Chapter 6: Archiving and Access Systems for Remote Sensing

John Faundeen
U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA

George Percivall
Open Geospatial Consortium, Wayland, MA, USA

Contributing Authors:
Shirley Baros
Earth Data Analysis Center, University of New Mexico, Albuquerque, NM, USA

Peter Baumann
Jacobs University, Bremen, Germany

Peter Becker
Esri, Redlands, CA, USA

J. Behnke
NASA Goddard Space Flight Center, Greenbelt, MD, USA

Karl Benedict
College of University Libraries and Learning Sciences, University of New Mexico, Albuquerque, NM, USA

Lucio Colaiacomo
European Union Satellite Center, Madrid, Spain

Liping Di
Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA USA

Chris Doescher
U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA

J. Dominguez
LizardTech, Seattle, WA, USA

Roger Edberg
National Computational Infrastructure, Australian National University, Acton, Australia

Mark Ferguson
National Archives and Records Administration, National Records Management Program, Records Management Services, Broomfield, CO, USA

Stephen Foreman
World Meteorological Organization, Geneva, Switzerland

David Giaretta
Giaretta Associates Ltd, and PTAB Ltd, Yetminster, UK

Vivian B. Hutchison
U.S. Geological Survey, Core Science Systems, Lakewood, CO, USA

Alex Ip
Geoscience Australia, Canberra, Australia

N.L. James
NASA Goddard Space Flight Center, Greenbelt, MD, USA

Siri Jodha S. Khalsa
University of Colorado, Boulder, CO, USA

B. Lazorchak
Library of Congress National Digital Information Infrastructure and Preservation Program, Washington, DC, USA

Adam Lewis
Geoscience Australia, Canberra, Australia

Fuqin Li
Geoscience Australia, Canberra, Australia

Leo Lymburner
Geoscience Australia, Canberra, Australia

C.S. Lynnes
National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, MD, USA

Matt Martens
Stinger Ghaffarian Technologies (SGT), Sioux Falls, SD, USA

Rachel Melrose
Geoscience Australia, Canberra, Australia

Steve Morris
North Carolina State University Libraries, Raleigh, NC, USA

Norman Mueller
Geoscience Australia, Canberra, Australia

Vivek Navale
Center for Information Technology, National Institutes of Health, Bethesda, MD, USA

Kumar Navulur
DigitalGlobe, Westminster, CO, USA

D.J. Newman
Aeronautics, Astronautics, and Engineering Systems, M.I.T. School of Science, Cambridge, MA, USA

Simon Oliver
Geoscience Australia, Canberra, Australia

Matthew Purss
Geoscience Australia, Canberra, Australia

H.K. Ramapriyan
Science Systems and Applications, Inc., Lanham, MD, and NASA Goddard Space Flight Center, Greenbelt, MD, USA

Russ Rew
UCAR Unidata Program, Boulder, CO, USA

Michael Rosen
LizardTech, Seattle, WA, USA

John Savickas
Earth Data Analysis Center, University of New Mexico, Albuquerque, NM, USA

Joshua Sixsmith
Geoscience Australia, Canberra, Australia

Tom Sohre
U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA

David Thau
Developer Advocate, Google Earth Engine, Mountain View, CA

Paul Uhlir
Scholar, National Academy of Sciences, and Consultant, Data Policy and Management, Annandale, VA 22003

Lan-Wei Wang
Geoscience Australia, Canberra, Australia

Jeff Young
LizardTech, Denver, CO, USA

1 INTRODUCTION

MRS-4 Chapter 6, *Archiving and Access,* focuses on major developments inaugurated by the Committee on Earth Observation Satellites, the Group on Earth Observations System of Systems, and the International Council for Science World Data System at the global level; initiatives at national levels to create data centers (e.g. the National Aeronautics and Space Administration (NASA) Distributed Active Archive Centers and other international space agency counterparts), and non-government systems (e.g. Center for International Earth Science Information Network). Other major elements focus on emerging tool sets, requirements for metadata, data storage and refresh methods, the rise of cloud computing, and questions about what and how much data should be saved. The sub-sections of the chapter address topics relevant to the science, engineering and standards used for state-of-the-art operational and experimental systems.

1.1 *Implementation focused experience area*

This chapter focuses on Archiving, Discovery, Visualization and Access, and Processing and Workflows. Each main topic has several sections addressed by subject matter experts from private industry, academia, or government, including:

- Major developments by CEOS, GEOSS, and ICSU (World Data System) at the global level;
- Initiatives at national levels to create data centers (e.g. NASA DAACs and international counterparts);
- Non-government systems (e.g. RGIS, CIESIN);
- Other major elements focus on emerging tools, requirements for metadata, data storage and refresh, cloud computing, interoperability;
- Science, engineering and standards used for state-of-the-art operational and experimental systems.

1.2 *Functional analysis of archiving and access*

Certain key components are essential to archive and access remotely sensed data successfully. Remote sensing information systems grow more useful as additional components are contributed. Examples of components include observing systems, data processing systems, dissemination systems, capacity building or

other initiatives. The GEOSS Strategic Guidance Document cites several components already contributed to GEOSS (Figure 6-1):

- Components to acquire observations based on existing local, national, regional and global systems to be augmented as required by new observing systems;
- Components to process data into useful information recognizing the value of modeling, integration and assimilation techniques as input to the decision support systems required in response to societal needs; and
- Components required to exchange and to disseminate observational data and information including data management, access to data, and archiving of data and other resources.

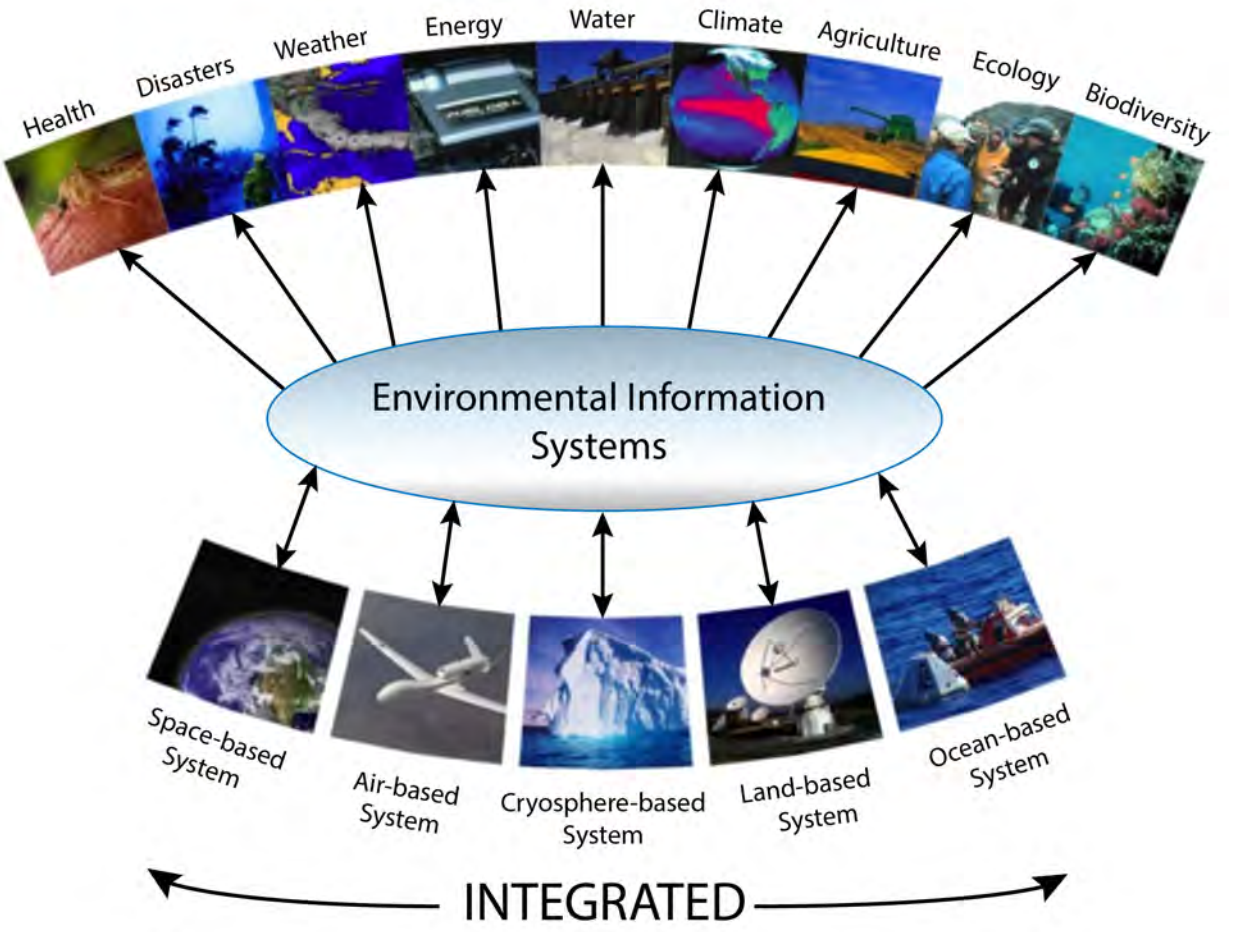


Figure 6-1. A Global Earth Observation System of Systems (GEOSS).

In analyzing any system for archiving and accessing remotely sensed and geospatial data, it is helpful to identify clearly the Functional Use Cases involved. Use cases are reusable functions observed repeatedly in remote sensing information systems. Use cases provide easy, familiar processes for providers to describe their functions, as well as for users to anticipate the application of remotely sense data. A summary of Use Cases is shown in Figure 6-2.

Figure 6-2. Functional use cases for archiving and accessing remote sensing.

Note the two archive and access actors in Figure 6-2 symbolized as "stick figures" for users and resource providers. Table 6-1 shows two additional actors, the domain integrator and the system operator, who are found typically in GEOSS use-cases. In addition to these two actors, there are also primary, secondary and administrative GEOSS Actors involved in use-cases as listed in Table 6-1.

Table 6-1. Archive and Access Actors.

Actor	Description	Role Type
User	Discovers, consumes, and exploits GEOSS resources	Principal
Resource Provider	Deploys, operates, registers GEOSS resources	Principal
Domain Integrator	Builds network of organizations and components to address objectives for a domain community	Secondary
System Operator	Operates the system components that provide the remote sensing archive and access functions	Administrative

2 ARCHIVING

This section addresses the elements involved with archiving remotely sensed and geospatial data. A wide variety of useful information is provided on topics such as archiving requirements and planning strategies. Retention and appraisal practices are inter-related and can help organizations focus their resources on the data most important to their particular mission. Disposing of archived data often is viewed as destructive, but this activity does not need to end that way. Transferring a collection to another organization may free up resources and better align a collection to a user community. Additional topics of how technology affects archiving, the importance of metadata, how storage practices and media formats affect longevity also are discussed. In summary, this section focuses on the elements of preservation with subject matter experts providing guidance and/or best practices learned through years of experience.

2.1 *Open Archival Information System (OAIS)*

OAIS (ISO 14721:2012) is a reference model for long-term preservation of digital (and other) assets. While this model does not specify the implementation or functional design, OAIS is an indispensable guide to creating an archive with an accepted set of roles, responsibilities and methods that encourage safe, long-

term archival and access of critical government information. The OAIS Reference Model was designed primarily to address three broad issues:

- that digital preservation is difficult for an organization to prove, and correspondingly difficult to test;
- that fundamental principles of digital preservation of any digital object, in particular scientific data, are far from clear; and that…
- there are many difficulties arising from differences in terminology used in different domains.

To address these three broad areas OAIS contains two distinct models and associated concepts. The first is the Information Model and associated ideas – conformance is linked to these; and the second is the Functional Model, which defines terminology but is not involved in conformance. These are described in turn.

The starting point for the Information Model is a set of definitions of interrelated fundamental concepts defining preservation in terms of continued usability and understandability. The advantage of this idea is that it is applicable to any type of digital object, and it is testable.

2.1.1 *Definitions*

The italicized text below is taken from OAIS (ISO 14721-2012) and takes a very general definition of its prime concern which, the "I" in OAIS suggests is information:

- Information: *Any type of knowledge that can be exchanged. In an exchange, it is represented by data.* An example is a string of bits (the data) accompanied by a description of how to interpret the string of bits as numbers representing temperature observations measured in degrees Celsius. Note that k*nowledge* is not defined in OAIS.
- Data: *A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.* Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.
- Digital Object: *An object composed of a set of bit sequences.*

One might wonder why data include physical objects such as a "moon rock specimen". The answer becomes clear later; but in essence, if one is to provide a logically complete solution to digital preservation, one needs to jump outside the digital, if only for example, to read the labels on tapes to know which tape to put in the reader.

Regarding the length of time for which one needs to be concerned, OAIS provides the following pair of definitions:

- Long Term: *A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future.*
- Long Term Preservation: *The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term.*

In other words, we are not only talking about decades into the future, but as is a common experience, we need to be concerned with the rapid change of hardware and software, and the time cycle which may be

just a few years. Of course even if an archive is not, itself, looking after the digital objects over the long term, even by that definition, the intention may be for another archive to take over later. In this case the first archive needs to capture all the metadata needed so that they can be included as well.

2.1.2 *Key concepts*

Three key concepts are embedded in “Long Term Preservation” and require further explanation/discussion; beginning with “Authenticity” and “Independently Understandable”.

- Authenticity: *The degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence.* Note that Authenticity is not a yes/no concept.
- Independently Understandable: *A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.*

Now we approach an element of what the "preservation" part of "digital preservation" means. To require that objects are able to be "interpreted, understood and used" is to make some very powerful demands. These could include playing a digital recording so it can be heard, or rendering an image or a document so that it can be seen. It also includes being able to understand what the columns in a spreadsheet, or what the numbers in a set of scientific data mean. This information is needed to understand, and in particular, to be able to use the data in an analytical program, or combining them with other data to derive new scientific insights. The "Independently" part “independently understandable” is to exclude the easy but unreliable option of being able to simply ask the person who created the digital object; unreliable not because the creator may be incorrect but, because the creator may be, and in the very long term certainly will be, deceased!

Another key concept in “Long Term Preservation,” is that of “Designated Community”.

- Designated Community*: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time.*

Why is this a key concept? To answer that question, one needs to ask another fundamental question, namely: how can one tell whether a digital object has been successfully preserved? This question might be asked repeatedly as time passes. Clearly, one can do simple things like checking whether the bit sequences are unchanged over time, using one or more standard techniques such as a cryptographic hash (often called a digital digest). However just having the bits is not enough. The demand for the object to be "interpreted, understood and used" is broader than that; and, of course, it can be tested.

But surely there is another qualification. Is it sensible to demand that *anyone* can "interpret, understand and use" the digital object, for example a four year old child? Clearly we need to be more specific. But how can such a group be specified, and indeed, who should choose? This seems a daunting task; who could possibly be in a position to do that? The answer that OAIS provides is subtle. Those who are able to "interpret, understand, and use" the digital object and who can test the success of the "preservation", are defined by the people who are doing the preservation.

2.1.3 *Information model*

2.1.3.1 What "metadata", how much "metadata"?

Metadata, often defined as "data about data," are required to properly exploit, for example, sensor returns from aircraft and space-borne sensors. A fundamental question to ask is 'what *"metadata"* do we need, and how much is needed?' The problem with *"metadata"* is that they are so broad that people tend to have their own, usually limited, view. OAIS provides a more detailed breakdown. The first three broad categories deal with (1) understandability; (2) origins, context and restrictions; and (3) the way in which the data and *"metadata"* are grouped together. The reason for this separation is that given some digitally encoded information one can reasonably ask whether it is usable, which is dealt with by (1). This is a separate question to the one about where this digital object came from, dealt with by (2). Since there are many ways of associating objects it seems reasonable to define how this is done in any specific case (3). The next few sub-sections introduce these different categories briefly. They are each discussed in much greater detail in OAIS.

- Understandability (Representation Information)

One type of *"metadata"* that can be identified immediately is the need to interpret, understand and use digitally encoded information. OAIS defines this as *Representation Information*; that is, *the information that maps a Data Object into more meaningful concepts* (Figure 6-3). An example of Representation Information for a bit sequence which makes up a FITS file might consist of the FITS standard, which defines the format plus a dictionary that defines the meaning of keywords in the file that are not part of the standard. The figure below indicates that the Representation Information is used to interpret the *Data* Object to produce the *Information* Object; something which one can understand then and use.

Figure 6-3. Representation Information.

The OAIS definition of Information Object is: *A Data Object together with its Representation Information***.** This is a very broad definition. When referring to something targeted specifically for preservation, the term Content Information is used. This is *information that is the original target of preservation or that includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information***.** In a little more detail, recognizing that the Data Object could be either digital or physical, one can draw Figure 6-4, which is a simple unified modelling language (UML) diagram. For further information see http://www.uml.org. Figure 6-4 shows that:

- an Information Object consists of a Data Object and Representation Information;
- a Data Object can be either a Physical Object or a Digital Object. An example of the former is a piece of paper or a rock sample;
- a Digital Object consists of one or more Bits;
- a Data Object is interpreted using Representation Information;

- Representation Information is itself interpreted using further Representation Information.

