

- the development and exploitation costs of space systems ;
- the conviction that planet Earth with all its complexity is to be considered as a World natural heritage.

The Committee on Earth Observation Satellites (CEOS) was founded in 1984 in order to bring up this needed cooperation. And, sometime later, within CEOS, the CEOS Working Group on Information Systems and Services was created to address the more specific problem of sharing space borne data and information.

About CEOS

The Committee on Earth Observation Satellites (CEOS) is an international coordinating mechanism charged with coordinating international civil spaceborne missions designed to observe and study planet Earth.

CEOS is recognized as the major international forum for the coordination of Earth observation satellite programs and for interaction of these programs with users of satellite data worldwide.

The three primary objectives of CEOS are as follows :

- to optimize benefits of spaceborne Earth observations through cooperation of its participants in mission planning and in development of compatible data products, formats, services, applications, and policies ;
- to serve as a focal point for international coordination of space related Earth observation activities ; and
- to exchange policy and technical information to encourage complementarity and compatibility of observation and data exchange systems.

Fore more information about CEOS, see: http://www.ceos.org.

About WGISS

The CEOS Working Group on Information Systems and Services (WGISS) is one of three subgroups supporting the Committee on Earth Observation Satellites.

WGISS promotes collaboration in the development of systems and services that manage and supply EO data to users worldwide.

The present *Interoperability Handbook* is produced and published by WGISS.

Fore more information about WGISS, see: http://wgiss.ceos.org.

Purpose

Work accomplished by the CEOS Working Group on Information Systems and Services (which replaced in 1994 the former CEOS Working Group on Data) is aimed at providing technical responses to the high level requirements expressed under the CEOS resolution and principles given hereafter.

Within WGISS, agencies join their efforts, identifying the concepts and the underlying technology by which these high level requirements could be satisfied.

This document will go more deeply into these concepts, subsumed under a single one: *interoperability*. Interoperability was once defined by CEOS as *the capability of the user interface and administrative software of one instance of a service to interact with other instances of same type of services*[1].

Services are said to be interoperable if they allow for interoperability as previously defined. And systems are said to be interoperable if they are effective implementations of interoperable services.

This document is the CEOS Interoperability Handbook.

It provides recommendations for the development of interoperable systems, drawn from WGISS 10 year experience. It is a handbook, not an academic essay devoted to the theory of interoperability. It is for immediate use by anyone willing to implement interoperable services in a way that preserves their interoperability.

Developing interoperable systems as implementations of interoperable services makes sense only if there is an existing or expected partnership between organisations developing and/or operating these interoperable systems. By definition, interoperable systems do not behave as if they were insulated. They become *de facto* members of systems of systems linking each of such systems to at least one of its *alter ego*. It results that the CEOS Interoperability Handbook is also for communities who may adopt it as guidelines for systems they want to be (or become) interoperable with other ones. The primary community for this handbook, obviously, is the CEOS community. But others may find benefit in applying it.

CEOS Resolution on Satellite Data Exchange Principles in Support of Global Change Research

CEOS members endorse the following principles relating to satellite data exchange in support of global change/climate and environmental research and monitoring and agree to work toward implementing them to the fullest extent possible. Principles for data exchange in support of other data uses beyond global change/climate and environmental research and monitoring will be developed for CEOS endorsement as a next step.

1. Preservation of all data needed for long term global change/climate and environmental research and monitoring is required.
2. Data archives should include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.

1. Strictly speaking, this definition was given by CEOS for catalogue services, which are a particular type of services.

3. International standards - including those generated by the CEOS Working Group on Data - should be used to the greatest extent possible for recording/storage media and for processing and communication of data sets.
4. Maximizing the use of satellite data is a fundamental objective. An exchange/sharing mechanism among CEOS Members is an essential first step to maximize use.
5. Non-discriminatory access to satellite data by non CEOS Members for global change/climate and environmental research and monitoring is essential. This should be achieved within the framework of the exchange and sharing mechanisms set up by CEOS Members.
6. Programs should have no exclusive period of data use. Where the need to provide validated data is recognized, any initial period of exclusive data use should be limited and explicitly defined. The goal should be release of data in some preliminary form within three months after the start of routine data acquisition.
7. Criteria and priorities for data acquisition, archiving, and purging should be harmonized.

CEOS Principles on Data Provision

CEOS members endorse the following principles relating to data provision in support of operational environmental use for the public benefit and agree to work toward implementing them to the fullest extent possible within available resources.

1. Criteria and priorities for data acquisition, processing, distribution, preservation, archiving, and purging should be harmonized to take into account the needs of users of data for operational environmental use for the public benefit.
2. Real time and/or archived data for operational environmental use for the public benefit should be made available on time scales compatible with user requirements and within agency capabilities.
3. CEOS data suppliers should provide (e.g., through the CEOS International Directory Network) easily accessible information about the data and related mission parameters, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.
4. Recognized standards, to be defined and developed in common, including those generated by CEOS Working Groups, should be used to the greatest extent practical for recording/storage media and for processing and communication of data sets.
5. To optimize the use of data for operational environmental use for the public benefit, CEOS Members should establish appropriate data provision mechanisms.
6. Programs should have no exclusive period of data use except where there is a need to provide for data validation. An initial period of exclusive data use should be limited and explicitly defined. The goal should be release of data in some preliminary form within three months after the start of routine data acquisition.

Abstract

While a stand-alone data system is usually built for a community with specific needs, an interoperable system is built as part of a broader system for different communities sharing similar needs. Interoperable systems are actually systems of systems, with each of their components providing at least one part of the full answer expected by their users.

An interoperable system cannot be implemented in the same way as a stand-alone system because it must be built upon the same underlying concepts as are the other interoperable systems with which it will interact. Thus, implementing an interoperable system requires to follow particular guidelines.

These guidelines applied to Earth observation systems are shown in this handbook as recommendations.

The handbook contains an introduction and several chapters.

The introduction is actually an overview of two basic concepts: the concept of data and information, and the concept of interoperability.

Chapter 1 - Data Archiving - deals with the question of keeping data and information in a state that permits to use them even many years after they have been created. This question must be brought up because interoperable systems are always long living systems.

Chapter 2 - Data Access - deals with the question of discovering, locating and accessing data which may be geographically distributed over many places within interoperable systems.

Chapter 3 - From Data to Information - deals with the question of deriving information - which is what the user actually wants - from the data that are accessible through an interoperable system. The process of deriving information from data is conducted by services. Chapter 3 presents the CEOS view on services applicable to Earth observation data.

Chapter 4 - Technologies enabling Interoperability - deals with the question of the underlying technologies that permit systems built upon them to become interoperable. It presents the technologies recommended by CEOS for the development of interoperable systems.

Chapter 4 is followed by a summary of CEOS recommandations for interoperability.

Introduction

WGISS has been experimenting interoperability over more than 10 years. Lessons learnt from these experiments are gathered in the following chapters and expressed in the form of recommendations.

To make these recommendations more accessible, two fundamental concepts need first to be introduced and clarified :

- the concept of *Earth Observation data and information* ;
- the concept of *interoperability*.

About Earth Observation Data and Information

Data is originally the plural form of the Latin word *datum* which means “something given”[2]. Thus, data are “things that have been given”. Though data is strictly speaking a word in the plural form, it is accepted in Modern English to use it also as a word in the singular form.

There is a link between data and information. For instance, clouds in the sky are data from which one may derive an information like “rain is coming”. Information is actually the result of a mental process on data as “things that have been given”. This mental process can also be said as the *process of becoming informed*[3].

There are many instances of data. As such, an object in a museum may be seen as data since it imparts some information. However, the only data that are considered in this handbook are data given as encoded expressions of natural or anthropic phenomena related to planet Earth observed through instruments, many of them being parts of satellite payloads.

These data are called *Earth Observation data* (in short : *EO data*). In general they are given as digitally encoded expressions because they can be submitted to computers only under this form. But they may also be given as traditional documents like printed books or photographs.

Data may be seen as expressions of pieces of a phenomenon. These pieces are selected because they provide a simplified but appropriate view of the phenomenon in a given context, through a modeling process. For instance, the wind over some area may be modelled as a field of vectors giving its local strengths and directions.

A simplified view of a phenomenon is sometimes called a *feature*. Categories of features, called *feature types* may be identified. For instance, the Golden Gate bridge in San Francisco is an instance of the feature type *bridge*. Features and feature types are categorized by *attributes*. The name of the bridge, its height, the material used to construct it are attributes. When a phenomenon is viewed as the variation of a function within a spatiotemporal domain (for instance the variation of a vector), its view is also called a *coverage*.

One peculiarity of EO data, in comparison with other types of data (like astronomical data), is that they always refer to some position on the Earth. This is due to the fact that a

2. The word *date* has a similar origin: it comes from Medieval Latin *data littera*, first words of the expression used in the past for giving the date when a letter was issued.

3. In some sense, the data “clouds in the sky” contains the information “rain is coming”. This explains why, very often, data and information are considered identical. For a thorough analysis of the concept of information, see: Information and Information Systems, Michael Buckland, 1993, ISBN 0-313-27463-0.

phenomenon appears at some place on the Earth. This place generally has a 2 or 3 dimensional spatial extension. In addition to its spatial domain, a phenomenon can also have a temporal domain given, for example, by a beginning date and an ending date. This makes EO data intrinsically complex data.

Data having similar characteristics (for instance wind vectors as measured by a single instrument) may be grouped. The result of such a grouping is called a *dataset*. On a computer, a dataset is often handled as collection of files. A dataset may be processed or formatted for future use. In this case it is called a *data product*.

An instrument rarely observes directly the phenomenon of interest. It usually only observes some basic phenomenon. The data relevant to the phenomenon observed by the instrument are later transformed by application of an appropriate algorithm into data which are then taken as data relevant to the phenomenon of interest. For instance, data relative to wind vectors over the ocean are derived from the data relative to an another phenomenon, namely the response given by a radiometric signal after reflection on the ocean surface. The response depends on the profile of the ocean waves which, in turn, permits to derive the wind vectors. Several algorithms may be applied in cascade to an initial data product to get the desired final product. It results that algorithms should be attached to the data since they contribute to the description of the phenomenon of interest.