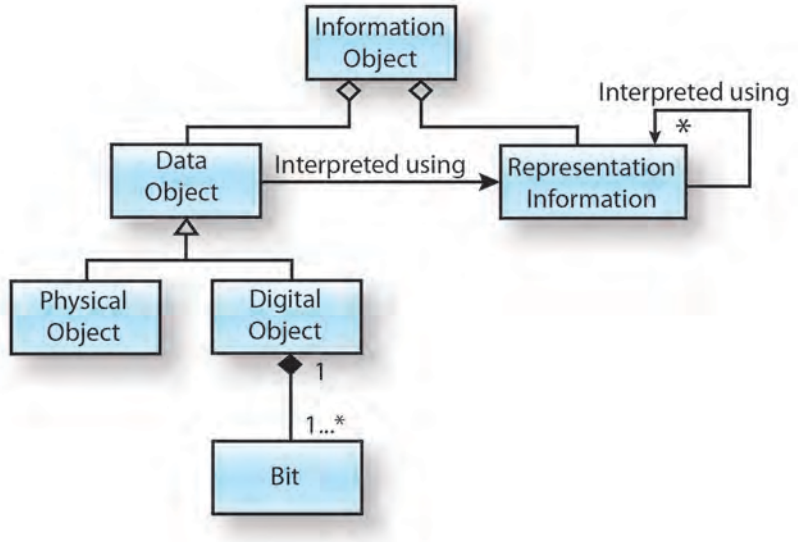


Figure 6-4. OAIS Information Model.

Figure 6-5 denotes that Representation Information may be sub-categorized usefully into several different types; namely Structure, Semantic and (the imaginatively named) Other Representation Information. This breakdown is useful because Structure Representation Information is often referred to as "format"; Semantic Representation Information covers items such as ontologies and data dictionaries; Other Representation Information is a catch-all for everything else.

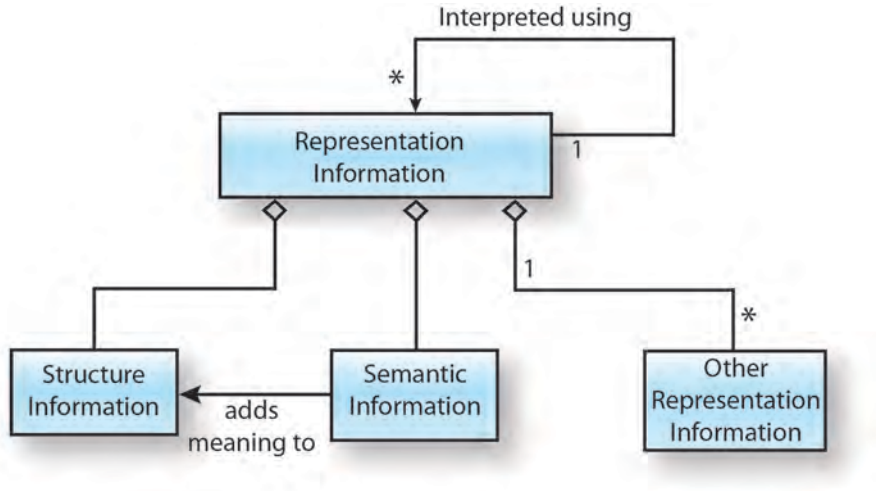


Figure 6-5. Representation Information Object.

When this is coupled with the fact that Representation Information is an Information Object that may have its own Data Object and other Representation Information associated with understanding that Data Object, as shown in a compact form by the *interpreted using* association, the resulting set of objects can be referred to as a Representation Network.

In the extreme, the recursion of the Representation Information will ultimately stop at a physical object such as a printed document (ISO standard, informal standard, notes, publications etc.). This allows one to

connect to the non-digital world. However, use of something like paper documentation would tend to prevent "automated use" and "interoperability". Also, complete resolution of the complete Representation Network (discussed in Section 1.1.3.2) to this level would be an almost impossible task.

As the final part of this OAIS concepts overview, we turn to '*how many metadata are needed?'* A piece of Representation Information is just another piece of Information; hence the name *Representation Information* rather than *Representation Data*. For there to be enough Representation Information it has to be understandable and usable by the Designated Community to understand the original data object. However, what if this is not the case?

The Representation Information may be encoded as a physical object such as a paper document, or it may be a digital object. In the latter case, one can simply provide Representation Information for that digital object. If the Designated Community still cannot understand and use the original data, one can repeat the process. This provides a way to answer the "how much" question; that is, the network provides representation information until there is enough for the Designated Community to understand the Data Object.

OAIS defines Representation Network as *The set of Representation Information that fully describes the meaning of a Data Object. Representation Information in digital form needs additional Representation Information so its digital forms can be understood over the Long Term.*

To complete the picture, we can then see a way to define the Designated Community, namely we define them by what they know (OAIS calls this their Knowledge Base); that is: *A set of information, incorporated by a person or system that allows that person or system to understand received information.* This definition includes tacit knowledge as well as formal qualifications and available software. Clearly this is impossible to capture completely; however, a broad description such as "an English speaking university EO researcher" is workable.

- Origins, context and restrictions (Preservation Description Information)

OAIS defines types of *"metadata"*, under the name of Preservation Description Information (PDI) (Figure 6-6), which broadly concerns knowing what and from where the digital object came.

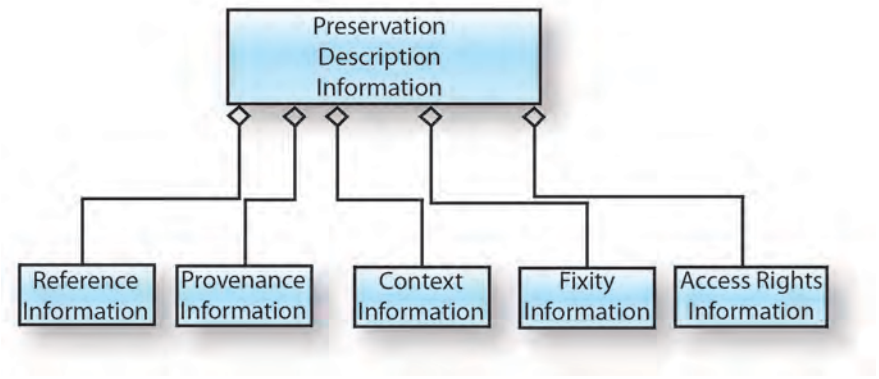


Figure 6-6. Preservation Description Information.

The idea is that one needs to name a way to identify the digital object; to know how and by whom and why the digital object is what it is; to know the broader context within which it exists; to be sure that the digital object has not been changed; and, finally, to know what rights are attached to it. Components illustrated in Figure 6-6 include:

- Reference Information: The information that is used as an identifier for the Content Information. It also includes identifiers that allow outside systems to refer unambiguously to a particular Content Information. An example of Reference Information is an ISBN. Clearly what are called persistent identifiers provide a form of Reference Information.
- Provenance Information: The information that documents the history of the Content Information. This information identifies the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. The archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity. Provenance may reasonably be divided into what we might term Technical Provenance, items that, for example, are recorded fairly automatically by software. This must be supplemented by Non-technical Provenance, for example the information about the people who are in charge of the Content Information – the people who could perhaps fake the other PDI. In other words, to judge whether we can trust the multitude of information that surrounds the Content Information, we must be able to judge whether we trust the people who were responsible for collecting it, and who may perhaps have been able to fabricate it.
- Context Information: The information that documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects**.** It is worth noting here that many traditional archivists would say that "context" is entirely important and trumps all other considerations. OAIS defines "context" in a rather more limited way, but on the other hands provides a greater level of granularity with which to work, although it does point out that Provenance, for example, is a type of Context.
- Fixity Information: the information which documents the mechanisms that ensure the Content Information object has not been altered in an undocumented manner. An example is a Cyclical Redundancy Check (CRC) code for a file. Digests are used often for this purpose, relying on the fact that a short bit sequence can be created (using one of several algorithms) from a larger binary object which it represents (essentially uniquely). By this we mean that it is, practically speaking, impossible to design a different file with a matching digest. This means that if we can keep the (short) digest safely then we can use it to check whether a copy of a (perhaps very large) digital object is what we think it is. This can be done by recomputing the digest, using the same algorithm, using the digital object which we wish to check. If the digest matches the original one we carefully kept then we can be reasonably sure that the digital object has the same bit sequence as the original.
- Access Rights Information: the information that identifies the access restrictions pertaining to the Content Information, including the legal framework, licensing terms, and access control. It contains the access and distribution conditions stated within the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer).

It also includes the specifications for the application of rights enforcement measures. Examples of PDI from different disciplines are given in OAIS.

- Linking data and "metadata" (Packaging)

The idea behind packaging is that one must somehow be able to bind the various digital objects together (Figure 6-7). Remember that Content Information is the combination of Data Object plus Representation Information, and PDI has various components. The figure below shows the other conceptual components of a package.

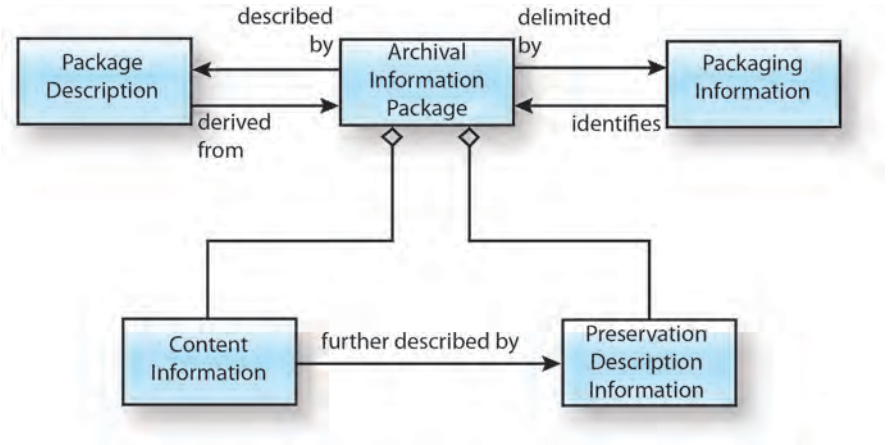


Figure 6-7. Information Package Contents.

It is important to understand that the package does not need to be a single file. The package is a logical construction; in other words one needs to be able to have something that leads one to the other pieces. One needs to be able to identify the composition of the package. Is it a file, a collection of files, or a sequence of bytes on a tape? This information is known as the Packaging Information. OAIS defines Packaging Information as "*The information used to bind and identify the components of an Information Package.* For example, it may be the ISO 9660 volume and directory information used on a CD-ROM that provides the content of several files containing Content Information and Preservation Description Information. The Information Package could indicate that the package is presented as a ZIP file (the file name extension probably is ".zip").

Additionally, the package has content that is described in the Package Description, which can be used to search for that specific package. The reader may notice that the additional concepts identified so far are called "Information". In most cases these will be digitally encoded. This leads us to a fundamentally important point discussed in section (2.1.3.2).

2.1.3.2 Recursion – A pervasive concept

An example of recursion is shown in Figure 6-8 which illustrates that a piece of Representation Information should have Provenance associated with it, while Provenance, which will be encoded in some way, must have its own Representation Information so that it is understandable and usable. A formal way of expressing this in OAIS is by showing that many of the concepts used are Information Objects (Figure 6-9).

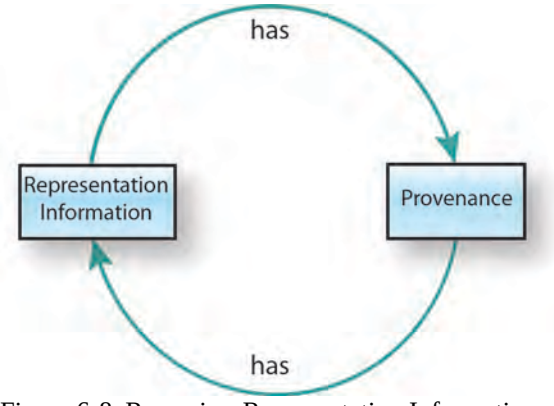


Figure 6-8. Recursion-Representation Information and Provenance.

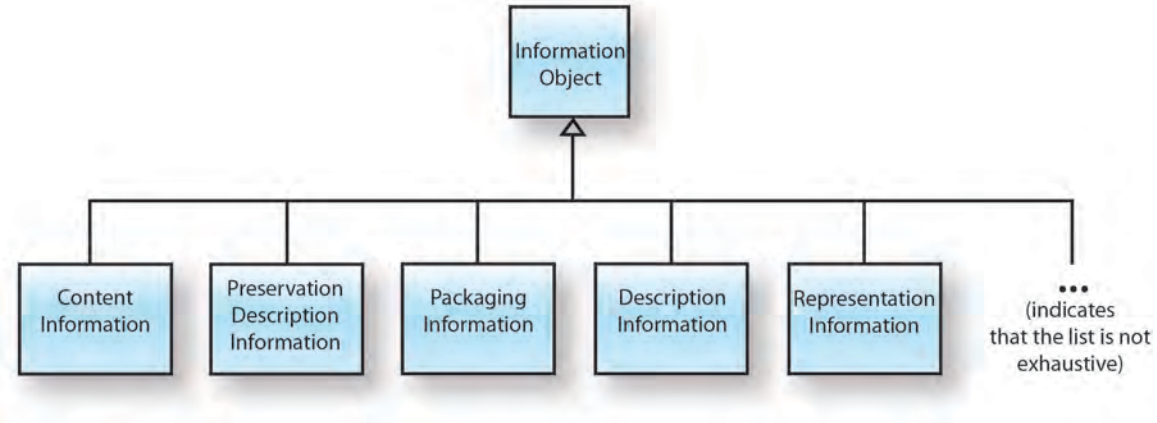


Figure 6-9. Sub-types of Information Object.

2.1.4 *OAIS Conformance, Audit, and Certification*

Conformance to OAIS is defined in the standard as supporting the Information Model and fulfilling the following mandatory responsibilities:

- Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.
- Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no *ad-hoc* deletions.

- Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.

While OAIS provides the key concepts of digital preservation, it is not written in a way that facilitates auditing. Instead, the standards ISO 16363 (2012a), *Audit and Certification of Trustworthy Digital Repositories*, and ISO 16919 (2014), *Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories*, (ISO, 2012b and ISO, 2014) were published. These were based on OAIS conformance supplemented by ideas about organizational sustainability and security. Further information is available at http://www.iso16363.org.

To create a system for auditing and certifying repositories according to ISO 16363, within the ISO system, a further standard was needed to specify who and how the audits and certification are done. The standard to address this has been published (ISO, 2014) which is the specialization of the ISO hierarchy of standards ensuring that the audits and certification decisions will be carried out with impartiality, competence, responsibility, openness, confidentiality and responsivity to complaints. The specialization sets out the specific competences that auditors must have.

2.1.5 *Functional Model*

Figure 6-10 is used often to represent the top level OAIS Functional Model. It shows six functional entities (Ingest; Archival Storage; Data Management; Administration; Preservation Planning; and Access) and related interfaces. Only major information flows are shown. The lines connecting entities identify communication paths over which information flows in both directions. The lines to Administration and Preservation Planning are dashed only to reduce diagram clutter.

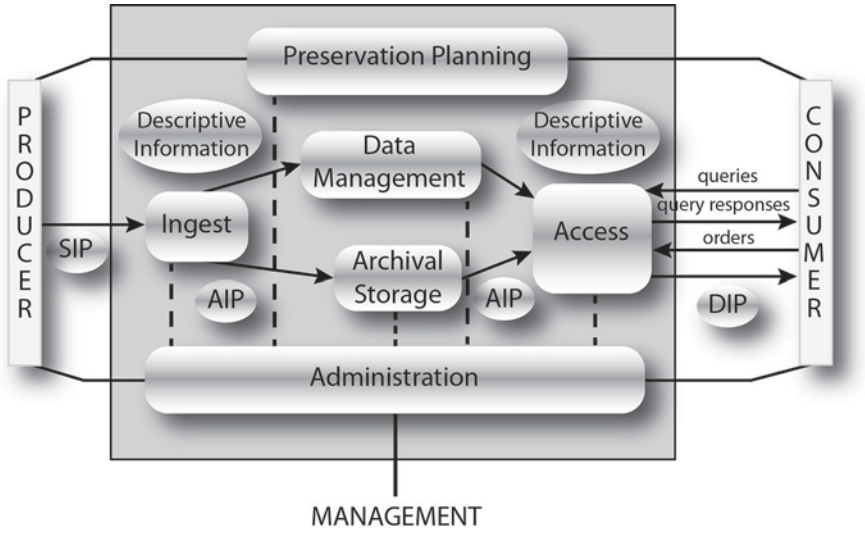


Figure 6-10. OAIS Functional Entities.

Many more details are provided in OAIS to create a rich collection of terminology and a checklist of ideas, which an archive should consider. The level of detail sometimes misleads readers into thinking that it is a design for an archive, or that the various functions are mandatory. Neither is true.

Mapping a repository to the OAIS Functional Model is neither necessary nor sufficient for conformance to OAIS. However it is extremely useful for describing the repository in a way that makes it easier to inter-compare repositories and to ask questions about what may be missing in terms of functionality. There may be good reasons for certain functionalities to be missing.

Despite the widespread use of OAIS, there are many collections of terms related to digital preservation used by different groups, organizations, and countries. It is difficult sometimes to understand how these terms relate to each other or to OAIS terminology. To address this source of confusion, a glossary has been created (APA 2014a), based on a Simple Knowledge Organization System (SKOS, 2006), that shows relationships between terms from six different collections and OAIS using the "broader", "narrower" and "related" relationships. It also provides a Uniform Research Identifier (URI) for each term: http://www.alliancepermanentaccess.org/index.php/consultancy/dpglossary/#Representation_Information (last date accessed: 11-20-2015).

2.1.6 *OAIS and Earth Observation Archives*

A few specific issues have arisen in Earth Observation (and other) archives. Some of these have been discussed in terms of limitations of OAIS; that is, areas where OAIS should be extended. The question is whether there are fundamental changes needed to the core concepts, or whether new terminology is needed. In the latter case, one can ask if additions are needed within OAIS, itself, or if entirely new or supplementary standards are needed.

2.1.6.1 Distributed Archives

Many archives are distributed – using distributed storage and/or contractors for parts of the storage operations. Suggestions have been made that distributed archives cannot be covered by the OAIS Reference Model and therefore OAIS must be replaced by an entirely new standard. Is this true? The answer is an emphatic no. Those who make these suggestions believe that the Functional Model provides the most important picture; that is, a picture of a monolithic organization. That view is incorrect. As we have seen, the Functional Model is not relevant for OAIS conformance. The Information Model and the mandatory responsibilities provide the key concepts, and careful reading reveals no requirement for a monolithic organization.

The Functional Model is used to provide useful terminology. In fact neither the terminology nor the diagrams imply a monolithic organization. Yes, OAIS introduces a few special terms related to federated archives. It is probable that the terminology does not cover all possible technological implementations – nor was it intended to. Therefore it could be useful to extend the terminology, either within OAIS itself if new concepts become as ubiquitous as Administration, Bits, and Consumers, or within a separate "daughter" standard if the terminology is sufficiently specialized.

For example OAIS itself, from its first publication in 2002, contained a "roadmap" of follow-on standards needed to address specific issues, and many of these, including ISO 16363 and ISO 20652 (ISO, 2006) have been published already. Further discussion of OAIS developments is available at http://www.oais.info.

2.1.6.2 Linked Data

An issue related to distributed archives is Linked Data (Berners-Lee, 2006) where data from many sources can be linked together from outside any organization structure, distributed or otherwise. As Linked Data are used increasingly, the question arises – how can these data be preserved? A detailed study has been carried out in the PRELIDA project (PRELIDA, 2014). This study shows that while there are many practical issues, the OAIS concepts provide perfectly adequate conceptual underpinnings.

2.1.6.3 Sustainability

The importance of data about the Earth, and especially long time series of unique data, is hard to overestimate and so its preservation is naturally a major concern. University libraries/archives have been described as the natural home for preserving digital objects because of the longevity of the parent organization and the mandate of the library/archive.

However the sheer volume and variety of Earth Observation and other "big science" data makes their preservation by any organization a great challenge with current technologies. It seems possible that such data will be handed from one organization to another and so the significance of the definition of "Long term" in OAIS becomes obvious – each link in the chain of preservation must capture the relevant information required by the Information Model and be prepared to hand Archival Information Packages to the next archive.

A view of digital preservation (Figure 6-11) within a broadly defined business process has been developed by the APARSEN project (APA, 2014b). This argues that there is a natural chain:

- Preservation is closely related to…
- Usability, on which…
- Value is based; and from which…
- Business cases can be developed and implemented through…
- Business models that produce benefits that, among other parameters supports preservation.

Figure 6-11. Integrated view of digital preservation (Source: APARSEN Project).

The key idea is that preservation based on OAIS requires working with the concepts of *Designated Community* and their *Knowledge Base*. However a repository can use these concepts to make their holdings useable to a broader community; for example by adding appropriate Representation Information, even if

not guaranteeing usability over the long term for that broader community – unless the definition of the Designated Community is changed. This potentially increases the value and hence the opportunities for exploiting business opportunities in the broadest sense. In this way a repository can address some of the key challenges of the Blue Ribbon Task Force (BRTF, 2010) on the economic risks to sustainability. Positioning preservation within an economic process also addresses key challenges identified in the report of the High Level Expert Group on Scientific Data (HLEG, 2010).

2.1.6.4 Authenticity

As part of the hand-over of information between archives and from the data creator(s), the evidence supporting claims of authenticity must be captured. This evidence may be technical, in particular Provenance about all the hand-over and preservation activities, including the associated Representation Information of those Provenance data. There also may be non-technical information such as those responsible for the preservation, as well as the security of the content, including among others the system log files which could contribute to the evidence that the data have not be tampered with or otherwise corrupted. Specific examples of such evidence are given in ISO 16363:2012 (ISO, 2012b).

2.2 *Retention*

Remote sensing and related records require retention controls to help ensure that unique and non-repeatable data and products are retained and preserved over time, in sustainable formats, and are available for continuing use now and into the future.

Retention is the length of time a record must be kept because it is needed for ongoing business, to document an action, or for statutory reasons. The establishment of systematic retention practices for organizational records can help realize economy and efficiency of operations by removing useless records and freeing up physical and virtual space; pre-identifying and protecting records of historical value that need to be preserved well beyond current business needs; making available those records that are needed to comply with business, legal, and scientific needs; and avoiding legal risk by complying with laws, regulations, and industry standards that govern the retention of records.

2.2.1 *Authority*

The legal foundation for retention is found in public laws and regulations that either mandate specific periods of time that certain categories of records must be kept; or, that require the establishment of a records management program that includes record retention policies. For instance, the U.S. Code of Federal Regulations (Part 1220.34) requires that all U.S. Federal agencies develop records retention schedules for all records created and received by the agency; and to obtain National Archives and Records Administration (NARA) approval of the schedules prior to implementation. Additionally, ISO 15489-1 (2001) Information and Documentation—Records Management, Part 1: General, says that a comprehensive records management program should include provisions for ensuring that records are retained for as long as needed.

2.2.2 *Development Criteria*

The 'record series' is the common unit by which records are managed, analyzed, inventoried, and ultimately retained prior to their final disposition. A record series is a group of logically related files that support specific business functions and which consist of multiple documents, folders, and items that are used and filed together (Saffady, 2004). In terms of electronic records, the scope may involve assigning retention periods in the context of an electronic information system (such as a database environment) with input, output, master file, and documentation. Retention is assigned to each of these system components.

Most records are temporary, meaning that one day they will be subject to destruction or other disposition action, but determining specific record retention periods should follow a systemic analysis of the value of each records series. For the majority of records there are three such values: administrative, legal, and fiscal (Franks, 2013). *Administrative* refers to how long the records are needed to perform current business and conduct operations. A permit for right-of-way on a public land parcel may have value to the creating agency for only as long as the permit is in effect. This period can be predetermined and is reflected in the retention schedule that covers the permit record series. Certain remote sensing records created by mapping agencies like the aerial photography collections of the U.S. Department of Agriculture, may retain their administrative and operational value for many decades. Fiscal value concerns the length of time the records are needed to document financial transactions and obligations. For example, the standard retention period for Federal accounting and procurement records is six years (NARA, 2014). Legal value concerns how long an organization must keep records for legal purposes, not only for itself but for its clientele and other stakeholders. Real property ownership records often have lengthy retention periods depending on how long the property is legally owned by the organization. Skupsky (1991) indicates that "legal research will uncover the laws affecting the retention of records for legal purposes. Some laws clearly state the legal requirement (minimum retention). Other laws that affect records either do not state a specific retention or merely give a legal consideration, such as a statute of limitations.

A small percentage of records perhaps three to five per cent, depending on the nature and mission of the organization, possess secondary values. This means that they have use for others outside of the creating organization. For example, long after Federal population census information ceases to be useful to the U.S. Census Bureau in preparing its reports for Congress and for publishing its statistics, census records continue to be a rich source of family history information for researchers and genealogists. Organizations and archival authorities may assign permanent retention to records that are deemed to have sufficient historical or other value to warrant their continued preservation in an archival repository (NARA, 2013).

2.2.3 *Retention Types*

Decisions about the retention of specific records series are captured in a document called a Records Retention Schedule. The records retention schedule lists all records created by an organization and provides the detailed and compulsory instructions for what to do with records that are no longer needed for current business. Schedules are developed through various processes that include identifying the business functions that generate the records, inventorying records wherever they are found across the enterprise, establishing dispositions for records based on business and archival needs, risk analysis, and statutory and regulatory requirement, and finally, obtaining required formal retention concurrences. Traditionally, retention

schedules are formatted in a granular, item-by-item listing of every records series and corresponding retention period. A more recently developed strategy called "flexible scheduling" provides for concrete disposition instructions that apply to larger groupings of categories of records also called "retention buckets". Unlike traditional scheduling, bucket-type schedules cover all records that support a work process or program area, such as an environmental assessment process. All records for the particular process fall under a single retention period (NARA, 2004).

2.2.4 *Retention Considerations*

There are several considerations that affect record retention periods. One such consideration is volume or quantity on hand, and the projected growth of both physical and virtual records. A high-volume series of records is expensive to maintain over long periods, especially for analog records. It is important to validate the perceived administrative/operational, legal, and/or scientific value of the records to justify their long term storage and maintenance. Another consideration is the frequency of use of the record series within and outside of the organization. Knowing how often the records are used (and at what point they fall into disuse) helps provide a baseline of retention for a particular series. For many records the reference activity is highest just after the records are created and over time, the frequency of access begins to diminish such that documents are seldom if ever accessed for current business. However some records, such as those containing observational data or scientific research information, can suddenly invite researcher interest many decades after they are first created (NARA, 2007).

Long term accessibility to records is another consideration. According to the International Organization for Standardization (ISO), records should remain accessible and useable throughout their lifespan. This means that they can be located, retrieved, presented and interpreted until final disposition." ISO also recommends that "in developing retention periods for each record series, the organization needs to ascertain whether or not the equipment and documentation needed to read and interpret electronic records will be available years or decades hence" (ISO, 2001).

Finally, steps should be taken to assign retention to record metadata so that they, too, are preserved for at least as long as the record to which it pertains. Metadata provide the accompanying details about records including contextual data (e.g., time and date of creation; their logical structure; and information needed to understand the record).

2.2.4.1 Retention Language

When preparing a record series description, one must ensure that there is enough information to provide the essentials as to the form, content, and basic purpose of the record. In addition to listing the representative types of documents, forms, objects, or other entities that characterize the record, the description should also state the purpose of the records, and their relationship to the program's mission, administrative activities, and business operations (Saffady, 2004). Table 6-2 provides two examples of retention for remote-sensing series created by the U.S. National Oceanic and Atmospheric Administration.

Table 6-2. Sample Records Retention Schedule Items.

Record Title / Description	Retention Instructions
Aerial Film Negatives Original aerial film roll negatives, taken with Single-lens, high precision cameras mounted on aircraft, covering the shoreline and other areas of the United States, Puerto Rico, and possessions. Collection includes approximately 5,000 rolls of color and B/W film of oceans, shoreline, and airport areas dating back to 1943. Film has been scanned for distribution. This series includes a finding aid.	A. Original negatives and related finding aids. Permanent: Transfer immediately to the National Archives the original films to which NOAA has made an in-house reference copy. Transfer all subsequent originals as NOAA continues to make reference copies. For those films where a reference copy is not made, transfer original films to the National Archives in 5-year blocks when the most recent record block is 40 years old. Transfer related indexes and finding aids with each block. B. Digital Distribution Copies. Temporary: Delete when no longer needed for reference or dissemination.
***In Situ* & Remotely Sensed Environmental Data Master Files** The data record environmental phenomena near to, and distant from, the location of the instrument. Metadata about the station's purposes location and instrumentation also are included. These data are a source of information on environmental parameters, such as weather patterns, vegetation and land cover, human activity, ocean climates, and geophysical descriptions of Earth phenomena.	Cut off files at the end of the calendar years in which the data are no longer needed for immediate/current research Destroy/delete 75 years after cut-off upon approval by NOAA and NESDIS stake-holders. A longer retention may be necessary for research purposes.

NARA 2001 and 2008. NOAA Request for Disposition Authority

For electronic records and for the more fragile analog formats, reference should be made in the schedule to any preservation activities that will be required during the retention period, especially for records with long-term temporary and permanent retentions. This means having built-in processes and operations to ensure the technical and intellectual survival of authentic records through time. Special media records and permanent records in all formats need to be stored in temperature and humidity controlled space to extend their life expectancy. Digital storage media in particular will deteriorate without special controls. Records will have to be copied to new media while they are still readable and moved (migrated) from one system to another during their retention period, while maintaining the records' authenticity, integrity, reliability, and utility (ISO-19011, 2011). Sustainability of records (the ability to access an electronic record throughout its lifecycle, regardless of the technology used when it was originally created) is critical when deciding on how to retain long-term, technically complex records. (NARA, 2007).

2.2.4.2 Retention Implementation

Conducting a full-scale inventory, appraising the value of each record series created by an organization, and memorializing this information in a records-retention schedule is a waste of time if the schedule is never implemented within the organization. Lack of implementation can also present legal risks especially when litigation reveals instances of irregular, inconsistent, and ad hoc use of the schedule to destroy records. This can lead to accusations of spoliation of evidence. Implementation should take place in the normal course of business by all units in the organization. Activities include the physical transfer of inactive analog records to offsite storage, or placing electronic records into an "archived" status; destroying and deleting records that have reached the end of their authorized retention period, have no continuing value, and are not required for legal holds and litigation; and the transfer of records of permanent value to an archival

authority or institution, such as the National Archives of the United States for Federal government agencies. In the Federal government, the Office of Management and Budget requires that all Executive Branch agencies to incorporate records management and archival functions into the design, development, and implementation of information systems, including retention (OMB, 2000). According to Circular A-130, Par. 8a(1)(k), agencies are required to provide training and guidance to all employees on their records disposition requirements and procedures and other significant aspects of the records disposition program. When a new or revised records schedule is issued, agencies are to provide specific guidance to employees responsible for applying the schedule (NARA, 2009).

2.2.5 *Retention Application*

Until recently, the systematic application of retention to records has been accomplished through predominantly manual processes; hardcopy schedules, paper forms, storage and retrieval of paper records staged in warehouses, etc. Today, there are number of programs and software products that accomplish the same disposition processes through automated means, especially where the records are "born electronic" and are captured in an electronic recordkeeping system. Examples of automated retention functions include: suspending retention period of records beyond their scheduled disposition in the presence of ongoing litigation; copying pertinent records and associated metadata to facilitate their transfer to a long-term storage or archival repository; identifying and presenting records that have met the end of their retention periods; electronic approval of disposal notification forms; and deleting electronic records approved for destruction in a such a way that physical reconstruction is impossible. Those responsible for carrying out retention must follow the instructions strictly in the authorized retention schedule. Using the example in Table 6-2, NOAA staff would periodically review, identify, and segregate original aerial film negatives that have reached 40 years of retention. When five years' worth of these negatives have accumulated, NOAA completes, signs, and transmits a formal instrument to the National Archives, effectively authorizing the transfer of legal and physical custody. Having been placed in proper containers along with associated indexes and finding aids to the discrete set of negatives, the records are sent to the National Archives. From then on, the National Archives stores, preserves, and makes available the NOAA aerial film to any researchers interested in the records.

2.3 *Long-term Data Preservation Challenge*

Earth observation data inherit the preservation challenges that apply to digital information in general, including the rapid rate of technological change and the potential for loss-of-utility of preserved data in the absence of sufficient documentation or the means to effectively render the data. In the case of Earth observation data, additional challenges are raised by the sheer volume of data, as well as by the presence of unique and often proprietary data formats (McGarva, 2009). Given appropriate care, analog formats are easy to handle and render information in such a way as to make it accessible, immediately, and have proven to be effective media for long-term preservation. Digital data, however, are produced in formats that have a very strong reliance on technology to both manage and access information (Digital Preservation Collection, 2015). The risk of losing data is much higher in the absence of forward-thinking on the part of data producers, as well as periodic preservation-oriented interventions designed to ensure persistent access to

data. Data that are described insufficiently, or are in obsolete formats, become unusable and risk being destroyed (FGDC, 2014a and 2014b).

Technological concerns aside, risk also arises from the reality that the preservation of older data is not always a priority for data creators; hence the creation of the Geospatial Multistate Archive and Preservation Partnership (GeoMAPP). Limited resources, budget constraints, or an insufficient understanding of the benefits of preserving older data can limit efforts to implement data preservation plans. As a result, older data are extremely susceptible to permanent loss, either through outright loss, or through loss of the facility to recover, access, interpret, and use the data successfully (GeoMAPP, 2011a; 2011b; and 2011c).

2.3.1 *Defining Long-term Digital Data Preservation*

Digital data preservation refers to the series of managed activities, policies, strategies and actions needed to ensure continued access to, and accurate rendering of, digital data for as long as necessary, even in the face of media failure and continuous technological and organizational change (IFDO, 2015). It is impossible to predict what technology will exist several decades into the future, or the manner in which data produced today might be accessed and utilized. Nonetheless, there are measures that can be taken immediately to mitigate the risk that data are lost, and to increase the likelihood that the data will remain discoverable, accessible, and usable in the future.

2.3.2 *Data Preservation Planning*

Data producers, data custodians, and archivists need policies designed to establish practices for data stewardship that ensure the quality, integrity, confidentiality, and security of digital resources over time. The policies should address the requirements of data organizations as well as future users, and agreements should be established to ensure that the policy is adhered to over time. The output of the archival planning process should, itself, be preserved over the long term to support future preservation efforts.

2.3.2.1 Building partnerships

The process of archiving and preserving digital Earth observation data should begin ideally at the point of data creation, not at the time of transfer to an archive. Collaboration across organizations, including data producers, data custodians, archivists, funding agencies, and standards bodies, is a key component in establishing a unified approach to preservation. Due to the complexities of Earth observation data and the unique processes required to preserve them, organizations seeking to preserve data stand to benefit from working in partnership to establish practices and policies for preservation. A high level of collaboration helps to prevent duplication of efforts and can lead to cultivating a common understanding that supports the development of recommendations, the implementation of policies and systems, and informs the establishment of best practices and standards (GeoMAPP, 2010).

2.3.2.2 Developing the business case

Assessing the value of preserving data for unanticipated future applications is not done easily. Nonetheless, support for digital preservation efforts can be strengthened by developing a business case that describes the value of historical and temporal data; adequately captures both the tangible and intangible benefits of preserving older data; and identifies the risks of inaction. Use cases can be employed to tell compelling

stories about why it is important to preserve certain Earth observation datasets and to help clarify the value of these resources (GeoMAPP, 2011a and b). The importance of data preservation efforts become clear to decision makers when they understand that providing sustainable policy and funding support for preservation activities is in alignment with organizational missions (FGDC, 2014).