In the past, CEOS has identified five levels of data products. They are recalled hereafter :

- **Raw Data** - Data in their original packets, as received from a satellite.
- **Level 0** - Reconstructed unprocessed instrument data at full space time resolution with all available supplemental information to be used in subsequent processing (e.g., ephemeris, health and safety) appended.
- **Level 1** - Unpacked, reformatted level 0 data, with all supplemental information to be used in subsequent processing appended. Optional radiometric and geometric correction applied to produce parameters in physical units. Data generally presented as full time/space resolution. A wide variety of sub level products are possible.
- **Level 2** - Retrieved environmental variables (e.g., ocean wave height, soil moisture, ice concentration) at the same resolution and location as the level 1 source data.
- **Level 3** - Data or retrieved environmental variables which have been spatially and/or temporally re-sampled (i.e., derived from level 1 or 2 products). Such re-sampling may include averaging and compositing.
- **Level 4** - Model output or results from analyses of lower level data (i.e., variables that are not directly measured by the instruments, but are derived from these measurements).

Each level represents a step in the abstraction process by which data relevant to physical information (raw, level 0, level 1) are turned into data relevant to geo physical information (level 2, level 3), and finally turned into data relevant to thematic information (level 4).

About interoperable services and systems

The process of transforming data may be provided as a *service*. The ISO 19119 standard[4] defines a service as the distinct part of a functionality that is provided by an entity

4. ISO 19119:2005 - Geographic information - Services

through *interfaces*. The same standard defines an interface as a named set of *operations*, and defines an operation as the specification of a transformation. An operation has a name and a list of input and output parameters. A service may consist of one single operation. It may also consist of the chaining of several operations.

The above definitions do not put any restriction on the way operations are implemented. Operations can be - and often are - manual operations. However, in this handbook we are considering only operations that are implemented on computers.

Different entities may implement in different ways the same service. In this case, each entity makes the service available through one of its own set of interfaces.

Two or more implementations of the same service are interoperable if the invocation of one implementation entails the invocation of the other implementations. The values of the output parameters of the last operations in each service chain are merged and make the complete output of the service that was first invoked.

Interoperable systems are systems that implement interoperable services.

To clarify the above statements, let us consider the following small catalogue service with one single operation:

operation name	get_metadata
input parameter name	time_period (mandatory)
input parameter name	keyword (optional)
input parameter name	metadata_description_type (mandatory)

The catalogue service operation retrieves descriptions of existing EO data products (these descriptions are also called *metadata*).

Descriptions are retrieved only if the corresponding data products have a temporal domain that intersects the time period given by the *time_period* parameter. A time period may be expressed, for instance, by the indication of a beginning date and of an ending date in the Gregorian calendar. Additionally, if the corresponding data products can be characterized by a keyword, the only descriptions that are retrieved are those whose corresponding data products match the keyword given with the *keyword* parameter. This constraint applies only if there is a *keyword* parameter. And finally, among the various descriptions that are retrieved, only the description with same type as the type given with the *description_type* parameter are definitively retrieved. Common description types are *brief*, for short descriptions, and *full* for comprehensive descriptions.

Such a catalogue service may be implemented in many ways. For instance data product descriptions may be stored in a relational database, retrieved using SQL requests built from the get_metadata operation input parameters and sent to the service user as a text file.

If such an implementation is to become interoperable with another one, several conditions must be taken into account :

- since the request by which one service is invoked must be forwarded to the other one, both implementation sites must be connected in some way;
- the incoming service request must be encoded by the first service implementation according to an encoding mechanism recognized by the second implementation ;
- conversely, the descriptions retrieved by the second service implementation must also be encoded according to an encoding mechanism recognized by the first implementation.

For instance, if both services are implemented as HTTP servers, and assuming there is an Internet connection between the server sites, the HTTP protocol structures could be used for the transportation of requests/replies between them. Both implementers will need to agree on common encoding schemes for the messages that will flow from one implementation to the other. They may select some encoding schemes amongst well established standards. For instance, dates may be encoded according to the CCSDS recommendations for date/time encoding, and metadata that are retrieved by the servers may be encoded according to the CEOS Directory Interchange Format (DIF).

Beyond the technical solutions which may be elected to achieve interoperability between the two service implementations, the point to emphasize is that interoperability requires the definition of additional *private* service interfaces by which systems implementing services privately provide interoperability services to each other. These private services (in our example, the ability for one service implementation to forward *get_metadata* requests to the other one) extend the public service interfaces by which the implemented service is publicly seen by the ordinary user. This is one reason why making an already existing service implementation fully interoperable with other ones is always a challenge.

The first of the above conditions deals with interconnection.

The second and third conditions deal with structure and encoding, i.e. with *syntax*.

There is a fourth condition, dealing with *semantics* : implementers need to share a common understanding of the service they want to implement.

For instance, implementers of the above catalogue service would need to agree on a common definition of data product descriptions, both in their brief form and in their full form. Such a definition will become a *de facto* standard for the two implementers. They will also need to agree on the handling of the optional *keyword* parameter value : should it be checked against a list of valid keywords or just handled as free text ? One implementer may also not completely implement the service. For instance, he may not implement the optional *keyword* parameter part of the catalogue service because the data product descriptions on his side all belong to a single category and discrimination by keyword does not make sense. He must however be prepared to semantically understand and syntactically recognize the *keyword* parameter in an incoming *get_metadata* request, in order to ignore it without rejecting the request.

And there is finally a fifth condition: *time*.

Most interoperable services are designed as long term responses to fundamental needs common to many user communities. Discovering data, locating data, accessing data, using data are examples of such fundamental needs of EO data users. Thus, most interoperable services are long living services, often defined by standards, and all objects

handled through interoperable services must be set up with a perspective of long term use and reuse.

In some sense, interoperability should be considered as goal to reach, rather than a property a system may exhibit or not. In the past, CEOS has defined levels of interoperability. They are recalled hereafter :

- **Level 1.** A user accessing one service is routed directly to another related service. For example a user identifying a relevant data set in a directory is routed directly to the associated inventory service, or a user of one inventory service is routed automatically to another related inventory to extend a search. At this level, the user interface differences between the routed services are not hidden from the user.
- **Level 2.** This is similar to level 1 except that a user accessing one service is routed directly to another related service, and context information is passed at the same time. The context information typically includes information concerning the user, their interests, and the activity on the original service (for example a search query) which can be interpreted by the new service to assist the user in making further queries. For example search query criteria entered at one site can be passed to another, translated and loaded as a search query for the user to activate. At this level the user is still exposed to the differences in operation between the two services, but related information can be shared.
- **Level 3.** At level 3, the user is hidden from the specific operations of each service. A service request is routed automatically to one or more services, and the results from each returned to the original access point. At this level (and levels 1 and 2) each service is likely to use a different data model and the query may need translation to reflect these differences. Similarly, the query results may also reflect the individual service data models and may or may not be translated before they are presented to the user. At this level the user is no longer exposed to the differences between service operation.
- **Level 4.** At the highest level of interoperability a single data model is assumed to apply to all services within an interoperable system. Thus a query entered by a user can be executed using a single distributed database. This level of interoperability permits direct database operations such as union and joins between metadata held by different catalogue services and for operations. The interoperable system is thus seen as a single database system by the user.

Several interoperability levels may coexist in a system made of interoperable systems :

- Some systems could just provide level 1 interoperability : the user of one system may be linked by this system to another system via a URL address but he will have to rebuild himself his request within the second system interface. This is the weakest level of interoperability but it is often the easiest way to make existing systems interoperable with other ones.
- The small interoperable catalogue service system example would provide interoperability at level 3, since the get_metadata request is automatically routed to the other site and the results sent back to the user.
- Some organizations could locally agree on a common data model for a given discipline and offer level 4 interoperability for data relevant to this discipline through a distributed database.

Chapter 1

Data Archiving

1.1 Introduction to Data Archiving

The primary purpose of data archiving is to preserve data over time. Preserving data over time consists in holding data in repositories in a way which enables data to be available, retrievable and usable even many years after the time it was produced.

Because data archiving is a long-term process, it is necessarily impacted by environmental changes. These changes, external to the archive, cannot generally been easily circumvented because the archive has no or little control over them. For instance the manufacturer of a storage device used by an archive may decide at some time that he will no longer maintain this device. The archivist must then select another device from the same or a different manufacturer, implement it in the archive, and organize the transfer of the data from the obsolescent device to the new one. Or, a format used to record data must be changed because the format is a proprietary format whose definition is not publicly available and because the software which reads this format is no longer maintained. In this case, it may be necessary to reformat - which means: transform - the whole dataset.

Data archiving is not only a long-term process, it is also a complex process with possibly many partners: data providers supplying data to the archive, data users willing to use the archive, archive managers organizing the archive, other archives with which interoperability may be sought. All these partners should use the same vocabulary.

In the following paragraphs, we will describe some of the elements contributing to archive interoperability:

- a shared vocabulary

  In a long-term process as complex as data archiving, a same vocabulary shared by archivists makes interoperability easier.

- technology for archives

  Data physically resides on storage devices. Storage device selection is difficult because there is a wide choice of existing storage devices and new ones are becoming available continuously.

  Data is recorded on a storage device according to some data format. Many data formats exist. They are either designed by sensor developers or by user communities for specific applications.

  Reading and writing data from/to a storage device requires storage management systems. For big archives, a virtual storage methodology is sometimes adopted, i.e. a combination of disks, tapes and hierarchical storage management (HSM) software, making the archive look as if it were a wide file system.

- organizing archives

  Archives can be broken up into several functional components. Examples are the data ingest component devoted to the process of ingesting data retrieved from

data providers, the data storage component devoted to the process of maintaing the data on storage devices, the data management component devoted to the organization of the data in the archive, the data access component, devoted to process of data retrieval from the archive by data users.

1.2 A shared Vocabulary

CEOS has agreed to use a subset, suitable for digital archives, of the "Glossary of Archival and Records Terminology", published by the Society of American Archivists.

This subset is available at: http://wgiss.ceos.org/archive.

The full glossary is available at the same address and at: http://www.archivists.org/glossary/index.asp[5].

Recommendation 1

Interoperable archives archivists should use a same glossary of terms and definitions applicable to data archiving.

Rationale

CEOS agencies share the same concern - long term data preservation - and they share the same problems, like the handling huge amounts of data, or the organization of data migration from one storage device to another one. CEOS agencies feel that using a shared vocabulary in the complex field of data archiving contributes to a better understanding and easier solving of all these problems.