2.3.2.3 Identifying data

The first step in selecting candidate data for long-term preservation is to determine what data actually exists. Data inventories can provide an assessment of what data are available and in what format; when they were created and by whom; who is responsible for them; and where the data came from. All of these elements are essential for subsequently selecting and appraising which data are currently at risk and which need to be considered for long-term preservation and access (GeoMAPP, 2010)

2.3.2.4 Selecting and appraising data

Long-term preservation is an economic issue as well as a technological challenge. Due to limitations in storage and processing capacity, organizations may not be able to preserve every dataset in their holdings and will often need to be selective, deciding what to archive and what to discard. To aid in the selection of data for preservation, organizations should develop a formalized, documented appraisal process to assess data for their archival worthiness based on legal, historical, business, and research value. As long-term benefits may be intangible, it may be necessary also to focus on short and medium term benefits (NARA, 2007; NDSA, 2013; FGDC, 2014).

2.3.2.5 Retention schedules

Records retention schedules can be an effective mechanism for ensuring that data worthy of long-term preservation are retained and transferred to the archives. Formal records retention schedules prompt data producers to think about what information they produce, which data need to be preserved, and how to make these data useful to others over the longer term. The retention period expected for each individual dataset or product should be defined, and the documentation that explains what has been selected for retention and why should, itself, be preserved (GeoMAPP, 2010; FGDC, 2014).

2.3.3 *Initial Archival Phase*

Digital data preservation systems must ensure the authenticity, reliability, integrity, security, and usability of the data. Data producers and custodians should define and document archival processes for ingesting, processing, and making data available, and identify mechanisms and tools to support the preservation of data to support access over the longer term. A number of measures can be taken in the effort to make archived data more adaptable to future, unknown technological environments and to increase the likelihood that data will be usable in the future.

2.3.3.1 Encapsulating data

Data resources should be self-contained and independently understandable. Data retained for the longer term should preferably be maintained in open, file-based systems rather than databases or complex computing environments (Rönsdorf, 2014).

2.3.3.2 Selecting data format

The issue of data format sustainability comes into play when addressing technology and obsolescence risks. A sustainable format increases the likelihood of data being accessible in the future. Sustainability criteria should be identified and applied, with priority emphasis placed on factors such as format openness, community uptake, data portability, and the ease of data migration and data access (Library of Congress, 2014; GeoMAPP, 2011a; GeoMAPP, 2011b). Efforts must be made to balance between the goals of capturing complex functionality and of ensuring the longer-term usability of the content, which might be achieved by using simple models. It may be desirable to restrict the number of formats and encodings to a widely agreed upon set of open and well-documented file formats. For long-term preservation, binary encodings should be avoided as much as possible, and any compression or packaging formats should be open, well documented and widely used (Rönsdorf *et al*., 2014).

2.3.3.3 Metadata and documentation

Preserved data should be accompanied by complete, standards-compliant metadata containing information that is critical to understanding the dataset for current use, as well as to enable discovery, rendering, comprehension, and use of data over the long term. At the point of data production, assigning a logical file name that incorporates such attributes as *geographic extent*, *data theme*, and *creation date* will aid in identifying and managing those data (GeoMAPP, 2011). Retention of a graphical representation alongside the logical representation of the data will increase the likelihood of at least a minimal understanding of their content when examining or evaluating them at a future time. Preservation of pertinent data specifications, standards, and definitions of coordinate systems also aid in future efforts to interpret and understand the data (Rönsdorf *et al.*, 2014).

2.3.4 *Long-term Preservation*

It may be useful for data custodians and archivists to consider preservation timeframes that account for short-, medium-, and long-term needs. The EuroSDR Archiving Working Group has suggested that preservation timeframes of one, ten, and one hundred years be considered, with different archival solutions and approaches for each frame:

- A short-term (one year) archive would focus on operational safekeeping, addressing short-term needs for present day management and access. Solutions may favor robust functionality needed for immediate use in the original context of the data.
- A medium-term (one decade) archive would focus on reusability of, and access to, data in a broader context, and over time to build a bridge between shorter-term data provider needs and archival needs.
- A long-term (one Century) archive would aim to conserve data and include robust measures to protect against corruption, or loss of utility of the data. In this case, data would preferably be held in simple, open formats (Rönsdorf *et al*., 2014).

2.3.4.1 Redundancy

To support long-term preservation, multiple copies of data should be retained in diverse storage systems that track the location and integrity of each dataset. Ideally, copies of the data would be stored in different

technology environments, housed by different organizations, and based in different geographic regions to protect against data loss that might result from failure of a particular technology, dissolution of a particular organization, or disaster impacting a particular geographic location.

2.3.4.2 Monitoring and auditing processes

To support long-term preservation, archivists must take proactive steps to keep hardware, software, metadata, and data formats current in order to minimize the potential for loss of digital data resulting from media or system failures, or from technological obsolescence (FGDC, 2014a; FGDC, 2014b). Data custodians and archivists will need to track, manage, and audit the preservation status of data. Monitoring processes should include integrity checks on the data to ensure at least bit-level integrity, and strategies for dataset migrations and for refreshing media should be considered.

2.3.4.3 Format migration and emulation

While archived data may be refreshed by copying data from older, less stable media and systems to newer, more stable media and systems, this action does not guarantee that the data will continue to be usable in new software and computing environments. Additional approaches to long-term preservation include format migration and emulation. If data created at earlier times cannot be read and used in newer environments, migration of those legacy data to a new format will almost certainly be required. As an alternative, emulation, which involves re-creating or sustaining an earlier computing system, may be attempted to allow for the use of data in their original software environment.

Given the likelihood that original data formats will not be supported and readable at some point in the future, some combination of migration or emulation may well take place. Since emulation is still in the early stages of development, format migration is likely to play a key role in efforts to keep data readable and usable over a long period of time. A key challenge in format migration lies in retaining the ability to display, retrieve, manipulate and use data in their original context while also making them useful in newer software and computing environments. Data custodians and archivists must be prepared to choose which significant properties of data to preserve as format migrations are implemented. Amendments to archived data, and the rationale behind preservation actions, should be recorded (Rönsdorf *et al*., 2014). The original data should be retained alongside newer, amended versions if at all possible.

2.3.4.4 Transfer

Over time it may be necessary to transfer data from one archive to another. In such cases, it is necessary to define data transfer validation mechanisms and procedures to review and validate data both before they are transferred and after they arrive at the receiving archival organization. Key factors to consider include virus checking, bit-level verification, functional review of the data using available software, and metadata validation (GeoMAPP, 2011a; GeoMAPP, 2011b; GeoMAPP, 2011c).

2.3.4.5 Providing long-term access

The end goal of long-term preservation is to support long-term data access and use. Preserved data resources should be made available for discovery on the open web and via general and domain-specific data portals or clearinghouses. As preserved data are discovered, accessed, and used, they may gain more value

for future user communities, and support for continued preservation efforts. To this end, undiscoverable and inaccessible "dark archives" should, as much as possible, be liberated in order to be openly accessed and used.

2.4 *Appraisal*

The practice of *appraisal* has its roots in archiving and records management disciplines. Utilizing this practice, even informally, is recommended to remove remote sensing and geospatial data that do not meet an organization's needs. Appraisal simply means to review and assign value to a collection of records. A review of records ensures that the data in question continue to align to the mission or objectives of the organization holding the data. Archiving, at a minimum, involves preservation and access activities, which require resources. Incorporating appraisal practices allows organizations to help determine where their resources are best expended.

Geospatial content plays a significant role in a wide range of applications that support planning and decision-making across a broad range of activities. While many applications rely on the most current available content, there is increasing demand for older content to support historical and temporal analyses related to change in Earth's natural and human landscape, including physical infrastructures. There are challenges to selecting and appraising digital geospatial data for their long-term value, but there are common elements that can be shared across government at all levels, academia, and the public and non-profit sectors to guide them in making proactive decisions on stewarding digital geospatial data.

2.4.1 *The Value Proposition for Appraising, Collecting and Preserving Geospatial Data*

Geospatial content, information and data (also known as geodata) includes resources such as geographic information system data sets, digitized maps, remote sensing data resources and tabular data that are tied to specific locations. Digital mapping provides the dramatic ability to create visualizations that resonate in our imaginations and enable us to tell compelling stories of change over time (Lazorchak, 2010), and this digital geospatial information has long-term, enduring value.

Geospatial content plays a significant role in a wide range of applications that support planning and decision-making in a broad range of activities. While many applications rely on the most current available content, there is increasing demand for older content to support historical and temporal analyses related to change in Earth's natural and human landscape, including physical infrastructures. Examples of applications that require historic content include climate change, disaster planning and environmental impact analysis, industry site location planning, and resolving legal challenges.

For some users, digital geospatial data lose value the minute they cease to be current. An analogy is found in the automobile industry. Motor vehicles lose "book value" the minute they leave the dealer's lot. But a car that runs well still holds significant value for its owner. Even a no-longer operational car that exists only in its scattered parts and documentation still holds value in that it tells us quite a bit about the transportation infrastructures, mechanical and computational advances and consumer preferences of its era.

In the same way, legacy and historical geospatial data store significant untapped value. Some of their value exists by default, through the mandates of legal regimes and records retention policies designed to

ensure that governments and organizations have a factual basis to support decisions (some decades old) and retain a record of government activity to support continuity of operations or disaster-recovery efforts.

Other value resides almost entirely in the material as historical records. The ability to do temporal analysis for land-use changes, property values, climate modeling, and such relies on having historic data available for comparisons. The cultural heritage value of historical geo-data get most of the attention because of the sheer pleasure we get from looking at older maps and aerial photographs. But an exclusive focus on the cultural heritage value of historic digital data, while still significant, fails to take into account the significant returns on investment that can be achieved through the thoughtful stewarding of data across the lifecycle of information. There are numerous examples that showcase the potential value of historical data at the federal, state and local government levels, as well as across a range of academic, non-profit, and private-sector organizations.

The U.S. Geological Survey's Federal Geographic Data Committee (FGDC) efforts to establish a category of data known as National Geospatial Data Assets (FGDC, 2014a), and to recognize geo-data as investments, showcases the value the federal government attaches to these materials.

The 2011 final report of the Library of Congress supported Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) notes the long understood use of geospatial information in decision-making processes and planning efforts by state and local governments, as well as their recognition of the value of historical geo-data:

State governments are also recognizing the importance of having access to older data as they allow them to explore societal, environmental, and economic change geospatially over time. There are compelling business drivers to support the preservation of older data: tracking population, land, or vegetation changes over time; providing a cultural record of place over time; or avoiding the cost of re-creating datasets that were not preserved, are just a few of many drivers spurring users to seek out and use superseded geospatial content (GeoMAPP, 2010).

Value is often in the eye of the beholder, but the GeoMAPP Geo-archiving Business Planning Toolkit (GeoMAPP, 2011b) provides a collection of guidance documents and tools to assist planning and development of a geo-archiving business plan that may also help organizations develop a supportable value proposition for their historic geo-data.

2.4.2 *Organizational Challenges of Appraising and Collecting Geospatial Data*

Dynamic digital geospatial data are widely accessible from almost anywhere in the short-term; but they are fragile and at-risk of loss unless efforts are made to preserve and keep them accessible over the long-term. How will we ensure that digital records of our contemporary landscape will remain accessible in the future? (Lazorchak, 2010).

We now live in a world where we benefit from the full advantages of digital data creation: portability, replicability, storage space savings and more. We no longer need to build large buildings to house acres of paper. Instead we build (slightly) smaller buildings with ever-larger arrays of digital storage that hold monumentally greater amounts of information per square foot than ever before.

The benefits of digital data are enormous, but with increasing amounts of digital data come a new set of preservation risks, the most pressing being the dynamic technology environment in which we live. While

the dynamics of technology advancement provide their own benefits, they are the main source of preservation risk for digital information. For example, as storage densities increase, the media used to house digital data have changed, leading to the obsolescence of a number of storage technologies, including ¼" and ½" disks, Zip disks and soon, even optical compact disks. In addition to storage and hardware media obsolescence, format obsolescence is also a significant concern. Software companies go in and out of business rapidly, and the software format of one tool can rapidly become orphaned and unsupported in a few short years if the company goes out-of-business or the tool falls out of favor.

Additionally, digital data present challenges regarding their provenance (where did they come from and what has been done to them?); authenticity (are the data I'm working with be exactly the same as the original data?); and metadata (do the data have enough descriptive information attached so I can understand them?). These preservation challenges are common to all digital data, and they are being addressed through digital stewardship processes largely being developed by the library, archive and museum communities.

The 2014 National Agenda for Digital Stewardship published by the National Digital Stewardship Alliance offers that "digital stewardship" is the series of "managed activities, policies, strategies, and actions that ensure that digital content of vital importance to the nation is acquired, managed, organized, preserved, and accessible for as long as necessary. Digital stewardship activities protect important content in spite of changes in technology, economic sustainability, or institutional capacity" (National Digital Stewardship Alliance, 2013a).

In addition to the common digital preservation challenges faced by all digital data, digital geospatial content faces its own special set of challenges. Geospatial data resources typically are supported by backup plans, designed to ensure near-term retention of data, but they are infrequently addressed by archival plans that allow explicitly for longer-term retention of data; especially the capture and retention of superseded versions of current datasets (National Digital Stewardship Alliance, 2013). The NDSA report, "*Issues in the Appraisal and Selection of Geospatial Data*" lists a set of unique challenges that affect geospatial data.

2.4.2.1 Frequently or Continuously Changing Data

Dynamic data represent a moving target for archival capture. At what points and how often is it appropriate or necessary to capture dynamic data or samples of dynamic data for long-term preservation and access purposes?

2.4.2.2 Commercial or Proprietary Data Formats

Decisions to capture data in native formats may introduce dependencies on proprietary technologies, while efforts to normalize the same data in formats based on open standards may result in data loss or reduced data usability.

2.4.2.3 Spatial Databases vs. Individual Datasets

Spatial databases are composed of datasets in combination with relationships, behaviors, annotations and models. It is challenging enough to manage the packaging of these individual pieces into a coherent whole, but the overall management of these complex datasets involves using technology not regularly available to stewarding organizations. One solution is to capture database snapshots on a regular basis, but these snapshots may not survive long without active management. Another data capture approach involves extracting

individual datasets and converting them to more stable forms, but at the potential cost of making the data less usable, or possibly even deleting essential information.

2.4.2.4 Complex, Domain-specific Metadata Needed for Appraisal and Subsequent Use

Metadata are a key to the long-term preservation of all manner of digital data, and geospatial data are more complicated and more in need of description than average digital data. Geo-data do have the benefit of an existing data-sharing culture and infrastructure that supports the development of quality metadata, though this is by no means universal.

2.4.2.5 Variety and Complexity of Data Representation Methods and Data Derivatives

Digital maps are rarely as simple as their representational counterparts in paper form. Digital maps are complex aggregations of information, often created on the fly from widely disparate data sources, with widely varying methods of creation and production. These widely varying methods of production make it challenging to settle on common digital stewardship practices widely replicable across disparate organizations.

2.4.2.6 Scale and Resolution

Dataset options for any particular geospatial theme may be available at different scales or resolutions. The challenge for stewarding organizations is to determine which scale(s) and resolution(s) are most appropriate for them to acquire and preserve from a long-term perspective, and also to ensure that the relationships between different scales and resolutions of the same data are retained over time. In the case of raster data, higher resolutions incur significantly higher costs in terms of digital storage because of the larger file sizes of the data.

2.4.2.7 Data Volume and Capacity Limitations

Expanding file size, both in terms of total bytes of data and in the proliferation of datasets supporting millions of smaller files, introduces special challenges to the stewardship of geo-data. Spatial databases can be comprised of millions of individual data points, while collections of geospatial imagery can create very large datasets that are challenging to manage with conventional computing infrastructures. It is rare that stewarding organizations have the storage capacity or computing power to work easily with datasets of these sizes. The good news is that there are common appraisal and selection methodologies to assist organizations in addressing the preservation challenges of geo-data and provide stewardship of the information over time.

2.4.3 *Common Elements of an Appraisal and Selection Process for Digital Geospatial Information*

One way to systematically address the challenges of preserving digital geospatial data is by establishing methodologies to select and appraise key geospatial data at each stage of the digital stewardship lifecycle from creation to disposition. This is a relatively new concept in that archival activities in the paper world have often taken place at the end of the data lifecycle, almost as an afterthought. Because of the fragility of digital data, archival actions have to take place at each stage of the lifecycle, providing for the continual stewardship of digital data to ensure that it remains understandable and accessible over time.

The NDSA document referenced above is one effort to draw attention to the need to address archival processes much earlier in the lifecycle. The Federal Geographic Data Committee's Users/Historical Data Working Group has been working on draft guidance to help federal agencies and data stewards define geospatial content of enduring value. This guidance suggests possible priority approaches on how resources might be allocated to support long-term preservation and access through appropriate *Selection and Appraisal* (S&A) processes in a challenging funding environment. The remainder of this chapter draws heavily on the group's guidance, currently published only in draft form.

It is neither possible nor desirable to preserve every bit of geospatial information created. Selection is typically associated with libraries and other collecting institutions and provides a comprehensive method to evaluate the materials that make up an organization's collection. Appraisal is associated with archival and records management processes and is defined as the evaluation of information to determine its ongoing value and its merits for long-term or permanent retention.

The long-term management of traditional non-digital resources has been under stress for years because of the limited resources available across the government for these types of activities. The rapid pace of change of digital technologies and the exponential increase in digital data volume adds urgency to a call for reevaluation of S&A and the engagement of content creators, aggregators and intermediary data stewards as early as possible in the processes of identifying, evaluating, managing and preserving digital geospatial materials of long-term value.

S&A guidance identifies a range of stewardship concerns that need to be addressed across the lifecycle to ensure that valuable information of importance to the nation remains accessible and usable. Organizational focus has driven S&A decisions, with data producing agencies, data managing agencies, archives and libraries each making decisions according to their individual needs, but it is worthwhile to consider the utility of a broad, national, multi-organizational focus in addressing S&A decisions.

The common elements of S&A process, as described by the FGDC Users/Historical Data Working Group, are summarized below. The elements are "common" because they are applicable to any set of digital geospatial resources. Each element illuminates key component of a comprehensive S&A process.

2.4.3.1 Data Inventory

Data inventory should be one of the first steps in S&A and part of a regular, ongoing data management. Data inventories can provide an end-to-end view of what is available and what may be at risk in order to support stewardship priorities. Data catalogs are intended to support data discovery and sharing by end users, and may be populated by data inventories.

The federal government's Data.gov infrastructure provides an authoritative process for assisting federal agencies in inventorying valuable geospatial content. The GIS Inventory System maintained by the National States Geographic Information Council (NSGIC, 2014) is a tool used by state and local governments to inventory their geo-data and to provide access to data resources within a specific geographic or thematic domain. It provides capabilities to assess the quantity of existing data, current formats, stewarding responsibility, creation dates and data origins as well as the status of geographic information system implementations in state and local governments. The Random Access Metadata for Online Nationwide Assessment (RAMONA) database is a critical component of the GIS Inventory. The GIS Inventory automatically

generates metadata that is minimally-compliant with the Content Standard for Digital Geospatial Metadata published by the FGDC. It posts the metadata to a web folder that is harvested by the Geospatial Platform and Data.gov (FGDC, 2014b).

2.4.3.2 Alignment with Organizational Mission

Proposals for the acquisition, development, and ongoing stewardship of geospatial data should justify how the data are aligned with the mission of the organization. These proposals should include statements on the relevancy of the data to the objectives in the organizational mission statement; how the data assists in attaining the long-term goals described in strategic plans; how the data meets the needs of the designated community the organization serves; and how the data contributes to or complements current or planned collections to meet the mission and objectives of the organization over time. Mission alignment and relevance can often be determined by reference to agency strategic plans. Additionally, stewarding organizations such as libraries often have "collection development policies" that suggest the organizational mission of the collections that drives acquisition and stewardship strategies (FGDC, 2014b).

2.4.3.3 Legal Rights, Restrictions and Mandates

In addition to the legal statutes governing records retention, selection and appraisal of geospatial data should consider any limitations, restrictions or mandates placed upon the data and rights or constraints for dissemination that have been specified in licenses or legal documents. Furthermore, security and confidentiality concerns may need to be applied to protect individuals, property, wildlife, locations, or inhabitants.

Legal jurisdictions must also be considered. As part of the review, evidence of rights or restrictions should be attained, examined, and retained to justify any decisions that are based on the review. Furthermore, the constraints and rights associated with the data should determine how the data are accessed, used, or distributed. Legal rights, restrictions and mandates should be documented in metadata (FGDC, 2014b).

2.4.3.4 Spatial Reference Information, Spatial Extent, and Temporal Information

An S&A for spatial reference, domain and temporal range ascertains that the location and time periods represented by the data fit clearly into the organizational mission and under its legal mandate. Descriptions of a data set's reference frame include physical information in terms of horizontal and vertical datum, coordinates, latitude and longitude resolutions, and geographic or planar projections. The spatial reference information serves as a point of orientation for the data set's location and provides information about the physical measurements of the spatial framework of the data set.

The spatial domain of a data set defines the areal extent bounding the geography of the data. The spatial domain can be described in terms of "bounding boxes"; the corner coordinates of a rectangular or polygonal geographic area of the data set; or, by various descriptions of geographies of scale, such as states, countries or continents.

Temporal range is the time period when the data were collected. In certain cases, the time period refers only to the publication date of the data set, rather than the creation date. Reporting the temporal range for a data set can vary from the most detailed information, including dates in terms of measured periods (calendar, single date), or as measured time (time of day, first hour, minutes), to generalized descriptions (multiple years, range of dates, event) (FGDC, 2014b).

2.4.3.5 Current Scientific or Cultural Heritage Value

The current scientific value of data is based on the concept that data were used to communicate the results of research studies and are required to continue research, create new science, or augment current research in other disciplines.

The current cultural heritage value of data is the importance of any particular set of digital information as determined by the aggregate of values attributed to it. According to the report "Assessing the Values of Cultural Heritage" (Getty Conservation Institute, 2002), the values considered in this process should include those held by experts (historians, archaeologists, architects, and others) and those brought forth by new stakeholders or constituents. Data associated with ongoing or current events of social significance should be preserved for ongoing and future research (FGDC, 2014b).

2.4.3.6 Technology and Obsolescence Risks

To adequately preserve digital geospatial data, proactive steps must be taken to prevent or mitigate the effects of technology obsolescence. Determining when to put these steps into action can be difficult, but a definition from the "Reference Model for an Open Archival Information System (OAIS) Magenta Book" (Consultative Committee for Space Data Systems, 2012) suggests that "long-term" stewardship may best be affected by dividing actions into shorter, punctuated durations with more regular monitoring. The report defines "long-term" as: A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future (Consultative Committee for Space Data Systems, 2012).

This definition does not provide a specific time period for when media should be refreshed, but suggests continual monitoring by data stewards of advances in computer hardware, software, firmware, and storage media. While upgrade cycles vary widely depending on the materials under consideration, geospatial data stewards should evaluate their technology refreshment lifecycle at least once during each five-year period. Software migration is often tied to operating system evolutions. When software is migrated, a key consideration is whether or not the new software can read media containing geospatial data created under previous versions.

The concept of "format sustainability" becomes important when addressing technology and obsolescence risks. Formats that are sustainable remain comprehensible and accessible throughout their lifecycle as technology evolves. The "Sustainability of Digital Formats Planning for Library of Congress Collections" (U.S. Library of Congress, 2014) site has addressed criteria for selecting formats based on their sustainability. These criteria, each of which is explored in greater detail on the site, include:

- Disclosure
- Adoption
- Transparency
- Self-documentation
- External dependencies
- Impact of patents
- Technical protection mechanisms

2.4.3.7 Cost-Benefit Analysis

Geospatial records considered for ingestion into long-term or permanent archives should be subject to a cost-benefit analysis as one component of an overall records S&A. When appraising existing collections, institutional policy determines the specific nature of data to be acquired and identifies any gaps in the collections that require filling. Ensuring that repositories have the right to reject data sets that fall outside their scope of collecting can help avoid acquiring data that may be too costly to maintain, both financially and in terms of staff resources.

In addition to the S&A of records resulting from processed data, data sets are candidates for long-term preservation if there is no realistic chance of repeating the experiment, or if the cost and intellectual effort required to collect and validate the data are so great that long-term retention is clearly justified. Funding streams for data-generating activities may wish to build-in adequate resources from the start to support end-to-end data management, including long-term stewardship if required, while understanding that the costs of capturing and storing data can and will fluctuate over time (FGDC, 2014b).

2.4.3.8 Tangible Media and Physical Condition

Tangible media, often called "physical media," is the generic name for external digital storage media, including 8, 5.25 and 3.5 inch "floppy" disks, CD-ROMs, digital video, blu-ray and other optical disks, memory cards, USB "flash" drives and external hard drives.

These devices may contain important digital files but should first be appraised in their physical form. These items present an elevated preservation risk in that the tangible media are, themselves, fragile; and that fragility endangers the digital materials housed on them.

Detailed guidance on managing digital materials stored on physical media in preparation for transfer can be found in the white paper "You've Got To Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media" (OCLC, 2012). Appraisal should include these steps:

- Count and describe all identified media. Retain the order (if one exists) of the original digital media and accompanying items.
- Count the number of each medium type, indicate the maximum capacity of each medium, calculate the total maximum amount of data stored in each medium, and then calculate the overall total for the collection. This will enable one to estimate storage needs keeping in mind that the media are rarely full, so the estimate will likely be far in excess of the actual storage needed.
- Detail the physical condition and overall quality of the tangible media.
- Record anything that is known about the hardware, operating systems, and software used to create the files. Leverage associated documentation if any exists.

Prioritize appraisal decisions for the tangible media collection by estimating the value, importance, and needs of the collection as a whole, the level of use, or anticipated use, of the collection and potential danger resulting from loss of content through degradation due to age or condition (FGDC, 2014b).

2.4.3.9 Metadata Availability, Quality, Completeness and Usability

Metadata are critical to the selection and appraisal (S&A) process. They comprise administrative, descriptive, preservation, rights management, structural, and technical information that provide context to data and help users comprehend and understand them. The availability of quality metadata addresses several

S&A elements already outlined in the guidance document: legal issues; spatial reference and temporal information; data provenance and lineage; media format; and many others.

The report "Utilizing Geospatial Metadata to Support Data Preservation Practices" (Geospatial Multistate Archive and Preservation Partnership, 2011a,b) describes the two primary geospatial metadata standards utilized by a majority of practitioners: the *Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata* (FGDC, 1998) and the ISO *Standard for Geographic Information Metadata* (ISO, 2003). The report includes a comprehensive checklist of important CSDGM fields that facilitate long-term geospatial preservation, though individual agencies will need to develop their own metrics on metadata completeness.

Legacy geospatial data often need remediation to upgrade metadata to a usable form. Beyond a particular dataset's original metadata, it is good practice to include additional information documenting the creation and transmission of the dataset. Items such as documentation libraries, guides, fact sheets, FAQs, instrument documentation, design reviews, lessons learned, hardware documentation, engineering models, computer models, platform documentation, algorithm documentation, URLs, and principle investigator contacts may be included, all generally falling under the category of "administrative" metadata (FGDC, 2014b).

2.4.3.10 Uniqueness

"Uniqueness" describes data that are the only or sole example, of their type. The "How to Appraise and Select Research Data for Curation" document defines "uniqueness" as: The extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere (Digital Curation Centre, 2010).

The Centre report poses these questions regarding S&A for uniqueness:

- Is the dataset the only source of its content and will it be preserved elsewhere?
- Does the dataset duplicate existing work?
- Do other copies of the data exist that are accessible and useable?
- If other copies exist, where is the most comprehensive or up-to-date version?

Are any other copies at risk of loss? If so, will they be preserved by their holding organization? (Digital Curation Centre, 2010)

2.4.3.11 Provenance

Provenance refers to an understanding of the context from which a set of geospatial data was created. Provenance captures where data came from, how they were derived, manipulated, and combined, and how they have been updated. Provenance information helps shed light on the original creation purpose of data and their potential use in future applications. They provide an understanding of the history of organizational data control over time and can provide significant assistance in determining the long-term management responsibility for any particular set of data. In appraising for provenance, stewards should examine the degree to which contextual information about the origin and ownership of the data in question is available (FGDC, 2014b).

2.4.3.12 Future Value Determination

One of the more challenging S&A points is a determination of the anticipated future benefits or secondary scientific or public policy value of geospatial data, especially the levels of service required to achieve these benefits. *The Procedure for Scientific Records Appraisal and Archive Approval: Guide for Data Managers* explores these types of evaluations and provides pointers to possible guidelines (NOAA, 2008).

The *Appraisal Policy of the National Archives and Records Administration* (NARA, 2006) is cited in the NOAA document and addresses the idea that data may have value to the agency, the Government, or to the public for unanticipated uses long after they have served their original purpose. What is of relatively low research use today may increase in value over time. In order to make an estimation on the issues and topics that will be considered of significance in the future, it is necessary to consider the kinds and extent of current research use and make inferences about anticipated use both by the public and by the Government.

The solution is a decision-making process that is iterative and ongoing, with data managers and stewards continually reviewing the data holdings under their purview to determine the appropriate level of service for each data set given legal and mission requirements, user needs, cost-effectiveness, and available resources.

Data managers should try to envision the needs of the future when making a decision regarding the long-term preservation of a dataset. It may be useful to research and document the current uses of the data in creating a rationale for preservation. However, this is only a part of the picture, and a sense of vision and imagination may be required to make the correct decision (FGDC, 2014b).

2.4.4 *What to Collect*

The issue of *what* to collect is distinct from guidance on *how* to collect. Here we pull back from the intricacies of assessment to a discussion about the value of digital geospatial data, especially the value of preserved geospatial data as part of a coordinated national collection. Of course, the perceived value of information depends on where one stands in government, academia or the private sector, but the concept of a "national collection" of digital geospatial data should transcend parochial interests and provide a useful framework for making collecting decisions that include representative samples of the most important data of our era. Current efforts make some of these value decisions less subjective.

The U.S. government is instituting greater oversight of its geospatial information by identifying key data as a *National Geospatial Data Asset*. It has begun implementing a management process for these assets, and it can be assumed safely that they are worthy of long-term stewardship and that the S&A processes described above be applied to them early in their lifecycle. The National Geospatial Data Assets are described more clearly in the National Geospatial Data Asset Management Plan (FGDC, 2014a). This plan supports the portfolio management guidance issued by the federal Office of Management and Budget under the Circular A-16 Supplemental Guidance (OMB, 2010).

The FGDC has identified 16 NGDA Themes (as of February 2013):

- Biota
- Cadaster
- Climate and Weather
- Cultural Resources

- Elevation
- Geodetic Control
- Geology
- Governmental Units, and Administrative and Statistical Boundaries
- Imagery
- Land Use-Land Cover
- Real Property
- Soils
- Transportation
- Utilities
- Water – Inland
- Water – Oceans and Coasts

State governments also collect digital geospatial data, often made centrally available through geospatial data portals such as that provided by the Utah Automated Geographic Reference Center. These data collections represent extremely valuable records of state, local and regional government activities and are ripe for the introduction of geospatial data stewardship models and S&A processes. State archival and library divisions are chiefly responsible for the collection and preservation of state agency digital geospatial data, but maturity levels vary widely from state-to-state, making a national collecting strategy for this information a priority.

There are also non-governmental and private collections that may be important as part of a comprehensive national collection of historic digital geospatial data. One key collection is *OpenStreetMap* (OSM), a public map comprised of open data created by thousands of individual contributors, including openly-licensed data from national mapping agencies and other sources licensed under the Open Data Commons Open Database License (*OpenStreetMap* Foundation, 2014). OSM represents an example of crowdsourced or "volunteered" geographic information similar to work being done with encyclopedias by Wikipedia.

While OSM essentially creates its own dataset from scratch, other private entities are aggregating publicly available open data. One example is ESRI's ArcGIS Open Data, part of the ArcGIS suite of tools and services. The open data portal provides a technology platform to aggregate data from other sources and provides benefits to creators such as a centralized distribution point (and the ability to leverage advanced ESRI technical infrastructure) using familiar workflows. The centralized Open Data portal may provide benefits for stewarding organizations by enabling the capture of data that might not otherwise come under long-term stewardship.

A third category of data for a national collection includes proprietary geospatial data from large mapping and technology companies such as Google, Digital Globe, Nokia, Microsoft, Amazon and others. There are intellectual property challenges and competitive advantage issues related to the ability to collect proprietary data, but there may be a willingness on the part of the creating companies to partner with archival entities to create escrow accounts to acquire data now but make them available at a future time.

The Digital Globe Foundation makes grants to individuals at military and civilian academic institutions who need access to their imagery provided they use them for non-commercial, research purposes only

(Digital Globe Foundation, 2014). There may be opportunities to leverage environments such as the Foundation to establish workflows for the capture of non-current private organization data.

2.4.5 *Conclusion*

The preservation of digital geospatial data depends on a number of factors. By identifying the value of data, developing appraisal and selection methods, developing a culture of stewardship across the lifecycle and by identifying partners in government, academia and the private sector, the preservation of a national collection of digital geospatial data can become a reality.

2.5 *Archiving Example: LP DAAC Ingest of ASTER Data from the Japanese Ground Data System*

The Japan Ground Data System (GDS) will stage the Advanced Space-borne Thermal Emission and Reflection Radiometer (ASTER) Level 1A data and the associated product delivery record (PDR), as they appear on the GDS DTF-2 archive tape, to the GDS L1 Online Server. GDS creates a product delivery notification (PDN) file that states the location of the staged data and PDR files (Figure 6-12).

The Land Processes Distributed Active Archive Center (LP DAAC) ASTER Data Transfer System (ADTS) validates the PDN file and transfers the PDR and L1 data files from the L1 Online Server. If an error occurs during data transfer, the transfer will be retried a configurable number of times to complete a successful transmission. If the ADTS cannot read the PDR, the LP DAAC will not attempt to ingest the data files and a PDR Discrepancy (PDRD) message will be generated and sent to GDS.

Once the PDR is read by the LP DAAC, and if no PDRD is generated, the ADTS will build a PDR and copy it to a location for ingest into the EOSDIS Core System (ECS) and the Long Term Archive (LTA). A Product Acceptance Notification (PAN) message will be generated when all Level 1A files, excluding browse, associated with the PDR are ingested into the LTA.

The ECS Polling Service pulls the PDRs from the polling directory and performs validation on the PDR. The ECS Processing Service copies the Level 1A data to a hidden directory into the Data Pool PAN for all valid PDRs. The Processing Service then performs checksum verification and translation of ODL files to XML files as well as validating the metadata using the XML Validation Utility. The validated science XML metadata are copied to a location in the StorNext Archive. Toolkit is used to determine the Day/Night Flag for the Level 1A data and is updated in the Inventory database. The Data Pool Ingest Utility (DPIU) registers the Level 1A granule in the Inventory database. The DPIU then copies the Level 1A granules to the StorNext Archive and the Level 1A browse files are published to the ECS Data Pool.

A PAN message is generated and sent to ASTER GDS when all Level 1A files associated with the PDR are ingested into the LP DAAC. If there are problems in the data archive with any of the files associated with a PDR, a failed PAN message will be generated and sent to ASTER GDS. If all the files are successfully archived or none of the files are successfully archived, a short PAN will be generated. If only some of the files failed to archive, a long PAN will be generated that lists all of the files and their status.