1.3 Technology for Archives

Archiving makes use of many components:

- data storage devices (media);
- software and firmware associated with the storage devices;
- application software that implement the data format;
- format, schema or layout of the data.

Strength and weakness of each of these components must be carefully analyzed before they are selected, and later periodically reviewed, because they impact the long-term reliability of the archive.

The three firsts of the above mentioned components apply to the materials used for reading and writing data. The decision for implementing such a material in an archive or for removing it results from many interlinked factors, like:

- data read/write speed;
- data storage device capacity;
- technology implemented in the material;
- material physical size;
- acquisition and maintenance costs;
- material expected commercial lifetime;

5. The scope of the full glossary is much broader than that of digital archives.

- commercial strength of the material.

As new technologies and product lines emerge, all these factors must be reviewed and this review will lead the archivist to decisions that may have serious consequences for the archive. For instance, an archive may decide to continue to operate obsolescent materials because their cumulative maintenance costs are still lower than the migration costs, or an archive may keep obsolete storage devices, like digital tape readers, because it is asked by its authority to be able to rescue data that are still recorded on this kind of devices[6]. Conversely, an archivist may decide to implement a new storage device, despite the potential data migration costs because it will increase the archive responsiveness to the user which may be felt as an important objective for the archive.

In order to help archivists to make well-argued decisions, CEOS has experimented a decision support tool that may help assessing some of the above mentioned selection factors. This tool is available at:
http://wgiss.ceos.org/archive/archive.xls/MediaToolVersion1.0.xls.

It should be viewed only as a storage device selection support tool, no more than providing a methodology for selecting a data storage device. It does not give solutions. These are the archivist's concern.

The remaining components are the data format and the application software that is used to control and read/write the data from/to the data storage devices. From the perspective of a long term archive, both data format and application software that implemented this data format must also stand the test of time. If the data format is complicated, and consequently the software complex, the task will become much more difficult as new technologies come up in the future. Even a simple operating system level change or a data format version change may impact the ability of the archive to guarantee data accessibility. In some cases, even the survivability of the archive may be threatened. CEOS has periodically surveyed the data formats that are being used by the CEOS agencies. These documents are available at: http://wgiss.ceos.org/archive/archive_docs.html.

All the above mentioned components, whatever they are pure technological or physical components, make an archive behave like the links in the chain between data and data users. The life expectation of the data dwelling in the archive relies on the strength of this chain, i.e. the strength of its weakest link.

1.4 Organizing Archives

Archive model

An archive is an organization which has the responsibility to preserve information for a community. It consists of persons, who operate the archive on a day-to-day basis, and of systems. The external environment of the archive consists of:

- producers who provide the information to the archive;
- consumers who search and retrieve information from the archive;
- managers who define the policy to be followed by the archive.

Producers and consumers may be persons or archive client systems. Managers are members of the authority which defines the rules applying to the archive. For instance, an academic institution may request an archive to preserve all information related to some discipline and restrict the distribution of this information to authenticated members of a designated community.

6. Some CEOS agencies have graveyards for such devices.

Information provided by a producer to the archive can be made available to the consumer in a way very different from the producer's contribution. For instance, while a producer may ship CDROMs to the archive, the consumer may see the same information through a relational database. This illustrates that, whilst the semantics of the information (i.e., its meaning) to be preserved by the archive always remains the same, its syntax (i.e. its structure) may change significantly as it flows through the various functional areas of the archive.

The Consultative Committee for Space Data Systems (CCSDS) has defined a general model for archives. This model is called the "Reference Model for an Open Archival Information System" (OAIS) and is available at: http://public.ccsds.org/publications/archive/650x0b1.pdf[7]. The term "Open" means that the recommendations contained in the standard have been discussed in open forums.

The standard describes a "reference" model providing concepts, terminology, techniques as frameworks for the implementation of a long-term information preservation process. One of the OAIS key concepts is the "Information Package", i.e. a conceptual container for the "Content Information" (the information to be preserved) and the "Preservation Description Information" (the information needed to preserve the "content information"). An OAIS compliant archive basically handles "Information Packages".

Recommendation 2

An archive system should comply with the Reference Model for an "Open Archival Information System" (OAIS).

Rationale

The reference model is suitable for almost every archive, including non digital archives like book libraries. It helps understanding what is required to preserve information for the long term. It breaks up an archive into several well identified functional areas, which makes the design, development and maintenance of an archive much easier and minimizes impacts of environmental changes. With the concept of "Information Package", it provides a sound information model.

Data appraisal

Not all data deserve archiving. For instance, an archivist may not want to archive data whose quality is questionable. But he may change his mind if it turns out that this data is unique and cannot be reproduced. The archivist may then archive the data with a flag indicating that the data quality could not be assessed.

The decision of archiving or not archiving data depends on the answers given to a number of questions about the data. Examples of questions are:

- Does the data fit with the archive policy?
- Are the data users known?
- Can the quality of the data be assessed?
- Is the data unique?
- Is the data documented?
- Is the archiving cost acceptable?

There are tens of such questions. They may be caracterized by several topics, like compliance to the archive policy, thematic description of the data, data quality, restrictions of

7. It has become an ISO standard: ISO 14721.

use, etc. It may therefore be helpful to gather all of them in a structured questionnaire for the use of archivists.

Recommendation 3

An archive should use a questionnaire to help archivists in appraising data that are candidates for archiving.

Rationale

Deciding if data must be archived or not must be the result of a rigorously conducted process. A questionnaire may be a supporting tool for this process.

CEOS agencies have experimented such a questionnaire available at: http://wgiss.-ceos.org/archive/index.html[8].

The producer Interface

Interactions between archives and information producers are always complex. Even if the information producer just provides CDROMs, there must be an agreement between the archive and the information provider about about a number of subjects, like the recording data formats, the pace at which the provider will supply the CDROMs, the documentation that describes the information on the CDROMs (for instance, the archive may accept only non-proprietary text formats). The producer may also want the information on the CDROMS be searched by the consumer file by file, and not just by CDROM identifier. Archive and information provider must also agree on quality controls to be applied during the ingestion process. For instance, if a CDROM is unreadable, the information producer may be asked by the archive to provide a new one.

The Consultative Committee for Space Data Systems (CCSDS) has analyzed all the possible interactions between archives and information producers and has published the "Producer-Archive Interface Methodology Abstract Standard" (PAIMAS) available at: http://public.ccsds.org/publications/archive/651x0b1.pdf[9].

Recommendation 4

An archive should apply the "Producer-Archive Interface Methodology Abstract Standard" (PAIMAS) standard to define its interactions with information producers.

Rationale

The recommendation breaks up the information ingestion process into different phases. Each phase has a dedicated objective (i.e. with expected results) and identifies the actions which must be carried out to reach the objective.

The purge alert

It sometimes happens that information is removed from an archive and destroyed. One reason can be that this information is no longer useful because it is available elsewhere with a higher quality. But another archive may be interested in retrieving this information because, for instance, it could not handle the higher quality information due to limited bandwidth. To make CEOS agencies aware that information is about to be removed, CEOS has set up the "purge alert" service. The purge alert service is available at: http://wgiss.ceos.org/purgealert/index.html.

Recommendation 5

CEOS agencies should use the purge alert service should be used before information removal from archives.

8. This tool has been designed by the U.S. Geological Survey.
9. It has become an ISO standard: ISO 20652.

Rationale

The purge alert service enables archivists to advise archives that data will be destroyed and offer these data to other archives before data destruction.

Chapter 2

Data Access

2.1 Introduction to Data Access

A user may want to access data, i.e. to retrieve data from the various repositories where they are stored. This implies that the user knows how to identify the data, locate the corresponding data repositories, and interact with these repositories to retrieve the data.

Generally, the user does not know directly the identity of the data but he is able to provide some characteristics of the data. The identity of the data may be obtained if there exist a list showing together existing data identities and their corresponding data descriptions to which the user provided characteristics may be compared. Such a list is called a catalogue[10]. A catalogue may be as simple as a printed document. It may also be organized as a complex computer system with sophisticated search and retrieve functionalities[11].

While the characteristics by which the user knows the data may be more or less precise, the descriptions that are found in the catalogue must be as precise and exhaustive as possible (the catalogue is actually a registry containing official records) and identify without ambiguity the data of interest to the user. In particular, the data description from the catalogue must contain the information needed by the user to request and get the data. Semantically, this information is the path which leads to the record(s) containing the data. It can be a simple URL in which case there is no complicated interaction (at least from the user point of view) with the repository hosting the data. It can also be a complex data order and data retrieval specification (sometimes including the payment of fees). In this case, the interaction with the repository may be complicated and need several more or less automated steps: contact the data custodian, order the data, pay the fees, retrieve the data with the appropriate protocol.

In the following paragraphs, we discuss in more details the three basic steps that are needed for data access:

- describing data
- searching data
- retrieving data

2.2 Describing Data

About Data Description

Describing Data is providing information about data.

10. According to the American Heritage Dictionary, a catalogue is *a list or itemized display, as of titles, course offerings, or articles for exhibition or sale, usually including descriptive information or illustrations.*

11. A catalogue system may comprise several single catalogues interconnected via an interoperability protocol. An example of such a catalogue system is the eoPortal Catalogue (http://catalogues.eoPortal.or g) whose catalogues are interconnected via the CEOS Catalogue Interoperability Protocol (CIP).

Examples of information about data are:

- the location of the data repositories
- the title by which the data is known
- the data content and the data purpose
- the data spatial and temporal domains

The technical word for data descriptions is *metadata*.

Describing data through metadata consists in providing views on the data. There may be many different views on data. Some of them may be very high level views, just providing information about the existence of data through a title and a short abstract. Others may go into details like, for instance, the various ways the data may be retrieved from each of its repositories.

Metadata provide a mediation between a user searching information and the repositories where the data containing the searched information may be found. In this sense, metadata act as data surrogates: while handling their metadata, the potential user can check if the data would satisfy his needs without having to physically extract them from their repositories.

This has some implications on the way metadata must be constructed:

- metadata must describe data in a way understandable by the metadata user; the words that are used in metadata must be part of the vocabulary expected by their users (vegetation specialists do not refer to the same concepts - and thus do not use the same words - as ocean specialists).
- there may be several metadata for the same data, depending on the metadata target audience; metadata for laymen are different from metadata for experts; if the same data are of interest in several disciplines, there may be several metadata for the same data, each one constructed according to the peculiarities of the related discipline.
- conversely, data that differ only at some level of details may be organized as a single collection; description would apply at collection level, not at data level.