When GDS receives a successful PAN, GDS will delete the data and PDR files associated with the PAN from the L1 -online server. The LP DAAC will not perform any modification to, or deletion of, the data placed on the L1 online server. If data need to be resent for any reason, GDS operators restage the data and create a new PDR. When the data are ready to be transferred to the LP DAAC, GDS will stage a new PDN.

ASTER GDS retains all products staged to the L1 online server until a successful PAN is received from the LP DAAC. The duration of the retention period is determined by the L1 online server resources allocated to the LP DAAC.

In the event that products staged for ingest by the LP DAAC are lost due to system anomalies at the LP DAAC or ASTER GDS, GDS will restage or reprocess the data and generate new PDR and PDN files.

The LP DAAC will track all ingest and archive failures and work with the ASTER GDS operations team to reconcile any persistent or recurring problems. Likewise, ASTER GDS will track all PDRD and PAN failures and work with the LP DAAC operations team to reconcile any persistent or recurring problem.

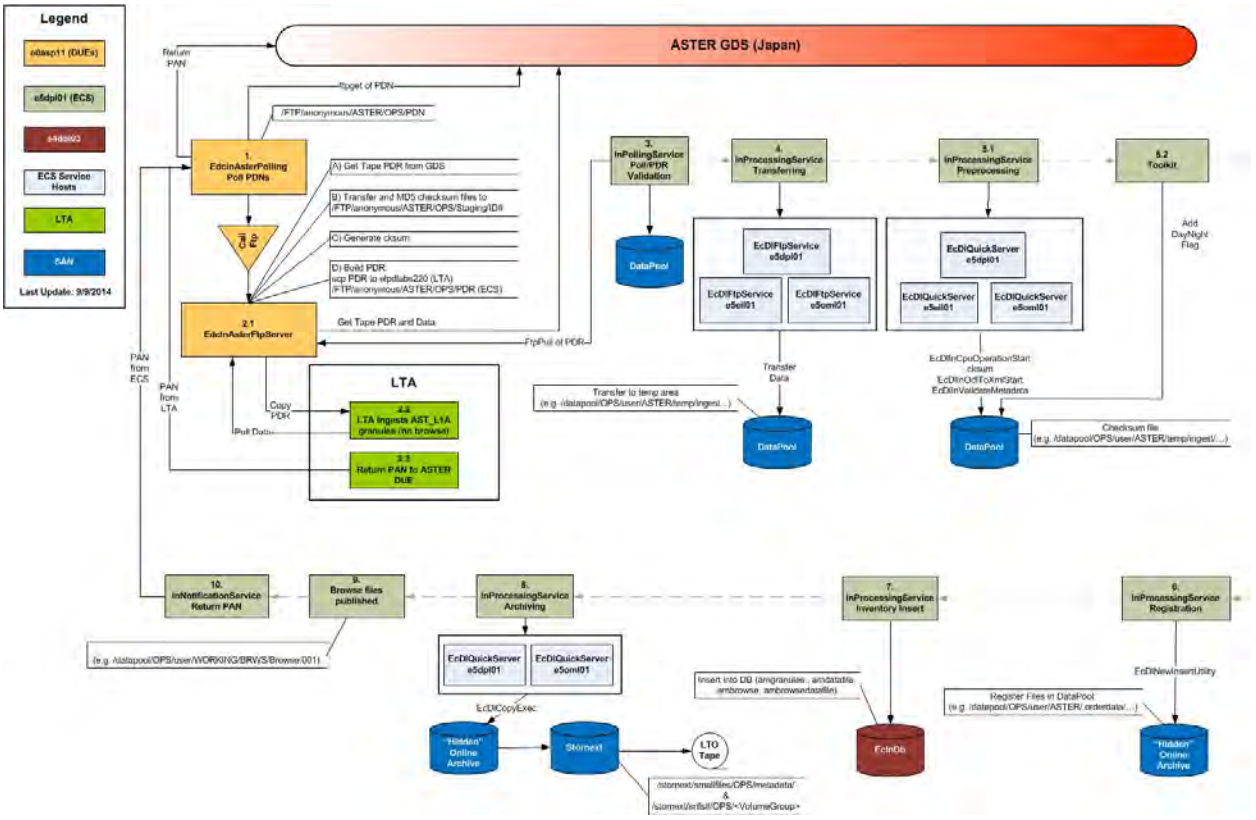


Figure 6-12. ASTER-ingest process.

2.6 *Disposal*

There are many misconceptions about what the word *Disposal* means or implies (Pearce-Moses, 2015). Simply put, the act of disposal relates to a record or a collection of records leaving your organization. These include transfers to another entity or possibly destruction. A good practice is to consider destruction as the last option for disposal. Seeking another organization, whose mission better aligns to the records and who is willing to take on the preservation and access activities, is preferred. There can be many reasons why an organization is seeking to dispose of records. Storage space costs real money, especially if the area is conditioned. Perhaps the organizational mission has evolved to the point where the records no longer fit well with the current vision of the organization.

The process of appraisal can aid in determining which records can be considered disposable. Once the decision to dispose of records has been made, the next step may be to seek a new home for the records. The Committee on Earth Observation Satellites (CEOS) has maintained a simple list server that announces intentions of a party to dispose of records. This list server is called *Purge Alert*. The simple list server allows the holder of the records to describe the collection such that other parties can determine if they have interest in pursuing acquiring the offered data.

CEOS established a Purge Alert service to help to ensure the long-term preservation of valuable Earth observation data. CEOS encourages use of this service to enable data archive managers to: 1) advise other archives of Earth observation data holdings scheduled to be destroyed, and 2) offer these data to other archive centers.

2.7 *Storage*

Storage considerations are part of preservation activities organizations pursue to protect investments made in remote sensing and geospatial records. Those records can be digital, analog or a combination, especially when ancillary analog documents are also considered part of a dataset. It is important to understand the proper storage guidelines to increase the longevity of the records being preserved. The U.S. National Archives and Records Administration (NARA) provide a comprehensive set of facility guidelines applicable to the storage of remotely sensed or geospatial records. The NARA guideline entitled, *Archival Storage Standards*, provides specific instruction on topics as varied as fire doors and floor load limits to air handling systems.

The environmental conditioning of storage areas is critical to manage and preserve analog and digital remote sensing and geospatial records properly. Table 6-3 provides recommended temperature and relative humidity ranges for various materials that records are stored on. The U.S. Geological Survey, working with the NARA, created the recommendations.

Table 6-3. Environmental Guidelines for Storing and Preserving USGS Science Records Media.

Records Media	Temperature Range	Relative Humidity
Paper – including files, maps, charts, drawings, posters	50º - 65ºF	30% - 50%
Magnetic/Electronic Media – computer tapes, disks, video tapes, audio tapes, optical disks	50º - 65ºF	30% - 40%
B/W Photographic Media (non-acetate/non-nitrate) – motion and still picture negatives, film, paper prints, x-rays, and microforms.	50º - 65ºF	30% - 40%
B/W Photographic Media (acetate) – motion and still picture negatives, film, x-rays, microforms, diazo, vesicular microfilm	0º - 35ºF	30% - 40%
Color Photographic Media – motion and still picture negatives, film, slides, prints, digitally produced prints (from inkjet, dye sublimation, electro-photographic, thermal)	0º - 35ºF	30% - 40%
Paper – Optimum preservation stacks primarily used in libraries[4]	35º - 65ºF	30% - 50% (±3%)

2.8 *Evolution of Archival Storage (from tape to memory)*

Over the last three decades, there has been a significant evolution in storage technologies supporting archival of remote sensing data. This section provides a brief survey of how these technologies have evolved. Three main technologies are considered: tape, hard disk, and solid state disk. Their historical evolution is traced, summarizing how reductions in cost have helped being able to store larger volumes of data on faster media. The cost per GB of media is only one of the considerations in determining the best approach to archival storage. Active archives generally require faster response to user requests for data than permanent archives. The archive costs have to consider facilities and other capital costs, operations costs, software licenses, utilities costs, etc. For meeting requirements in any organization, typically a mixture of technologies is needed.

2.8.1 *Active Archives*

Active archives are defined as facilities that store data that are in active use by the community. Typically active archives ingest raw data and/or derived digital products (hereafter simply referred to as "data") from active remote sensing missions, even though serving a user community with the stored data may proceed well beyond the life of active missions. Given the active user community, it is necessary for active archives to be responsive to the community requirements. Typically, relatively fast access is needed. Data need to be backed up and restored promptly while the system continues to operate in a responsive manner. Support staff is needed to assist data providers in setting up mechanisms for product generation and delivery to the archive, as well as any problems that arise during production and ingest operations. Also, user services are needed to help consumers of data with answers to questions they may have about either the mechanics of obtaining the data or scientific questions about the data themselves. When the missions are active, expert consultation is generally available to user services staff from scientists associated with the mission. It is also the active archive's responsibility to prepare for permanent archive of data at the end of the active archive phase, whether the data are transferred to another organization or continue to be held at the same organization.

From a hardware standpoint, an active archive might use a tiered storage mechanism to optimize performance. Tiered storage provides access to data across a virtualized storage system (http://en.wikipedia.org/wiki/Active_Archive). The data migrate between several systems that use different types of media for storage. Data that are accessed more frequently and need to be provided fast would reside on more expensive media and storage systems, while other types of data would be on less expensive hardware. The migration among the systems is handled automatically. Metadata keep track of where the data are. The data are available on primary, secondary and tertiary systems, providing on-line or near-line accessibility.

2.8.2 *Permanent Archives*

Permanent archives store data "forever", long after the data cease to be in active use. Quick access to data may not be an essential requirement. However, it should be possible to obtain the data when needed; for example, for retrospective studies that might occur, say, 30 years after the active usage ended. The level of service to users may not be as high as it is in active archives. Experts directly involved in the missions

would no longer be available for consultation. Thus one needs to depend on archived documentation, which must be complete to enable a diligent user to comprehend how the data had been generated.

Thus, from a hardware standpoint, a permanent archive may use less expensive and less responsive storage systems than an active archive. However, in both the active and permanent archives, preservation with no loss is equally important. Preservation requires:

- No loss of bits
- Discoverability and accessibility
- Readability
- Understandability
- Usability
- Reproducibility of results

From the point-of-view of hardware, the first and third bullets above are significant. The remaining bullets are also very important for preservation, but the actions to be taken to enable them are not within the scope of hardware solutions. Migration to newer media and reader technologies is essential to ensure no loss of bits over time, and readability of data. Changes in technology over time provide less expensive and faster storage with greater capacity, enabling us to archive ever-growing volumes of data. However, they also require frequent (perhaps continuous) migrations of data to newer media.

2.8.3 *Storage Technologies*

There are several recent publications tracking the evolution of storage technologies and the reductions in costs per unit of archival storage over the last three decades. An interesting history of computer devices, including storage, can be found in Computer History Museum (2015). Magnetic tapes, Hard Disk Drives and Solid State Disks/Drives are the major technologies that have been used for bulk storage. These will be discussed briefly below. It is to be noted that the names of companies and products given below only are meant to be illustrative of the technologies and capacities achieved as a part of storage technology evolution. Clearly it is beyond the scope of this section to cover all the storage products that have been made available in industry. The interested reader is encouraged to pursue the references provided for more details.

2.8.3.1 Magnetic Tapes

Magnetic tape storage technology, first patented by German engineer, Fritz Pfleumer, in 1928 (Zetta, Inc. 2015), has been evolving and is still in use for bulk storage applications due to its low cost, portability, and unlimited off-line capacity. Magnetic tapes were used for audio recording in the 1930s and were first used for data storage by UNIVAC in 1951. In circa 1970, IBM introduced the 10.5 inch standard tape reels. This standard lasted for over 25 years with various lengths (1,200 feet, 2,400 feet and 3,600 feet), numbers of tracks (7 and 9), and recording densities (ranging from 200 characters per inch to 6,250 characters per inch). Digital Equipment Corporation's (DEC) CompacTape Cartridge replaced the 1960s tape technology and was later standardized as Digital Linear Tape (DLT). The DLT technology evolved from 92MB capacities in 1984 to 800GB capacities in 2006 (super DLT Format). Cartridges and cassettes, consisting of tape reels that are completely enclosed in a plastic casing have come into common use since audio compact cassettes were used in home computers as inexpensive storage in the 1970s and 1980s. As of 2014, various

cartridge formats were in use for Digital Data Storage (DDS), a format for storing computer data on either Digital Audio Tape (DAT), Digital Linear Tape (DLT), or Linear Tape – Open (LTO). Steady increases in cartridge capacities are exemplified by the evolution of generations of LTO. LTO-1, in the year 2000, had a capacity of 100GB, while LTO-6 in 2012 had a capacity of 2.5TB. It is anticipated that LTO-10 will have a capacity of 48TB.

Magnetic tapes can be stored off-line or in "near-line" tape libraries. With off-line storage, a human operator needs to mount a tape on a tape drive in order to read or write data. Near-line tape libraries include a robotic device that is controlled to access and mount the tape of interest on a tape drive to permit reading and writing. The IBM 3850 mass storage system, announced in 1974, was one of the earliest examples, consisting of a number of cylindrical cartridges held in a hexagonal array of bins. Data were transferred automatically between higher-speed disk drives (on-line storage) and the cartridges. The capacity of the mass storage systems ranged from 35.3GB to 472GB, depending on the model. This series was discontinued in 1986. Since then, use of near-line libraries has become common. In late 1990s through mid-2000s, the NASA Earth Observing System (EOS) Data and Information System (EOSDIS) used robotic tape silos for near-line storage of most of the EOS data (several petabytes) and derived digital products in its Distributed Active Archive Centers (DAACs). Access to data from near-line storage can be significantly slower than that from on-line spinning disks. Today, the DAACs use on-line spinning disks for most of the archive storage while using near-line capacity for back-up. Of course, the access with near-line robotic tape silos would typically be much faster and less subject to human errors than from off-line storage requiring tape mounts by operators. As of 2014, there are near-line mass storage systems with multi-exabyte capacities.

2.8.3.2 Hard Disk Drives (HDD)

A history of the evolution of hard disk storage and a detailed time line from 1956 to 2014 can be seen in Wikipedia Contributors (2014). The following is a short summary of highlights from that article. The first commercial hard drives were introduced by IBM in 1956 with a capacity of five million 6-bit characters. In 1965, the IBM 2341 was introduced with removable disk packs of 11 disks for a total capacity of 29MB. In 1975, the IBM 3350 "Madrid" was brought into market, re-introducing disk drives with fixed disks. The capacity of Madrid was 317.5MB per drive, for a capacity of over two Gigabytes per string consisting of eight 14" disks. In 1980, Seagate Technology (then Shugart Technology) introduced the ST-506, the first 5.25 inch hard disk drive, which had a capacity of 5MB. Also in 1980, the IBM 3380 came on the market. It was the world's first gigabyte-capacity disk drive (2.52GB), was the size of a refrigerator, and weighed 249 Kg. In 1988, PrairieTek 220 was introduced as the first 2.5 inch hard drive, which had a capacity of 20MB and suitable for portable computers. In 1997, IBM introduced the Deskstar 16GP "Titan" with five 3.5 inch disks with a capacity of 16.8GB. This was significant in that it was the first commercial use of Giant Magnetoresistance heads. Also in 1997, Seagate brought into market the Medalist Pro 9140 (ST39140A) with a 9.1GB capacity, the first hard drive with fluid bearings. There were several key developments in 2005. The first 500GB hard drive was shipped by Hitachi GST (HGST), Serial ATA three Gb/sec was standardized, Seagate introduced Tunnel Magneto-Resistive Read Sensor and Thermal Spacing Control, faster Serial Attached SCSI was introduced, and Toshiba shipped the first perpendicular magnetic recording hard disk drive (1.8 inch, 40/80GB). The years 2007 through 2011 saw a few firsts in the

capacities of hard disks, starting from one TB (2007, HGST) to four TB (2011, Seagate). In 2013, HGST announced a helium-filled 6TB hard disk drive for enterprise applications. In 2014, Seagate shipped the first 8TB hard drives.

The costs of hard disk drives from 1980 to present are summarized by Komorowski (2009 and 2014). Figure 6-13 is adopted from his articles. Note the logarithmic scale in the figure. It shows that the cost per gigabyte has dropped from $700K in 1981 to between $0.03 and $0.06 in 2014. Komorowski shows a regression model indicating doubling of storage capacity per unit cost every 14 months. Other examples of such cost trends are compiled by Smith (2014) and McCallum (2014).

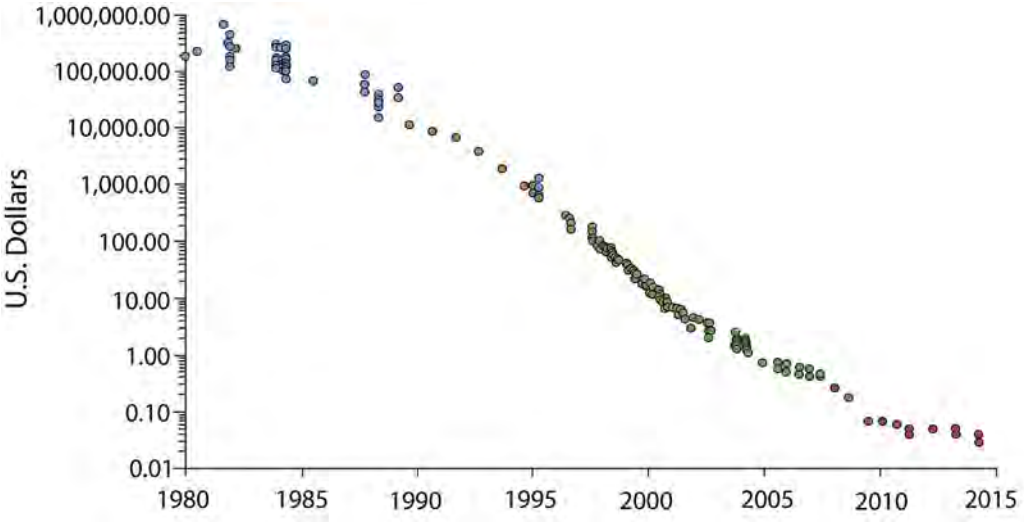


Figure 6-13. Hard drive costs per gigabyte – 1980 to 2014 (Credit: Komorowski, 2014).