In most cases, metadata may be seen as a document linked to the data (or data collection) they describe. Such a document consists of chapters and of paragraphs, each one devoted to a particular view on the data. It mainly contains printable text but may also contain (or be linked to) graphical representations like images. Metadata that are digitally encoded may be electronically exchanged. XML has become a very common technology for the digital encoding of metadata.

A community may agree on a metadata template. Such a template indicates the way metadata should be constructed by data providers and data custodians for the members of the community, i.e. the template gives the names and the topics of the chapters and paragraphs the metadata should or may contain. A large community may define a very general metadata template. This template may be further reused and profiled for their specific needs of smaller communities that are part of the larger one.

Metadata templates may become international *de jure* or *de facto* standards and there may also be metadata encoding standards.

There should always be a link from the metadata to the data they describe and, conversely, there should always be a link from the data to their metadata.

Metadata may be recorded sequentially in a single file which can be printed or displayed. Metadata may also be gathered in a database (usually, a community that owns hundreds or more metadata makes them available only through a database against which the user applies queries).

CEOS recommended metadata standards

Recommendation 6

The CEOS Directory Interchange Format (DIF) should be used for the description of science Earth Observation data collections.

Rationale

The DIF, which was created through international consensus has become the *de facto* standard for describing Earth Observation datacollections across many science disciplines. The DIF not only standardizes the metadata structure but also the metadata content: for example, the specification of personnel or the specification of dates. It is developed and approved through the CEOS Interoperability Forum.

The DIF is particularly suited for the descriptions of data collections, i.e. groupings of data that have commonality. Usually, a data collection comprises a series of datasets, for instance all datasets pertaining to the same scientific experimentation. The DIF can also be used for the description of single datasets.

Robust software is available for writing DIF compliant metadata. DIF metadata are now encoded as XML texts and may easily translate to other metadata standards like Dublin Core, ISO 19115, CSDGM.

Recommendation 7

The CEOS International Directory Network (IDN) should be used to host DIF compliant metadata.

Rationale

DIF compliant metadata may be encapsulated into the CEOS International Directory Network (IDN). The IDN is an operational database for DIF compliant metadata with powerful search and retrieve capabilities. It was developed to assist researchers in locating and discovering information on available datasets. It currently holds more than 18,000 metadata[12].

Recommendation 8

Description of geographic data should be done in conformity to the ISO 19115 metadata standard.

Rationale

Whilst there is a wide overlap between them, Earth observation data and geographic data should be considered distinct. Geography is the *study of the earth and its features and of the distribution of life on the earth, including human life and the effects of human activ-*

12. Limited views to the IDN metadata content, for instance, a view to an organization's metadata, are offered through dedicated IDN portals. Hence an organization can easily build a dedicated catalogue containing DIF compliant metadata without having to start from scratch. More information about the IDN: http://idn.ceos.org.

ity[13]. Thus, geographic information may be very complex, sometimes much more complex than ordinary scientific information.

It results that the descriptive power of a metadata format like the DIF may not be sufficient for geographic data. “ISO 19115 - geographic information -- metadata” is the current international standard for the description of geographic data.

Note

A community may develop an ISO 19115 metadata standard profile. A profile is an extension of the standard[14]. A community may need to develop a profile in order to increase, for its particular needs, the expressiveness of the standard. Conformity to a registered profile entails conformity to the ISO 19115 metadata standard.

2.3 Harvesting Metadata

About Data Metadata Harvesting

Data descriptions - or metadata - can be created from scratch by a metadata editor and individually inserted in some metadata repository (e.g. a metadata database).

But metadata can also be created by the data owner or data custodian, in which case it may be more convenient to retrieve these metadata from their dedicated repositories and not create them again. This process of retrieving metadata is called metadata harvesting.

Metadata harvesting implies that data owners or data custodians agree to expose their metadata (or a subset of them) for later retrieval by some metadata repository.

The metadata exposed by the data owner or data custodian are contained in some repository acting as a metadata server which can be queried for metadata retrieval according to some harvesting protocol used by a client called the “harvester”. Interoperable harvesting, and thus metadata exchange, is achieved when such a protocol is shared by several metadata servers and metadata harvesters

With the Protocol for Metadata Harvesting (PMH), developed through the Open Archives Initiative (OAI), it is possible to harvest metadata content without knowledge of the metadata server architecture.

An OAI-PMH compliant metadata server holds metadata, each one being identified by a unique identifier. Metadata can be recorded under various formats (there may be more than one record for the same metadata, e.g. Dublin Core and DIF). A metadata server may group metadata for the purpose of selective harvesting.

The harvester retrieves metadata from the metadata server as XML formatted texts. The harvester may retrieve all the metadata exposed by the metadata server or only the subset identified by one of the metadata groups set up by the server. The harvester may also retrieve top level information about the metadata server and about the metadata held by the server.

All OAI-PMH requests are expressed as HTTP requests[15].

13. Definition given by the *American Heritage Dictionary.*
14. The standard gives rules for creating a profile.
15. More information about the OAI-MPH can be found at : http://www.openarchives.org/.

Recommendation 9

The Protocol for Metadata Harvesting (PMH), developed through the Open Archives Initiative (OAI) protocol should be used for metadata harvesting.

Rationale

The Protocol for Metadata Harvesting (PMH), developed through the Open Archives Initiative (OAI) helps reduce many barriers for the exchange of metadata. By providing a standardized protocol for retrieving metadata from a network server, it is possible to harvest metadata content without knowledge of the server's database architecture. Simplifying the harvesting process is critical because it will reduce the time it takes data producers to deliver metadata to the IDN.

By using a harvester client capable of issuing OAI-PMH queries to an OAI-PMH server, all or portions of an external metadata database can be retrieved in a rapid and efficient manner.

The IDN has developed an OAI-PMH compliant metadata harvester client for interoperating with its partners.

2.4 Searching Data

About Data Search

We assume that the user who wants to retrieve data at least knows some characteristics of the data. These characteristics are generally based on the user's knowledge and expectations. For instance, a user interested in the radiation flux received or emitted by the Earth at top of the atmosphere over Europe may characterize the data he wants to retrieve by the words “Earth Radiation Budget” and the words “Europe”: “Earth Radiation Budget” is the name of the scientific discipline which studies the radiation received and emitted by the Earth; “Europe” is the name of a geographic region but “Europe” could have also been defined (perhaps with less ambiguity) by the geographic coordinates of a polygon surrounding the geographic region known to be Europe. Hence, the data characteristics provided by the user is made of a keyword (“Earth Radiation Budget”) and a spatial extent (given either by the geographic name “Europe" or by a sequence of geographic coordinates). Both the keyword and the spatial extent characterize the data. Hence, keyword and spatial extent are implicitly linked by the Boolean operator AND. If the user is only interested in short wave radiation, he may want to add the keyword “Short Wave Radiation” and link it to the previous ones with the Boolean operator AND.

The user must then find a catalogue that understands these data characteristics. Obviously, if no catalogue knows the keyword “Earth Radiation Budget”, the user will never find what he wants, even if the catalogue contains very detailed descriptions of data relevant to the Earth radiation budget but unfortunately only known by the catalogue only under the keyword “Atmospheric Radiation”. The consequence is that a catalogue should always be queried with the keywords (or, in a broader sense, with the vocabulary) the user is used to use within his community.

The technical name of the index of all terms in use within a community is “thesaurus". A thesaurus is a structured index, showing all the relations between terms (e.g.: a term may be the synonym of another term, a term may a specialization of another one, etc.), thus

suggesting the possible navigation schemas through the terms[16]. A thesaurus containing geographic terms (names of countries, cities, etc.) is sometimes also known as a "gazetteer".

CEOS Recommendations for Data Searching

Recommendation 10

A catalogue should always link each of the keywords available for data searching to the appropriate thesaurus.

Rationale

Since they are the most common way to browse through catalogues, keywords must be very strictly controlled in order to avoid as much as possible ambiguity. If there is an ambiguity in the definition of keywords, the user may retrieve unwanted data or, which is worse, not find valuable data.

Note

Keywords from different thesauri may be used in a single search.

Recommendation 11

For science data, catalogues should use to the far extent possible the keywords defined by the CEOS International Directory Network.

Rationale

The keywords defined by the CEOS International Directory Network are well accepted by most of the scientific communities because they have undergone a rigorous selection process under the auspices of top representatives of these communities.

Note

Using the keywords defined by the CEOS International Directory Network does not imply for a catalogue to comply with the CEOS Directory Interchange Format for the metadata.

2.5 Retrieving Data

About Data Retrieval

A catalogue does not necessarily directly give access to the data but at least it gives the guidelines the user should follow to access the data. These guidelines can be found in the metadata[17]. To access the data, the user must go through a special entity: a data distributor.

A data distributor may distribute only one kind of data. Or he may distribute several different kinds of data. There may also be several different data distributors for the same kind of data. Sometimes, several data distributors make a federation of data distributors,

16. Guidelines for the establishment of monolingual thesauri are given by the international standard ISO 2788. Guidelines for the establishment of multilingual thesauri are given by the international standard ISO 5964.

17. For instance, the ISO 19115 metadata standard has a section entitled "Distribution Information" giving details about the distribution of the data to the user, including the type and format of the media used for the transfer of the data to the user.

in which case the user interacts with only one of them who forwards the requests for data retrieval to the other ones.

Whatever the data distributors' organization, the user will always have to prepare a data order request, submit it to a data distributor, wait for the order processing and retrieve the data.

Order Preparation

The user selects the data from the metadata and prepares an order to be submitted to one of the existing data distributors. The identity of the data distributor can be found in the metadata.

Order preparation may be as simple as just selecting a URL if the data are available on line from a HTTP server. But usually order preparation is much more complicated and includes the following steps:

- identify the data distributor

  There may be several distributors for the same data. The user must select at least one data distributor from the metadata[18]. A data distributor may distribute data on line or off line. Large data may be distributed only off line.

- get the order instructions

  The order instructions should be found in the metadata or, at least, the metadata should give the identity of a resource holding these instructions (the resource may be the web site of the data distributor).

- set up the order according to the order instructions

  The order contains items like the identity of the data and, if there is a choice, the medium that will be used to transfer the data to the user with the appropriate format. The medium may be a physical medium if the data is to be transferred off line (e.g. CDROM, digital linear tape). It may be an electronic link to an FTP server which will host the data. The FTP server may be an FTP server owned by the data distributor (in which case the data distributor must allow the user to be a registered FTP server user) or it may be an FTP server owned by the user (in which case the user must allow the data distributor to be a registered FTP server user and he must send him the required log in information).

  The order also contains items like the address where the medium will be delivered or the address where the data distributor will send the invoice - if any - related to the ordering.

Order Submission

Once it is prepared, the order must be submitted to the data distributor in the way indicated by the order instructions, e.g. by mail, by electronic mail, by fax or directly by updating the data distributor's web site. User authentication may be requested by the data distributor and the order may be enciphered during the transfer to the data distributor. The order is evaluated by the data distributor who accepts the order unless he cannot process it for some reason, in which case the order is rejected. The user should be informed of the acceptance or rejection of the order, in a way indicated in the order instructions (sending an Email to the user is a common way). The order is given an identifier by the data distributor for further order retrieval by the user.

Order Processing

Once it is accepted, the order is processed by the data distributor. Order processing may be as simple as sending to the user a web page containing the data that were ordered.

18. As mentioned earlier, the data catalogue may sometimes act as a data distributor.

However, order processing by the data distributor is generally much more complicated and includes following steps:

- find in the data repositories the various items that make the data ordered by the user; these items are known as granules, i.e. the smallest parts of the data retrievable from a repository, and are often packaged as files.
- extract the granules from the repositories and create the dataset expected by the user, i.e. the collection of all the data ordered by the user. This collection must also contain the data descriptions that are needed for the utilization of the data. These data descriptions provide the syntactic view of the data (whilst the metadata provide the semantic view). They actually are the *instructions for use* of the data. If they are not available in the dataset, the user cannot use the data.
- transfer the dataset to the user as indicated by the order.

Additionally, the data distributor maintains a record of all orders received from the user with their status. The user may access this record, check the current order status. He may also cancel an order being processed. An order may have one of the following status:

- accepted - the order is accepted by the data distributor
- rejected - the order is rejected by the data distributor
- being processed - the order is being processed by the data distributor
- waiting - the order is being processed but cannot be completed unless some event has occurred[19]
- suspended - the order is suspended because something prevents its completion[20]
- cancelled - the order has been cancelled (either by the user or by the data distributor)

Order completion

The user is informed by the data distributor that the order has been completed. Data should be made available to the user immediately after transfer.

CEOS recommendations for retrieving data

Recommendation 12

The development of a data ordering facility should follow the recommendations of the CEOS Catalogue Interoperability Protocol for data ordering.

Rationale

At the end of the 1990's, CEOS has experimented the CEOS Catalogue Interoperability Protocol (CIP). The CIP was the result of an initiative to develop a protocol so that a number of international agencies could make their Earth Observation data, and related data, available in coherent manner to their users communities[21].

19. A data distributor may offer a subscription service, i.e. an order submitted by a user may be related to data that do not already exist but are known to be available in the future. In this case, the order cannot be processed immediately.

20. If the data must be transferred to an FTP server and there is not enough space on this server, the order processing must be suspended.

21. Source: CEOS Catalogue Interoperability Protocol - (CIP), Specification - Release B

CIP documentation contains a very detailed and exhaustive description of the full ordering process from all the points of view that are needed for the implementation of a data ordering system: user requirements, system specification, system design.

Recommendation 13

Data that may be accessed online should at least be accessible via the FTP protocol.

Rationale

The FTP protocol is widely used in the world. It should be available by default for online data retrieval[22].

Recommendation 14

The data descriptions provided with the dataset should follow the recommendations of the ISO 15889:2003 standard (Space data and information transfer systems -- Data description language -- EAST specification)

Rationale

EAST is a powerful data description language defined by space agencies within the Consultative Committee for Space Data Systems (CCSDS). It allows to describe data at a very detailed level (including endianness). Tools exist to create or interpret datasets from their EAST descriptors and generate automatically the related documentation.

22. See chapter 4 for more details about the FTP protocol.

Chapter 3

From Data to Information

3.1 Introduction to Services

Earth Observation data are often required to be served in the form of information for end users. Interoperable technologies will enhance the capabilities to deliver information to the end users generally through application systems which interact with dispersed systems to perform services.

Providing information derived from EO data is achieved by applications that may be characterized as services. The ISO 19119 standard provides an abstract definition of a service:

- service

  distinct part of the functionality that is provided by an entity through *interfaces*
- interface

  named set of *operations* that characterize the behaviour of an entity
- operation

  specification of a transformation or query that an object may be called to execute

According to this definition, a service is a named set of operations identified as an interface. Each operation specifies an action which is performed by, or on behalf of the entity providing the service. Each action has a result which is part of the value added by the functionality.

Since an operation is just a specification, standardizing this specification will enable interoperability between all existing implementations of the corresponding action. For instance, two organizations may implement each in its own way the action consisting in the extraction of an image from a repository of Earth images. However, if the specification of this action is the same for both organizations, a user requesting an image from one organization will be able to request an image from the other organization in exactly the same way.

An interface is just the aggregation of all operations which contribute within the same service to a common objective. Whilst most services have only one objective and thus only one interface, complex services (often resulting from the aggregation of other services) may have more than one interface.

A service is always known by its specification. To achieve service interoperability, a service must have an implementation neutral specification. This implementation neutral specification is further mapped to platform specific specifications for each implementation of the service.

Services can also be categorized according their types which refer to some service taxonomy.

And finally, a service can also be seen as a resource and thus be described by metadata. Service metadata are stored in registries.

The following paragraphs will go more deeper in the analysis of services. The two last ones will give an overview of the concept of service oriented architecture and of the services defined by the Open Geospatial Consortium.

3.2 Service Specification

The ISO 19119 standard provides schemas for service specifications.

- **implementation neutral specification**

  A service has a name, and one or several interfaces. Each interface has a type and contains a static portion which contains the specification of each operation and a dynamic portion which contains the restrictions that apply to operations (e.g. an operation may depend on another operation).

- **platform specific specification**

  For each implementation neutral specification, there is at least one platform specific implementation. Each portion of the implementation neutral specification should have its counterpart int the platform specific specification. The mapping from an implementation neutral specification to a platform specific implementation can be made easier if the selected platform is a well known platform like CORBA/IDL or JAVA/EJB.

3.3 Service Taxonomy

Because there are many categories of services, each service has a type which refers to some service taxonomy.

The ISO 19119 standard contains a taxonomy for circa 80 geographic services listed according to the following categories:

- Geographic human interaction services
- Geographic model/information management services
- Geographic work-flow/task management services
- Geographic processing services, subdivided into:
  - Geographic processing services - spatial
  - Geographic processing services - thematic
  - Geographic processing services - temporal
  - Geographic processing services - metadata
- Geographic communication services
- Geographic system management services

The International Directory Network also contains a taxonomy with following top level categories:

- Data Analysis and Visualization
- Data Management/Data handling
- Education and Outreach
- Environmental Advisories
- Hazards Management
- Metadata Handling
- Models
- Reference and Information Services

Each category is further divided into sub categories.

Recommendation 15

An EO related service should always be characterized by a service type having a reference either to the ISO 19119 service taxonomy or to the International Directory Network service taxonomy.

Rationale

Both taxonomies are well established in the field of Earth observation. The ISO 19119 standard is the reference for geographic services. The International Directory Network already hosts about 2000 service metadata.

Both taxonomies offer a reach collection of categories for service classification.

3.4 Service Metadata

Service metadata describe service instances. They may be recorded in a registry for retrieval by a future service user. As data metadata, service metadata must provide enough information to allow a user to understand what the service is about and eventually to launch the service.

The ISO 19119 standard provides a schema for service metadata. According to this standard, service metadata comprise three parts:

- a description of the service (including information about the service provider and the service identification);
- a description of the service operations (including, for each operation, the descriptions related to each of the various service operation parameters);
- a description of the data the service operates on.

The ISO 19119 service metadata schema reuses many of the ISO 19115 metadata schema elements.

The International Directory Network also provides a schema for service metadata, the Service Entry Resource Format (SERF). The SERF schema is simpler than the ISO 19119 schema.

Recommendation 16

EO related services which belong to one of the categories of the ISO 19119 service taxonomy should be described according to the ISO 19119service metadata schema.

EO related services which belong to one of the categories of the International Directory Network service taxonomy should be described according to the International Directory Network service metadata schema.

Rationale

Services whose specifications are compliant with the ISO 19119 standard should also be described by ISO 19119 compliant service metadata.

The International Directory Network is a well known repository for service metadata.

Note

It may sometimes be useful to describe services twice, using both schemas.

3.5 Service Architecture

Services are implemented as sequences of actions, each action being specified by an operation and implemented as a piece of software running on a computer.

A service may have states, which depend on previous interactions, or be stateless. Simple services are stateless services.

It may useful to make new services by aggregating existing services in chains. In this case, such a new service does not necessarily have an associated service type (the service may not have a reference in a service taxonomy).

Service chaining can be visually represented by a directed graph. Nodes in the graph represent operations and edges represent service interactions between operations. For instance an edge may represent the output of the service represented by the source node which becomes the input of the service represented by the destination node, indicating that the second service cannot be activated unless the first service has terminated. Thus directed graphs are useful to represent service dependencies in a service chain. A directed graph may be acyclic or cyclic. If the graph is acyclic, a service can be activated only once in the chain. If the graph is cyclic, a service may be activated several times as part of a service loop. In this case, control mechanisms must be implemented to avoid endless looping.

The automation of the mechanisms which control a service chain is called a *work-flow*. A work-flow can be implemented as special services called a work-flow services.

Examples of work-flow services are

- the work-flow enactment service to define, invoke and control service chaining;
- the service chain service validation service;
- the resource reservation service, needed for the reservation of the resources used by service operations.

The ISO 19119 identifies three kind of service chaining:

- transparent chaining

  The user has full control over the service chain. He activates all of the services in the chain. No work-flow service is needed.

- translucent chaining

  The chain is mostly controlled by a work-flow service. The user is aware of the chain but has only limited control over the chain (he interfaces with the work-flow service). For instance he may suspend the execution of the chain or supply parameters which are needed by some services for execution.