2.8.3.3 Solid-State Drives or Solid-State Disks (SSD)

Devices called Solid-State Drives or Solid-State Disks (SSD) are neither drives nor disks. They are storage devices like HDDs, but use integrated circuit assembles for persistent storage of data. They also use electronic interfaces that are compatible with Hard Disk Drives (HDDs), but provide significantly higher input/output performance. SSDs differ from HDDs, floppy disks or tape drives in that they do not have any moving components and thus are resistant to physical shocks and run silently. Most SSDs use NAND-based flash memory, which can retain data without constant need for electric power. For faster access, Random Access Memory (RAM) SSDs can be used, but they require some source of electric power to retain data (Wikipedia Contributors, 2015a,b).

The following is a brief summary of SSD highlights (Kerekes, 2015). In 1976, Dataram brought into market an SSD called BULK CORE. It emulated hard disks and had a capacity of 2 MB. In 1978, one GB of RAM SSD would have cost $1M. Texas Memory Systems introduced a 16 KB RAM SSD for accelerating field seismic data acquisition for oil companies. In 1982, Semi-Disk Systems shipped SSD accelerators for the Personal Computer market, initially with a capacity of 512KB and later with a capacity of two MB. In 1990, NEC introduced 5.25 inch SCSI SSDs that used RAM technology and backed up with internal batteries. By 1996, with ATTO Technology's introduction of SiliconDisk II, the RAM SSD capacities had gone up to 1.6GB with a throughput of 80MB/s and 22,000 input/output operations per second (IOPS). In 1999, BiTMICRO introduced a flash SSD with a capacity of 18GB. By the end of 1999, there were at least

11 manufacturers of SSDs. In November 2000, BiTMICRO launched the first hot-swappable 3.5 inch SCSI SSD. In 2001, Winchester Systems introduced FlashSSD as an option in its OpenRAID Storage Area Network products for use on a small percentage of "hot files" that account for a majority of disk access requests. FlashSSD provided consistent performance of 12K IOPS and 40 MB/s throughput. Also in 2001, Texas Memory Systems began promoting its RamSan-210, a RAM SSD with 32GB capacity, 100K IOPS and $20\mu s$ access times. In 2003, SSDs with a capacity of one TB became commercially available. In 2005, M-Systems announced that the industry's highest capacity 2.5 inch Serial Advance Technology Attachment (SATA) SSD with 128GB storage capacity was available. Also, Texas Memory Systems launched SSDs with a 4Gb/s Fiber Channel interface offering up to 128-gigabytes capacity and 500,000 random IOPS performance. In 2006, 1.8 inch 32GB flash SSDs from Samsung hit the market. In 2008, the number of Original Equipment Manufacturers (OEMs) of SSDs reached 100. In 2010, Texas Memory Systems announced the availability of the RamSan-630 SSD with four- to ten-TB capacity, 500,000 IOPS, and 8GB/s bandwidth. Foremay announced its 2-TB 3.5 inch and 1-TB 2.5inch SATA flash SSDs with read/write speeds of up to 200MB/s. Fusion-io set speed records of achieving one million IOPS and 6.2GB/s bandwidth, and offered capacity up to 5.7-TB. In late 2011, BiTMICRO announced a new generation of enterprise SSD controllers that could deliver up to 400K IOPS and a capacity of 5-TB for availability in the first half of 2012. In 2012, HGST demonstrated the first 12GB/s Serial Attached SCSI SSD in industry. In 2013, Micron announced a new model of hot swappable 2.5 inch Peripheral Component Interconnect Express (PCIe) SSDs with up to 1.4-TB multi-level cell (MLC) capacity, which could deliver 750K IOPS. Samsung entered into the 2.5 inch PCIe SSD market with its NVMe SSD which had up to 1.6-TB capacity, read throughput of three GB/s and up to 740-K IOPS. In 2014, Samsung provided a comparison of speeds of 2.5 inch SSDs using SAS and PCIe technologies and showed that its PCIe SSDs were three times faster than the SAS SSDs. SanDisk started sampling 2.5 inch four-B SAS SSDs. Skyera launched its 136-TB, 1U (i.e., 1.75 inch high) rack-mounted SSD called the SkyHawk FS.

A comparison of average prices per gigabyte of HDD and SSD over the period 1996 through 2012 is shown in Figure 6-14 (Royal Pingdom.com, 2011). The cost of hard disks and drives has dropped significantly over the past three decades, making their use feasible for petabyte scale data archives. Where high throughput performance is a requirement, Solid State Disks (SSDs) have been used in recent years. SSDs are significantly more expensive than HDD's as shown in Figure 6-14. However, the cost differential between HDD and SSD has been dropping significantly over recent years. Cost of SSD per GB was 120 times that of HDD in 2007, while in 2011 the same ratio was 32. In 2014, this ratio had dropped to about 25. Vendor advertisements in January 2015 showed a ratio of 8 to 10. It is difficult to predict whether the two costs will become comparable in the future. See Baxter (2014) for a comparison of HDD and SSD and a discussion of pros and cons.

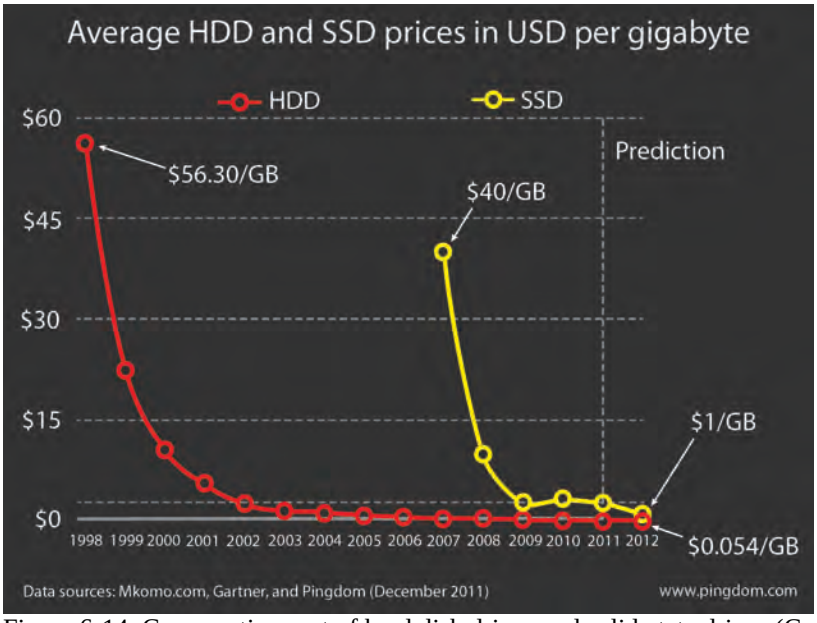


Figure 6-14. Comparative cost of hard disk drives and solid-state drives (Credit: Royal Pingdom.com).

2.8.3.4 Case Study - NASA's EOSDIS

NASA's Earth Observing System Data and Information System (EOSDIS) is a large system with 12 Distributed Active Archive Centers across the United States. EOSDIS manages most of NASA's Earth science data from satellite missions, aircraft investigations, field campaigns and other sources. At the end of 2014, EOSDIS archived approximately 10PB of data. The EOSDIS Core System (ECS) provides common "core" hardware and software capabilities to three DAACs – the Atmospheric Science Data Center at NASA Langley Research Center, Hampton, Virginia; the Land Processes DAAC at USGS EROS in Sioux Falls, South Dakota; and the National Snow and Ice Data Center at the University of Colorado in Boulder, Colorado. For purposes of illustration, the storage technologies employed in the ECS archives are discussed below. The ECS has been in operation since late 1999 and has been supporting archiving and distribution of the EOS satellite missions. It has evolved from near-line tape-based robotic archives to on-line disk-based archives. Behnke *et al*. (2005) describe the technology changes since the beginning of the ECS design in 1995 through 2005. Initially, all of the data were stored in robotic tape silos, with on-line disk storage being used to cache the data. As the cost of disk storage decreased, it became feasible to provide some of the data on-line. The concept of "data pools" was introduced in 2001 (Moore and Lowe, 2002). Data Pools were large (tens of TB) caches of popular datasets that could be directly downloaded by users, thus reducing the latency in meeting user requests. With further reductions in disk costs, most of the data are now held on-line. This also helps in providing other on-line services to users such as sub-setting, re-projection, visualization, etc. upon request. Regarding the utilization of disk and tape technologies for back-ups of archives at all the EOSDIS DAACs, the following observations can be made. There is an equal mix of tape and disk based on product count. Disk is a popular medium for smaller volume products, and tape for larger volume products. Transition from tape to disk based backup has been driven by reduced disk costs; improved restore input/output speeds from disk; and lower error rates in stored disks. A small number of products are backed up on CD/DVD/Blu-Ray (optical media). The DAACs have automated systems to manage ingest, archiving

and back-up of data. In particular, the back-up system used at the three DAACs mentioned above that have the EOSDIS Core System is a tiered storage management system using StorNext. This provides seamless access to data held on disk and tape media. The data are ingested on to a set of archive disks. They are then copied from the archive disk to tape for a complete back-up. Copies to tape are determined by a set of configurable policies and generally occur after a set period of time, a data volume threshold is reached for specified datasets, or when an archive disk reaches a capacity threshold.

2.8.4 *Conclusion*

This section provides a discussion of three main technologies for archival storage and traces their historical evolution, summarizing how reductions in cost have enabled storing larger volumes of data on faster media. The cost per GB of media is only one of the considerations in determining the best approach to archival storage. Active archives generally require faster response to user requests for data than permanent archives. Archiving costs have to consider facilities and other capital costs, operations costs, software licenses, utilities costs, and contingencies. An example of such an analysis by the San Diego Supercomputer Center is given in Moore *et al*. (2014) They demonstrate that the annual operating cost per TB of storage at their facility is a factor of three less for tape than for disk storage. However, for meeting requirements in any organization, typically a mix of technologies is needed. There has been a very significant change over the past 30 years in the capabilities that active archives can provide for their users. One could not have imagined 30 years ago that scientists using remote sensing data would today have most of the data available to them on-line and be able to "work from anywhere". Several technological advances have contributed to this change, including evolution of archival storage discussed in this section, as well as inexpensive storage available for users' laptop or desktop computers and faster performance of networks. While it is difficult to predict the technological environment of 30 years into the future, it can be expected that SSD's will become sufficiently inexpensive to support major archival operations and help with much faster access to data from users around the world.

2.8.5 *Acknowledgement*

Section 2.8 was supported by National Aeronautics and Space Administration (NASA) under contract number NNG12HP08C with Science Systems and Applications, Inc. H.K. Ramapriyan gratefully acknowledges encouragement, support and review comments by Ms. Jeanne Behnke of the Earth Science Data and Information System Project at NASA's Goddard Space Flight Center, Greenbelt, MD. He also thanks Mr. Chris Doescher of the United States Geological Survey's USGS-Earth Resources Observation and Science (EROS) Center for discussions on storage systems at the Land Processes Distributed Active Archive Center and back-up approaches.

2.9 *Stability and Life Expectancy of Magnetic and Optical Media for Data Archiving*

Advancements in digital storage devices, and innovation in recording formats, have increased significantly the speed and reliability of processing, accessing, storing and archiving several hundred Terabytes of digital data. A primary driver for these advancements resulted from improvements in magnetic and optical media composition of the digital recording medium. Several factors can affect the reliability of data

stored on magnetic and optical media; environmental stress factors like temperature and relative humidity can change the life expectancy of the media significantly. Laboratory methods use "accelerated aging" techniques to simulate changes in composition under elevated temperatures, and time dependent models are used to analyze the data to calculate the end-of-life of media. The values calculated serve as average trends rather than absolute end-of-life predictions.

Over a thousand times increase (from Gigabytes to terabytes) in storage density of digital has occurred since the 1990s. Increase in capacity has been made possible by using newer materials and associated technologies. For example, the chromium dioxide recording material in IBM 3480 tapes was replaced by metal particulate (MP) materials used in DLTIV and several later-generation, higher density tapes. More recently Barium Ferrite tapes have single tape storage capacity in the range of 10-TB.

2.9.1 *Magnetic Tape*

The primary mode of data storage and retrieval from magnetic media is by utilizing the ferromagnetic properties of the substrate layer contained within a polymer binder, supported by a back coating layer. These components - the magnetic substrate, binder and the backing layers can be potential source of failure for media. Magnetic substrates (γ-iron oxide, metal particulate (MP), chromium dioxide, cobalt-nickel metal evaporate) show decrease in magnetic remanence (MR) over time. At higher temperature/humidity conditions, binder (polyester polyurethane) hydrolysis reactions can result in "sticky tape" phenomenon, characterized by low tape coat modulus, high friction, and/or gummy tape residues (van Bogart, 1994). The MP substrate has been successfully commercialized, to increase the storage capacity of single tape cartridge (from a few hundred megabytes to two Terabytes). Under ambient temperature over a fourteen year period, about twelve percent decrease in MR was observed for the MP tapes (Hanai, 2002). At 60°C and 90 percent relative humidity, MP tapes with iron-Nickel composition showed 40 percent decrease in MR, whereas only 1-5 percent decrease for iron-cobalt tape substrate was observed (Hanai, 2013). The MR loss for Barium Ferrite (BaFe) tapes, compared to MP tapes is even less at higher temperature and humidity conditions (van Bogart, 1994, Weiss 2002). The stability of BaFe, coupled with increased areal density, has led to production of BaFe tapes that have storage capacity to 8.0 -TB for single tape cartridge (Peters, 2014).

2.9.2 *Optical Media*

CD, DVD's and Blu-Ray disks are comprised of a polycarbonate substrate, a data layer and a metal layer. The data layer can be molded (read only memory, ROM), organic dye (recordable (R) purposes), and phase changing metal alloy film rewriteable (RW). DVD's can be single sided or double sided. The storage capacity of DVD's that are double sided and double layered is quadruple that of a single side/double layered DVD. The organic dye is the most fragile of the components, and composition of the dye can significantly influence the longevity of the medium. Pthalocyanine, cyanine and Azo have been used as organic dyes in CD-R and DVD-R disks with gold, silver, or silver alloy as the reflective surfaces. Gold is noncorrosive, very stable and long lasting under extreme environmental conditions but has lower reflectivity compared to silver. The polycarbonate substrate is susceptible to moisture and airborne acids, oxidation of the metal reflective layer, UV light damage of disk surface, edge abrasion, de-bonding of the adhesive layer, all of which contribute to the reported "disk rot" phenomenon for optical media.

Therefore proper care and handling of CD's and DVD's is critical for archiving data on media (Byers et.al, 2004). Compared to magnetic tape, typical single DVD's storage capacity is 4.7 GB, Blu-Ray disks (BD) is 25-50GB, and BDXL 50-100GB. The BDXL optical library system can provide storage capacity of 50TB (Watanabe, 2013).

2.9.3 *Stability of Higher Density Magnetic Tapes*

A systematic study examining the magnetic properties and microstructure of the recording media, physical and recording characteristics, binder chemistry, and error correction/tape drive performance was conducted under the direction of the National Archives and Records Administration Media study (Weiss, 2002).

Several different stress conditions were induced by higher temperature and relative humidity (RH) conditions. Samples of each of the higher density magnetic tape types under the "accelerated aging" conditions were taken every 3-4 weeks for a period of 100 to 3500 hours. Every tape type showed a loss in magnetization under induced stress conditions of higher temperature and relative humidity.

Figure 6-15 illustrates the variability in life expectancy (LE) when temperature increases from 20 to 100°C. LE predictions for SDLT, DLTIV and IBM 3590 tapes are displayed at a constant relative humidity (RH) of 30 percent. At lower values in temperature, for example 20°C (68°F at 30% humidity), the predicted LEs for the tapes examined are in the range of 100+ years. At temperatures in the range of 50-60°C, the predicted LEs are as low as one year (Navale, 2005).

It is clear that relative decline in LE values is significant when the temperature increases from room temperature conditions to 50°C (122°F).

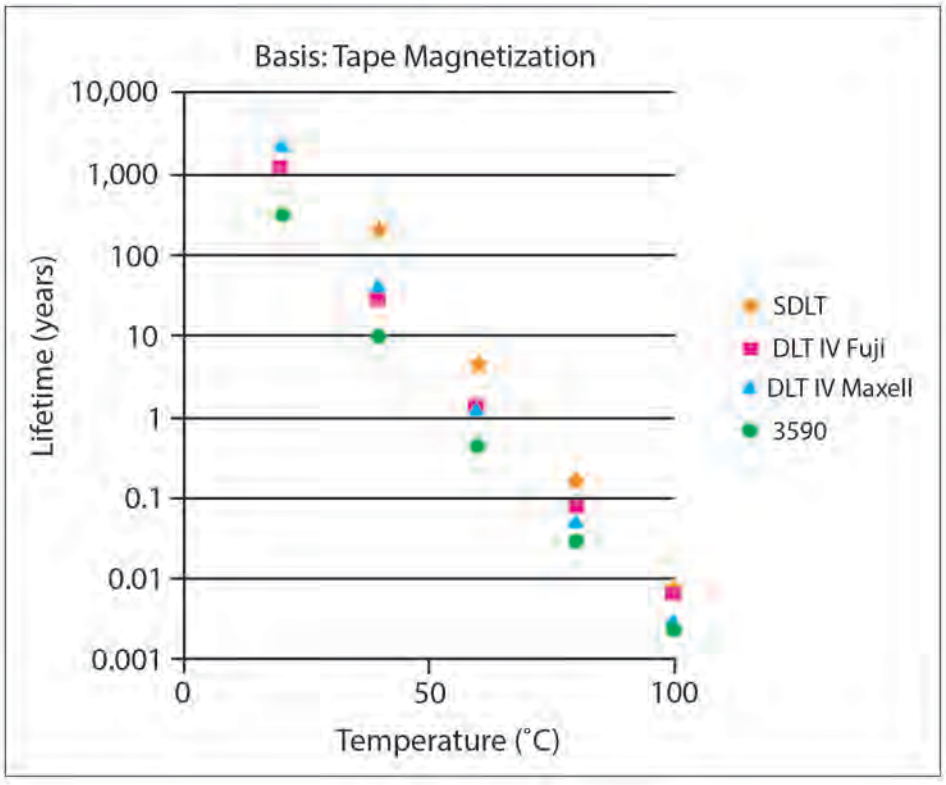


Figure 6-15. Life expectancy of high density magnetic tape vs temperature at 30% relative humidity.