- opaque chaining

  The user may not know that the service he is activating is actually a service chain. He sees the chain as an individual service. He has no control over the services in the chain.

An architecture that easily accepts services, work-flow services and service chains is called a *Service Oriented Architecture* (SOA).

3.6 The OGC Services

About OGC services

The Open Geospatial Consortium (OGC) has defined a number of services applicable to geographic data compliant with the above mentioned ISO 19119 standard.

Specifications in the geospatial domain, such as those developed by the Open Geospatial Consortium (OGC) and the ISO 19100 series are also applicable to Earth observation data. These standards allow EO data and information to be retrieved in a interoperable manner among dispersed service systems on the Internet.

Among OGC standards, implementation specifications for data providing purposes are specified in WMS[23] (Web Map Service), WFS (Web Feature Service), and WCS (Web Coverage Service) specifications. Those specifications define interfaces to provide data, low level processed EO data and highly processed data which are turned into thematic information such as thematic image derived from EO satellite data.

A data providing system provides data held within the system, according to a request received from a client system, by transforming the data into a requested geospatial configuration through on the fly re-projection, re-sampling, and sub-setting. The results are ingested into client systems for general purposes such as to display maps in an interactive web mapping system, or for further processing in a dynamic data processing system.

OGC services Overview

A service, in this case a WMS or WFS or WCS, consists of multiple operations. To activate a service, operations are invoked from a client system and received, parsed and performed by the service system to return a result to the client. The main operations of the WMS, WFS and WCS are the *GetCapabilities* operation and the operations to provide data such as the *GetMap* operation in the WMS.

A *GetCapabilities* operation returns an XML document. This document provides information about the data available from the service (i.e. metadata) and about the parameters that are required to formulate subsequent valid operations.

For instance, a WMS *GetCapabilities* operation returns the name of a coordinate reference system and the name of a data format. These names can be used as the parameters of a subsequent *GetMap* operation by which the user requests data of interest to displayed according to the specified coordinate reference system and rerieved using the specified data format.

23. WMS has become an international standard: ISO 19128 Geographic information -- Web map server interface.

Other parameters available for *GetMap* operation:

- server URL,
- layer-name (data name),
- supported geographical range (boundary box)
- supported time range
- background transparency

Application systems perform these operations implicitly on behalf of the end user who has an appropriate service interface to set-up his demand.

The main character of WMS, WFS, WCS in OGC specifications are summarized as follows:

- A WMS (Web Map Service) produces maps. A WMS client system may retrieve maps from multiple WMS servers. A map is defined as a visual representation of geospatial data. Maps are generally images encoded using a raster format like PNG or GIF, or a vector format like SVG.
- A WFS (Web Feature Service) produces geographic features.

  A WFS client may retrieve geographic features from multiple WFS servers.

  A geographic feature is a representation of real world phenomenon associated with a location relative to the Earth[24]. Geographic features are usually discrete entities. A WFS compliant geographic feature is encoded using the Geography Markup Language (GML).

  Unlike maps which are just images and usually require only limited processing by the client system, GML encoded geographic features may require complex processing by the client system for the rendering of the feature. For instance, a geographic feature like a river may represented by a series of attributes, including, *inter alia*, the name of the river and a sequence of coordinates, each coordinate couple identifying a location on the Earth surface. With GML, the representation of the river is actually an XML text. The rendering of the river on a map, will require the client system to interpret this text, and then compute and draw a curve using the sequence of coordinates and performing linear interpolation between to consecutive locations. Once drawn on a map, the curve will be the actual rendering of the river for the final user. Such a processing may be time consuming.

  For more information about GML, see paragraph below.
- A WCS (Web Coverage Service) produces “coverages".

  A coverage is a geographic feature that acts as a function which returns values from its range for any direct position within its spatial and/or temporal domain[25]. A coverage consists of a domain set, describing the domain of the coverage, and a range set, describing the range of the coverage. A coverage may be discrete or continuous (the range values may be constant or variable on each spatio temporal object in the domain). The coverage schema can be used for describing the properties of a phenomenon over space and/or time.

  A WCS system provides coverage data in a way which allows them to be rendered to the coverage user, or to be used as inputs to scientific models.

  WCS compliant coverages are encoded using geospatial data formats like GML, GeoTIFF, DTED, HDF, or EOS.

24. Source : Terminology of the ISO Geographic Information/Geomatics Technical Committee.
25. Source : Terminology of the ISO Geographic Information/Geomatics Technical Committee.

about OGC WMS, WFS, WCS is available at
http://www.opengeospatial.org/specs/?page=specs)

Recommendation 17

WMS, WFS and WCS compliant services provide space/time subsetted and projected data from interoperable servers distributed over the Internet. Data retrieved via these services can be used for purposes like mapping or dynamic data processing.

WMS is recommended for the retrieval and processing of mapping information.

WFS is recommended for the retrieval and processing of geographic features information.

WCS is recommended for the retrieval and processing of geographic coverages information.

Rationale

Using WMS, WFS and WCS compliant services enables a client system to retrieve mapping, geographic feature, and geographic coverage information from interoperable servers.

Adopting WMS, WFS and WCS standards as the basis for implementing data providing services, provides following benefits:

- the user is hidden from software/hardware issues between client systems and server systems;
- compatibility is increased between service entities implementing these standards.

CEOS has implemented a WMS based map server.

Constraints

From WGISS experiences, several limitations have been observed which could refrain from utilizing WMS, WFS, WCS services.

- **Differences in implementation**

  Although interfaces between a client system and a service system are defined by standards, ways to implement operations may vary among different systems. For example, to provide time series data, a system may use *TIME* parameter to specify time condition, while other system may determine time dimensional information within the *LAYERNAME* parameter by combining layer name with time stamp, sometimes of user defined time stamp. Therefore, it is possible that implementation vary among systems. Service metadata should reveal these differences.

- **Differences in standard versions**

  There exist several versions of OGC services. Upward compatibility is not guaranteed

3.7 The Geography Markup Language

The Geography Markup Language (GML) is an XML encoding for the spatial and non-spatial properties of geographic features.

Many geographic features have properties restricted to simple geometries with coordinates defined in 2 dimensions. These features are called “simple features”.

However, the real world cannot be completely represented with “simple features”. Geographic features have sometimes complex properties, like 3-dimensions geometry, spatial and/or temporal properties (including the definition of the spatial/temporal reference systems), or style properties for the geographic feature rendering. For geographic features like coverages or observations[26], units for physical measurements may also be required.

There are currently two coexisting versions of the Geography Markup Language: GML version 2 and GML version 3. GML version 2 deals with simple geographic features. GML version 3 extends GML version 2 to complex geographic features. Upward compatibility holds between both versions, but some GML version 2 clauses are claimed in GML version 3 to be deprecated and will be removed from GML version 3 in the future.

GML version 2 is much simpler than GML version 3 and can be used for systems dealing only with simple geographic features. In comparison, GML version 3 may be thought very complex and requires indepth expertise to be mastered.

GML version 3 has become an ISO standard: ISO 19136 - Geographic information -- Geography Markup Language (GML).

GML is a recommended language for the description of geographic features in the frame of OGC compliant web services.

26. A GML observation models the act of observing. An observation feature describes the “metadata” associated with an information capture event, together with a value for the result of the observation. Source: ISO 19136 Geographic information -- Geography Markup Language (GML). For instance, for an observation consisting in measuring a temperature, some properties are the time of the measurement, the region which is the object of the measurement (e.g. the name of a measurement station), the result of the measurement (i.e. the temperature), the measurement unit (e.g. Celsius degree).

Chapter 4

Technologies Enabling Interoperability

4.1 About technologies enabling interoperability

This chapter deals with the technologies CEOS has been experimenting with the objective of interoperability between geographically dispersed entities. The basic assumption was that all these entities would be widely dispersed and thus always be linked by a wide area network.

Most of the technologies analyzed by CEOS can be identified as protocols. A protocol is defined as the set of rules which must be observed for the transmission of data over a network in order for processes running on possibly different machines to communicate with each other.

The following and increasingly complex protocols will be analyzed in this chapter:

- TCP/IP and their companion protocols;
- OpenDAP;
- Z39.50;
- the grid protocols.

This list does not pretend to be exhaustive. Other protocols exist which would meet the same interoperability objective but they have not been analyzed by CEOS so far[27].

CEOS conclusion from the experimentation of the above mentioned technologies is that any of them can be used for the development of interoperable data and information systems. The final choice to be made by the developer only depends on the future system user requirements.

4.2 Methodology

As for the development of any other data system, the design and development of an interoperable data system should follow some well established methodology.

ISO/IEC 10746 *Information technology - Open Distributed Processing - Reference Model* is an ISO standard providing such a methodology for distributed systems implementing information processing services in an environment of heterogeneous IT resources and multiple organizational domains. These are characteristics of CEOS interoperable systems. This is the reason why application of this standard may be recommended for the design of interoperable systems.

Recommendation 18

A standard like ISO/IEC 10746 *Information technology - Open Distributed Processing Reference Model* should be used for the design and development of an interoperable data system.

27. CORBA (“Common Object Request Broker Architecture”) is an example of technology enabling interoperability which has not yet been analyzed by CEOS.

Rationale

The reference model described in ISO/IEC 10746 standard comprises four fundamental elements:

- an object modelling approach to system specification (a system is seen as a set of objects interacting at interfaces; objects have behaviour identified by roles, they may be classified according to their types; composition and decomposition may be applied to objects; objects and sets of objects may be described by templates);
- the specification of a system in terms of separate but interrelated viewpoint specifications; each viewpoint is a subdivision of the complete system specification; the standard defines the following five viewpoints:
  - the enterprise viewpoint, specifying the objectives and policy constraints of the system in terms of roles (e.g.: users, providers);
  - the information viewpoint specifying the semantics of the information and information processing within the system, namely through schemata;
  - the computational viewpoint, specifying the decomposition of the system into objects by the definition of object interfaces and interactions between objects (the system is seen as a single virtual machine);
  - the engineering viewpoint, specifying how the objects previously defined may be grouped in view of their effective interacting (the system is seen as a distributed set of capsules, each capsule being defined as a set of clusters of objects; all objects in a cluster can be manipulated through one single operation, e.g. moved from their current physical location to another location)[28];
  - the technology viewpoint, specifying the implementation of the system in terms of hardware and software;
- the definition of a system infrastructure providing distribution transparencies for system applications; the standard defines a set of transparencies (e.g.: location transparency, which avoids the use of information about the location of the binding to an interface);
- a framework for assessing system conformity (i.e. : the relationship between a specification and a real implementation).

4.3 Some technologies enabling Interoperability

TCP/IP

About TCP/IP protocols

TCP/IP (“Transmission Control Protocol/Internet Protocol”) is the generic name of a family of protocols for the transmission of data over Internet[29]. This name comes from the two main protocols of the family: IP which defines the transmission of data packets, and TCP which coordinates the transmission of data packets.

28. The two -or three tiers client server model is a very popular engineering viewpoint.
29. Internet protocols are available from the Internet Engineering Task Force (IETF): http://www.ietf.org.

TCP/IP protocols are categorized by layers. Following layers are defined within TCP/IP:

- access - defines the physical organization of the data on the network
- Internet - defines the data packets
- transport - defines the data packet transmission
- application - defines standard network applications

HTTP (“HyperText Transfer Protocol”) and FTP (“File Transfer Protocol”) are the two primary TCP/IP protocols at the application layer.

- HTTP

  A machine called “the client” requests another machine called “the server” to retrieve some information and send it back to the client. The server is identified by its physical address on the network, called its “IP address”. However, since IP addresses are difficult to memorize, the server may also be identified by a symbolic name, called its “domain name”. For the network, IP addresses and domain names are equivalent.

  The client provides via the HTTP request the kind of information it expects to be sent back by the server (the protocols defines several ¿methods" to do provide this information to the server). For instance, the client may provide the name of a database known by the server together with an SQL (Structured Query Language) query to apply against it. The information expected by the client is sent back by the server in a file transmitted to the client. All the file formats available from the MIME (Multipurpose Internet Mail Extensions) are accepted by the HTTP protocol.

  It should be noted that the HTTP is a stateless protocol: the server does not keep track of the requesting client (the server answers the client and forgets the client immediately). It should be noted also that the underlying TCP protocol is actually a stateful protocol. As a consequence, the chaining of several HTTP requests entails as the opening and closing of as many TCP sessions.

- FTP

  A machine called “the user” requests another machine called “the server” to transmit files. Files may flow bidirectionally between the machines (a file may be downloaded from or uploaded to a server).

  Two channels are opened to handle FTP requests:

  - a “control channel”, which is used to control the FTP commands; this is done via the TELNET protocol (another TCP/IP protocol) which opens an FTP session; FTP commands include commands like: “retrieve file”, “store file”, “rename file”, “delete file”, “create directory”, “delete directory”; there is also a “logout” command which closes the FTP session.
  - a “data channel”, which is used to transmit the files between the machines.

  The control channel is always opened between a user and a server.

  The data channel may be opened between a user and a server, or between two servers. If the data channel is opened between a user and a server, this user may not be the user holding the control channel. In other words, file transfers may take place between two machines on request of a third one.

Security can be added both to HTTP and FTP. For HTTP, this is achieved using the TLS (“Transport Layer Security”) protocol. TLS actually adds a “session” on top of the TCP/IP transport session and below the application session. If TLS is used, HTTP becomes HTTPS (HTTP Secured). For FTP, security is achieved using a secured variant of the

TELNET protocol, called SSH (Secure Shell). If SSH (version 2) is used, FTP becomes SFTP (Secure File Transfer Protocol).

Recommendation 19

The HTTP (HTTPS) protocol is recommended for simple (secured) stateless services implementation over TCP/IP networks. The FTP (SFTP) protocol is recommended for simple (secured) file transfers between a client and a server.

Rationale

The HTTP protocol opens an easy way to implement a service designed according to the 2- or 3 tiers client/server model. The client is usually an Internet browser. HTTP requests go to a server which handles the requests. The 3 tiers model enables interoperability because the server may forward the receiving request to other servers, collect the responses from these additional servers, and concatenate these responses into a single one which is then sent to the client.

Most geographic services defined by the Open Geospatial Consortium (OGC) suggest an HTTP implementation.

A well known HTTP implementation for data information and access services is OpenDAP.

The FTP protocol remains the ordinary way to implement a file transfer service over TCP/IP networks.

OpenDAP

About OpenDAP

OpenDAP is the acronym for “Open Source Project for a Network Data Access Protocol". It refers to a client/server mechanism for transparent access of data across the Internet. Data held under various formats are retrieved by OpenDAP clients from OpenDAP servers.

An OpenDAP client is often a data analysis package (e.g. GrADS[30], Ferret[31]) with an extension to use OpenDAP software for data access. Data retrieved from OpenDAP servers are then seen by the the data analysis package user as if they were local data. But the OpenDAP client may also just be a simple Internet browser.

An OpenDAP client requests services from an OpenDAP server using the above mentioned HTTP protocol. Available OpenDAP services are:

- data information services

  In addition to general information about the dataset (e.g. date of creation, file name and file type) and about the originating server, data information services provide the structure of the dataset (the “Dataset Descriptor Structure“ in the OpenDAP terminology) and the attributes that are related to the variables contained in the dataset (the “Dataset Attribute Structure” in the OpenDAP terminology).

- data access services

  Data access services provide access to the full dataset or to a subset of the dataset. The subset is defined by the client through a constraint expression which

30. Grid Analysis and Display System (GrADS): http://www.iges.org/grads/grads.html
31. Ferret: http://ferret.pmel.noaa.gov

is sent with the HTTP request to the server. A constraint expression consists of two parts:

- a projection, i.e. the list of the variables that are to be returned to the client;
- a selection, i.e. a function applied to one or several variables of the dataset.

Only the variables in the projection are returned, with the constraints indicated by the selection. If the selection is omitted, all the variables in the projection are returned. If the projection is omitted, all the variables in the dataset are returned but with the constraints indicated by the selection.

Data information services and data access services require that OpenDAP clients and OpenDAP servers agree on a number of file formats. File format supported by the current version of OpenDAP include HDF format, netCDF format, and WMO[32] format for gridded binary data.

The software needed to build OpenDAP clients and OpenDAP servers may be downloaded from the OpenDAP website[33].

Recommendation 20

OpenDAP is recommended for the implementation of simple stateless data retrieval and data analysis systems using well known file formats like HDF, netCDF, or the WMO format for gridded binary data.

Rationale

OpenDAP was initially developed by the oceanography community and was known as DODS ("Distributed Oceanographic Data System") until its name was turned into OpenDAP. Though it is still mainly used by the oceanography community, many servers from other domains are OpenDAP servers (cf. the IDN OpenDAP server portal). OpenDAP has been used by the CEOS in the frame of the WGISS Test Facility for Coordinated Enhanced Observing Period project (WTF-CEOP)[34], a distributed data integration system project lead by JAXA.

Installing dedicated OpenDAP clients and servers is an easy task.

Z39.50

About Z39.50

Z39.50 is the name of a standard given by the American National Standards Institute (ANSI) to an application service and its related protocol specification for interoperable retrieval of information located in databases[35]. The latest Z39.50 revision has been released in 2003.

In its abstract, the standard states:

> *This standard defines a client/server based service and protocol for Information Retrieval. It specifies procedures and formats for a client to search a database provided by a server, retrieve database records, and perform related information retrieval functions. The protocol addresses communication between information retrieval applications at the client and server; it does not address interaction between the client and the end user.*

32. World Meteorological Organization
33. For more information about OpenDAP: http://www.opendap.org.
34. More information about WTF-CEOP: http://jaxa.ceos.org/wtf_ceop
35. Z39.50 has become an international standard identified as "ISO 23950:1998 Information and documentation -- Information retrieval (Z39.50) -- Application service definition and protocol specification".

Communication between the client and the server is performed via an explicit association (called a Z association in the Z39.50 terminology). An association is always established by the client. It may be terminated by the client, or by the server, or by loss of connection. There may be multiple consecutive associations between a client and a server for a connection and there may be multiple consecutive as well as concurrent operations within an association. Thus, unlike other protocols like HTTP protocols which are stateless, the Z39.50 protocol is a stateful protocol.

Z39.50 clients search server databases. Since databases may be implemented in many different ways, Z39.50 defines an abstract database model against which individual database models will have to be matched. In the Z39.50 terminology, a database refers to a set of records containing related information. A database access point is a key (e.g. a document title) that can be specified in a search for records in a database. A search for records in a database consists in applying a query to the database. The query specifies the values to be matched against the database access points, i.e. the query specifies access clauses, which may additionally be linked by logical operators. The client may indicate several databases to be queried through a single query.

The result of a query is a subset of database records (the query result set). A result set may itself be referenced in a subsequent query.

In order to retrieve a record from the result set, the client must supply:

- a database schema identifier

  A database schema abstractly defines the database record structure (an abstract record structure in the Z39.50 terminology). An abstract record structure results in an abstract database record when applied to a database record. Each element of the database schema (e.g. “author”, “title", “abstract") is uniquely identified by a sequence of tags. This sequence identifies the path which must be followed in the database schema to reach the element, starting from the database schema root.

- an element specification

  A element specification defines the elements to be extracted by the server from the abstract database record (e.g. the sequence of tags identifying the elements to be extracted)

- a record syntax identifier

  A record syntax is applied by the server to an abstract database record, resulting in an exportable structure (called a retrieval record in the Z39.50 terminology). The simplest record syntax is the “Simple Unstructured Text Record Syntax” (SUTRS). In a SUTRS record, elements are just character strings separated by ASCII line terminators.

Database schema, element specification and record syntax are a common knowledge about the database structure shared between the client and the server. Usually, the client retrieves this knowledge by searching a special database, called the “explain” database in the Z39.50 terminology. This “explain” database behaves like any other database handled by the server but it has a pre defined record syntax and pre defined search clauses by which the above mentioned knowledge may be easily retrieved and handled by the client.

The Z39.50 standard defines several services that are established between client and server within a Z association (services are carried out by the exchange of messages), for instance:

- the “Init” and “Close” services invoked to open and close a Z association;
- the “Search” service invoked to start a search operation applied to one or several databases (this service is also used to search the ¿explain" database);
- the “Present“ service invoked to request response records from a specified result set;
- the “Scan” service invoked to scan a result set;
- the “Sort” service invoked to sort a result set;
- the “Extended Services” service invoked to create task packages; these packages are held in an “extended services” database which is handled in a way similar to the “explain” database. Task packages provide non Z39.50 services within a Z39.50 compliant system. For instance, a Z39.50 compliant information retrieval system could include a data ordering service in addition to the standard Z39.50 services.

The Z39.50 standard is maintained by the Library of Congress (USA)[36]. The Library of Congress also maintains a list of registered Z39.50 profiles, a profile being defined as the specification of the standard use for the specific needs of a community (a profile includes items like the definition of the database schema to be used by the community, the definition of the search service parameters, etc.).

There are two well known profiles in the field of geographic data:

- the Application Profile for Geospatial Metadata as defined by the U.S. Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata; this profile, commonly known as the GEO profile, is broadly used in the USA;
- the Catalogue Interoperability Protocol for Earth Observation Metadata as defined by CEOS; this profile, commonly known as the CIP profile is used within the eoPortal.

Recommendation 21

The Z39.50 protocol has been used for the implementation of powerful stateful information retrieval services implying sophisticated queries over several distributed databases. However it is not recommended for the implementation of new systems, unless compatibility is sought with legacy Z39.50 based systems.

Rationale

The Z39.50 protocol is being used heavily for the implementation of distributed catalogue services in the world of digital libraries. It is being used also for the implementation of distributed catalogue services for geographic data. In particular, Z39.50 was selected to provide search interoperability among the different servers of the US National Geospatial Data Clearinghouse Network.

IDN metadata are available through the FGDC Clearinghouse and the Geospatial One-Stop (GOS) portal. The metadata in the IDN are exchanged using a legacy Z39.50 harvester used by the Clearinghouse and also by Geospatial One-Stop (GOS). A customized stylesheet is used to translate IDN DIF metadata to FGDC compatible metadata. These

36. The standard and other information may be retrieved from: http://www.loc.gov/z3950/agency/ .

efforts support the National Spatial Data Infrastructure (NSDI) initiative and the ISO standards.

However, CEOS no longer recommends the use of the Z39.50 protocol for the implementation of new systems because other technologies like HTTP and web services are now preferred by CEOS agencies because, despite most of them are based on stateless protocols, they are much easier to implement, especially over the Internet where firewall crossing is required. If compatibility is sought with legacy systems using Z39.50, implementing a Z39.50 interface is acceptable.

Grids

About Grids

A grid is a collection of pooled resources distributed over the Internet (it is a distributed system). Resources encompass computers, storage facilities, data, software. The user of a grid resource does not necessarily know the physical location of the resource[37].

The objective of a grid is to make resources available to as many users as possible, i.e. to avoid that resources are being under utilized. For instance, if an application is able to run simultaneously on several computers, it will run, once submitted to a grid, on as many computers from the grid as are available.

There are three broad categories of applications for grids:

- computation intensive applications (aggregation of computing resources);
- data intensive applications (sharing of large scale databases);
- distributed collaboration applications (sharing of complex models).

A grid may be deployed inside a same enterprise. It may also be deployed over several enterprises. In this case, the grid is a means for these enterprises to reach a shared objective. For instance, space agencies could agree, in the frame of a natural hazards monitoring joined project, to share tsunami data and their related applications via a grid.

Grids, as repositories of pooled and distributed resources, are by nature technically very complex: resources of all kinds must be identified, named, described, registered, published, activated, monitored, managed, secured. The approach taken by the grid community to overcome the complexities of all these capacities was to define, and recommend for grids, a set of standards:

- OGSA

  The “Open Grid Service Architecture” (OGSA) defines the requirements for grid capabilities, including security requirements, and thus the underlying grid functions.

- GridFTP

  GridFTP is a secure data transfer protocol, based on the Internet FTP protocol with extensions that make it suitable for data transfers over large area networks with large bandwidth.

- WSRF

  The “Web Service Resource Framework” (WSRF) specifies a web service resource as the relationship between a web service and a resource.

37. In this sense, the World Wide Web is a grid. The World Wide Web resources distributed over the Internet are the web pages available to web browsers. Conversely a cluster of computers is not a grid in this sense because its resources (the cluster computers) are not distributed over the Internet.

About web services: A web service is an application than can be reached over a network through a application interface described according to the standardized model defined by the "Web Service Description Language" (WSDL). WSDL defines a web service as a set of operations, their bindings to network addresses through a network protocol, and the message formats for the flow of each operation input and output parameters. A WSDL compliant web service description is actually a textual document which can be stored in a web service registry from where it can later be retrieved by a client application and processed to generate the application interface required to launch the web service.

About web service resources: A web service always acts upon some resource. The resource is implicitly indicated through operation input parameters and may be modified according to operation output parameters. The WSRF identifies a new entity - called a web service resource - by explicit pairing of a web service and the resource which it acts upon at some time. The pairing of a web service and a resource is achieved within the WSRF by an extension of WSDL compliant web service description document making it able to hold the resource identity and the resource properties (for instance, if the resource is a file, its identity may be the name by which it is known, and one of its properties may be its size). It must be noted that a web service may be linked to more than one web service resource and similarly a web service resource may qualify more than one web service.

A web service and a web service resource are both network applications but there is a main difference between them :

- a web service is inherently permanent (it is not bound to a particular resource) and stateless (it does not know by itself about the state of the resource it acts upon) ;
- a web service resource is inherently transient (it is bound to a particular resource and exists only during the resource lifetime) and stateful (it knows the state of the resource to which it is bound through the resource properties).

The WSRF is not specific to grids. It can be used in other contexts. But it greatly helps defining a grid architecture because it provides, by means of a resource and service abstraction mechanism, a way to handle the complexity and heterogeneousness of all the resources (and their evolutions) that make a grid.

About grid security: grid security can become a concern, especially for larger grids that span many locations and cross enterprise boundaries. This is why the grid community has established a sophisticated "grid security model" to secure grid services, in response to the OGSA grid security requirements. This model also provides the security services that are needed to build the corresponding grid service security functions (authentication, authorization, service security policy representation, and trust negotiation in a web services context).

Recommendation 22

The grid technology is recommended for the development of complex interoperable service oriented infrastructures for the sharing of many collaborative applications amongst geographically dispersed organizations.

Rationale

The grid technology has emerged after 1990 and has now reached a level of maturity compliant with the requirements of modern applications. Though the grid technology remains very complex, many grid software suites exist over the world and a number of grids are running under operational conditions.

The Globust Toolkit software

The Globus Toolkit is being developed by the Globus Alliance. It is an open source software for building grids. It comprises development kits for the implementation of grid services and of the grid security infrastructure, as well as a variety of basic grid services, based on open standards.

As suggested by its name, it is a toolkit. Not all components need to be installed, but only those required by the characteristics of the grid being built.

Latest version (4.0) is OGSA and WSRF compliant.

Recommendation 23

The Globus Toolkit (version 4.0) is recommended as a developing tool for building service oriented grids.

Rationale

The Globus Toolkit version 4.0 is a tool for developers. It provides the basic building blocks and development kits for building grids as service oriented distributed application systems. It is world wide known and has now reached with version 4.0 a high level of maturity. It has been experimented by CEOS agencies.

Note

Globus Toolkit previous versions are not service oriented unlike version 4. Version 2 can be used if building a service oriented grid is not the primary goal.

CEOS Recommendations for Interoperability

NO	Recommendation
1	Interoperable archives archivists should use a same glossary of terms and definitions applicable to data archiving.
2	An archive system should comply with the Reference Model for an "Open Archival Information System" (OAIS).
3	An archive should use a questionnaire to help archivists in appraising data that are candidates for archiving.
4	An archive should apply the "Producer-Archive Interface Methodology Abstract Standard" (PAIMAS) standard to define its interactions with information producers.
5	CEOS agencies should use the purge alert service should be used before information removal from archives.
6	The CEOS Directory Interchange Format (DIF) should be used for the description of science Earth Observation data collections.
7	The CEOS International Directory Network (IDN) should be used to host DIF compliant metadata.
8	Description of geographic data should be done in conformity to the ISO 19115 metadata standard.
9	The Protocol for Metadata Harvesting (PMH), developed through the Open Archives Initiative (OAI) protocol should be used for metadata harvesting.
10	A catalogue should always link each of the keywords available for data searching to the appropriate thesaurus.
11	For science data, catalogues should use to the far extent possible the keywords defined by the CEOS International Directory Network.
12	The development of a data ordering facility should follow the recommendations of the CEOS Catalogue Interoperability Protocol for data ordering.
13	Data that may be accessed online should at least be accessible via the ftp protocol.
14	The data descriptions provided with the data set should follow the recommendations of the ISO 15889:2003 standard (Space data and information transfer systems -- Data description language -- EAST specification)
15	An EO related service should always be characterized by a service type having a reference either to the ISO 19119 service taxonomy or to the International Directory Network service taxonomy.

NO	Recommendation
16	EO related services which belong to one of the categories of the ISO 19119 service taxonomy should be described according to the ISO 19119 service metadata schema. EO related services which belong to one of the categories of the International Directory Network service taxonomy should be described according to the International Directory Network service metadata schema.
17	WMS, WFS and WCS compliant services provide space/time subsetted and projected data from interoperable servers distributed over the Internet. Data retrieved via these services can be used for purposes like mapping or dynamic data processing. WMS is recommended for the retrieval and processing of mapping information. WFS is recommended for the retrieval and processing of geographic features information. WCS is recommended for the retrieval and processing of geographic coverages information.
18	A standard like ISO/IEC 10746 *Information technology - Open Distributed Processing Reference Model* should be used for the design and development of an interoperable data system.
19	The HTTP (HTTPS) protocol is recommended for simple (secured) stateless services implementation over TCP/IP networks. The FTP (SFTP) protocol is recommended for simple (secured) file transfers between a client and a server.
20	OpenDAP is recommended for the implementation of simple stateless data retrieval and data analysis systems using well known file formats like HDF, netCDF, or the WMO format for gridded binary data.
21	The Z39.50 protocol has been used for the implementation of powerful stateful information retrieval services implying sophisticated queries over several distributed databases. However it is not recommended for the implementation of new systems, unless compatibility is sought with legacy Z39.50 based systems.
22	The grid technology is recommended for the development of complex interoperable service oriented infrastructures for the sharing of many collaborative applications amongst geographically dispersed organizations.
23	The Globus Toolkit (version 4.0) is recommended as a developing tool for building service oriented grids.

Index

A

C

D

E

F

G

H

I

J

L

M

N

O

P

R

S

T

U

W

X

Z