A combined high stress condition of higher temperature and higher RH is illustrated in Figure 6-16. The overall trend for all the tapes examined under similar conditions, reveal a decrease in the LE values with an increase in RH conditions. The data show that LEs of the tapes at 50°C and at 50% RH are in the range of 1-3 years.

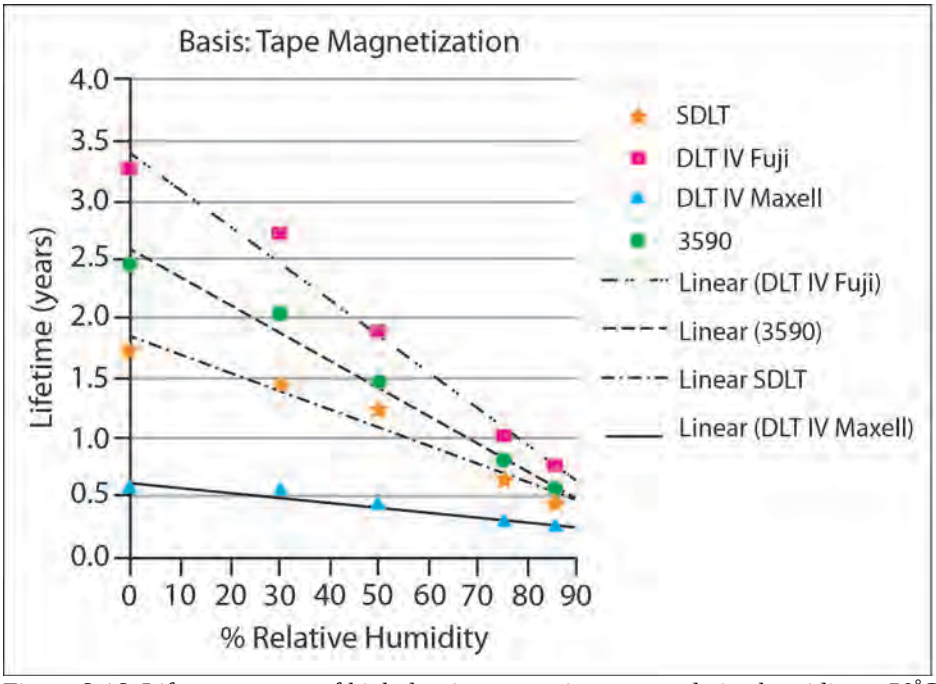


Figure 6-16. Life expectancy of high density magnetic tape vs relative humidity at 50°C.

2.9.4 *Stability of CD-ROMS*

One hundred and sixty prerecorded compact disks were randomly sampled from duplicates present in the collections of Library of Congress holdings. The CDs were subjected to environmental stress conditions (temperatures of 60, 70 and 80°C, relative humidity from 55-85%) over a time period of 500-1000 hours (Shahani, 2005).

The rate of deterioration of each specimen was determined by measuring block error rate (BLER) according to techniques specified in ANSI/NAPM IT9.21-1996. The estimated time to reach end-of-life for each disk subjected to a particular stress condition was studied and lognormal distribution pattern provided the best fit to the experimental data.

Figure 6-17 shows the probability of failure as a function of time at 25°C and 50% RH. It predicts a mean life time of 1592 years for CD-ROMs stored under those conditions.

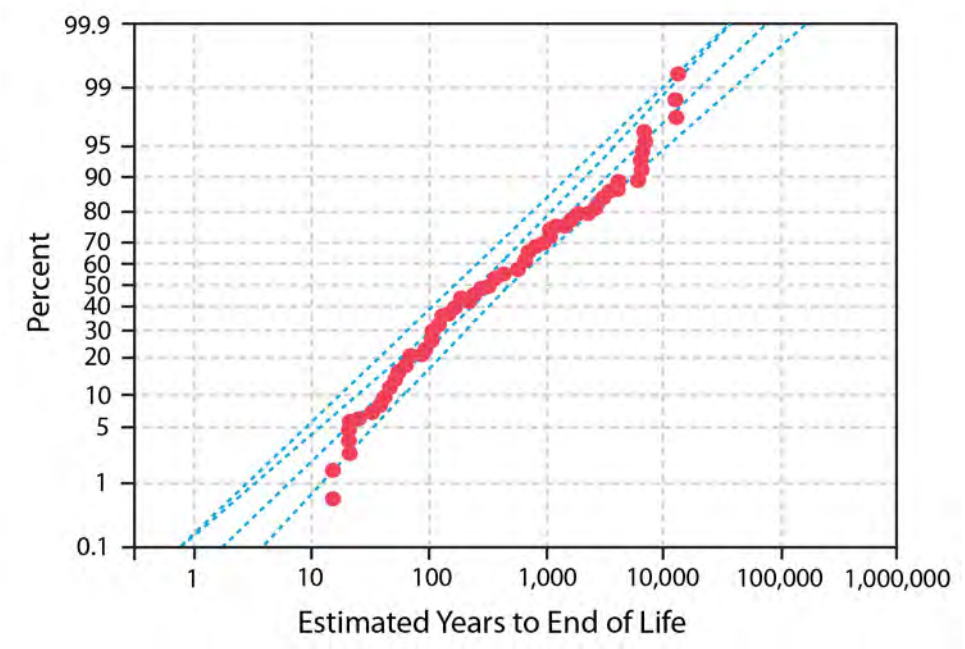


Figure 6-17.CD-ROM failure probability over time at 25˚C and 50% relative humidity (based on lognormal distribution pattern).

2.9.5 *Stability Study on CD-ROMs and DVD-ROMs*

Typically, CD-Rs are composed of five components (polycarbonate, the organic dye, the reflective layer, acrylic lacquer, and the label) that can affect the stability of the media. The organic dye is the most fragile of the components, and the composition of the dye can influence the longevity of the media significantly. The commonly used dyes in CD-Rs are phthalocyanine, cyanine, metal stabilized cyanine, and azo. During the National Institute of Standards and Technology (NIST) work, random samples of commercial CD-Rs that included the major dye types were analyzed.

Table 6-4 provides a list of coating and dyes for the various samples numbered S1-S7 that were used in the NIST study.

Table 6-4 List of Coating and Dyes for the Various Samples Numbered S1-S7 Used in the NIST Study.

Sample	Coating and Dye
S1	Unknown, Super Azo
S2	Unknown, Phthalocyanine
S3	Unknown, Super Azo
S4	Silver + Gold, Phthalocyanine
S5	Silver, Metal stabilized cyanine
S6	Silver, Phthalocyanine
S7	Silver, Phthalocyanine

Results on the accelerated aging tests conducted on commercially available recordable media (CD-Rs and DVD-Rs) during the study (Slattery, 2003).

Changes in Block Error Rates for CD-Rs and parity inner errors (PIE, referring to the number of rows in a data block that contain errors) for DVD-Rs were used as a measure of the quality of the specific

commercial media types analyzed during this work. The study also examined the effect of direct exposure to light, including some UV exposure. Samples were exposed to extreme conditions (temperature 60-90°C and 70-90% RH) with BLER and PIEs measured as a function of increasing exposure time (450-850 hours). BLER maximum has been stated at 220 errors per second in the CD-R specification (Watanabe, 2013).

2.9.6 *Effect of Light on CD-ROMs*

Tests demonstrated significant variation in BLER values over time when continuously exposed to direct light, as shown in Figure 6-18 (a,b). Phthalocyanine based samples S2 and S4 had the lowest error rates and remained below the end-of-life threshold for the entire testing period, whereas the azo dye based samples (S1 and S3) showed sharp increases in BLER values before 400 hours of light exposure was reached, the shortest of any of the samples. The two phthalocyanine samples using only silver in the reflective layer (S6 and S7) performed well in direct light exposure until approximately 600 hours, while the silver with metal stabilized cyanine (S5) showed relatively low BLER values until an exposure time of 800 hours.

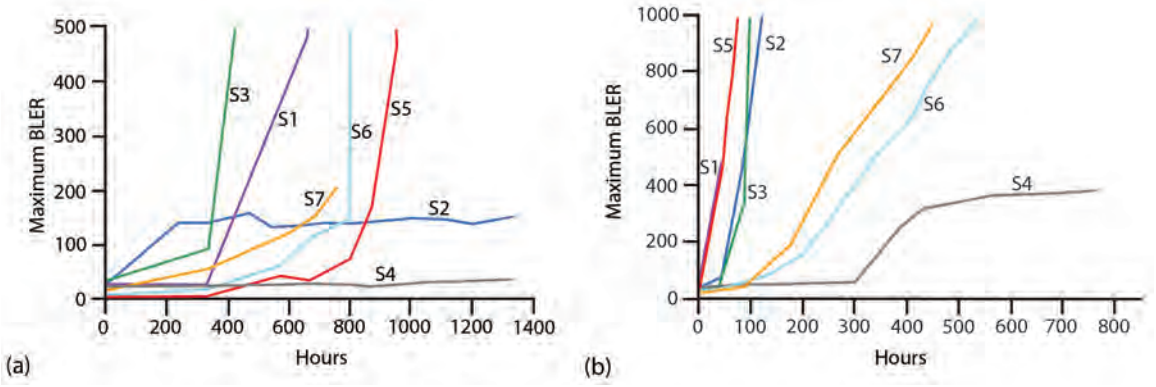


Figure 6-18. (a) BLER changes in CD-ROMs after continuous light exposure by using a metal halide lamp; (b) BLER changes in CD-ROM test samples after continuous exposure to higher temperature (90º C) and humidity (90% RH).

2.9.7 *High Temperature and Humidity Effects on CD-ROMs*

Considerable variability in BLER values was observed after exposure to high temperature and humidity conditions (90°C and 90% RH). Sample set S4 had the lowest BLER values in the extended light exposure test, and also had the lowest BLER values in this test. On the other hand, sample S2, which had the second lowest BLER value in the light exposure test was in a group of samples (also including sets S1, S3 and S5) that spiked high BLER values after less than 100 hours of exposure to high temperature and humidity. Samples S6 and S7 had identical dye and reflective layers with a common manufacturing source, evident by similar overall BLER trends—low values until 100 hours followed by an increase to 1000 by about 400 hours of exposure (Figure 6-19a,b).

2.9.8 *End-of-Life (EOL) of Media*

A primary goal for measuring a medium's end-of-life is to identify parameters that can be used for calculating media life expectancies (LE's). The standard method is to model the rate of change of an EOL variable as a function of temperature by using the "Arrhenius" equation and/or the "Eyring" equation (that includes temperature and relative humidity data), to predict LE values of media under various environmental conditions. The assumption these experiments make is that higher temperatures and relative humidity

(RH) over short time intervals (e.g. hours or days) can accelerate physical and chemical decomposition, mimicking changes that may occur at longer time intervals when media are stored at ambient environments. Normalizing accelerated aging LE values to ambient conditions (e.g. 25°C, 50% RH) can be used to determine the probability of media failure over time.

Changes in MR over time have been modeled, and LE predictions for environmental conditions, 20°C and 50% RH (for MP tapes manufactured during 1990'), ranged from 30-40 years (Bogart, 1994). At 40°C and 50% RH, LE predictions for the higher density (MP tapes manufactured after 2000) were in the range of 6-15 years (Weiss, 2002). These studies have clearly shown that LE's of MP tapes decrease significantly under higher temperature and RH conditions.

Magnetic tape drive error handling methods (cyclic redundancy checking, error correction code (ECC), and partial response maximum likelihood (PRML encoding) assist in correcting "soft errors" that can result from variability in media composition. Based on ECC application, "block error rate'" calculation provide a parameter that accounts for the system (device and medium) performance. LE's based on block error rate changes for tapes (DLT's, IBM 3590) at specified conditions (40°C, 50% RH) were reported to be higher, compared to LE's based on MR changes under similar conditions (Navale, 2005).

Comparison of LE predictions based on ECC corrected BLER versus predictions based on changes in magnetization, without use of ECC are shown below Figure 6-19a,b.

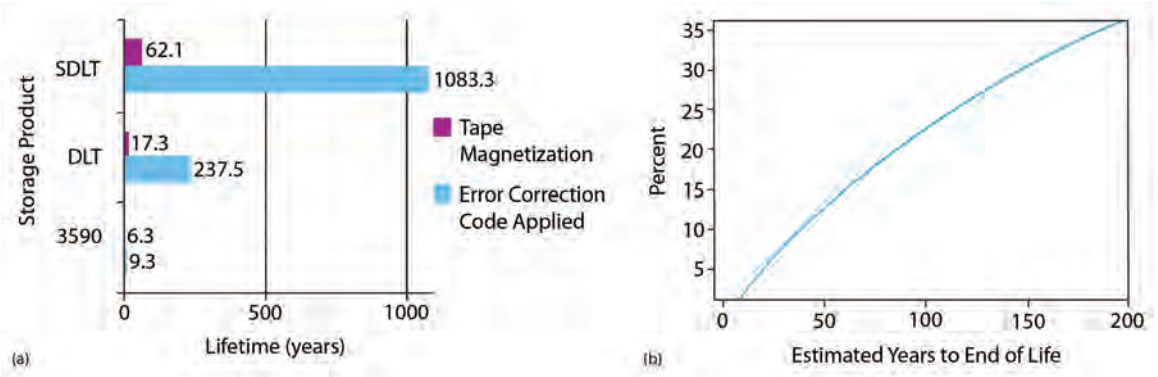


Figure 6-19. (a) Life expectancy @ 40°C/50% relative humidity (basis: tape magnetization vs. error correction code); (b) Cumulative failure rate for CD-ROMS @ 25°C and 50% relative humidity (based on lognormal distribution pattern).

Block error rate (BLER) changes have been measured also for CD-ROM's, and the mean EOL for CD's tested at 60°C and 85% RH was in the range of six years (Shahani, 2005). The probability for CD-ROMs to fail when maintained at 25°C and 50% RH during the first ten years is estimated to be less than 2% and may increase to 5% over a 20-year period. Figure 6-19b provides an indication of the expected percentage failure rate during the first 200 years.

It is important that life-expectancy predictions should be viewed more in terms of being able to predict average trends for different media types rather than as an absolute measure of end-of-life values. Recent performance tests with commercially available DVD drives showed that only 38% DVD-R's achieved the initial recording error rate standard of 140 or less (Inui, 2013). A life-expectancy prediction of 100 years for a specific medium will serve little purpose if an error condition violates the ECC capabilities of a system within a year.

2.9.9 *Summary and Conclusions*

Prior information on uncorrectable bit error rate (UBER) is an indicator for magnetic tape reliability (Peters, 2014). The UBER reported by vendors for enterprise tape media (BaFe) is in the range of 1x10E bits; for mid-range tape systems it is 1x10E, whereas for disk drives (HDD), it ranges from 1x10E bits for desktop SATA, 1x10E bits for enterprise SATA, to 1x 10E for enterprise class Fibre channel. Based on UBER, the enterprise tape media (BaFe) has three orders of magnitude higher raw reliability than enterprise HDD. Although "BLER" for magnetic tapes and CD's are measured differently, initial and periodic BLER change metrics over time can be useful for determining the reliability of archived media. For long term DVD-R storage, initial and periodic (at least once in three years) monitoring of parity inner error (PIE) should be carried out, and if PIE's are above 280, then data from DVD-R's should be migrated to newer media.

BaFe tapes have higher areal densities, higher signal to noise ratios, lower byte errors rates, and are suitable for long term data archiving. For archiving data on DVD-R's guidelines provided by international standards ISO/IEC 29121 should be followed.

LE comparisons between magnetic and optical media are not straightforward because error rates are determined differently, and LE predictions of a hundred years will serve little purpose if the device error correction thresholds are exceeded. Accelerated aging of media is more useful when it can be correlated with LE values determined from natural aging.

2.10 *Number of Copies, Copy Locations, Media Choices*

Choosing a medium on which to store your remote sensing and geospatial data is both easy and hard. Easy in that there are many choices that work across many platforms; hard in that there are several factors to consider before choosing a particular medium. One of the first factors to consider is the size of the dataset being considered. If volume is relatively small (under a terabyte), consider making several copies, storing the relevant metadata with the data, and keeping at least one copy offsite. This is fairly straightforward for small datasets, especially if the data are static. Most organizations, however, will have growing datasets, even from older aerial photography missions if they are digitizing analog collections. So, how to weigh archive media choices? What are the factors that should be considered? Since 2001, the U.S. Geological Survey (USGS) has conducted an "Offline Archive Media Trade Study" to help determine media choices for its remote sensing and geospatial archives. These trade studies are conducted approximately every other year, which has worked well to document changing media technologies. The studies are intended exclusively to guides USGS activities, but the methodology and factors used may be of value to organizations wrestling with archive media decisions.

The factors chosen by USGS include the tape design criteria, transfer rate, capacity, cost analysis, real world scenarios, vendor analysis, and drive compatibility. The USGS Offline Archive Media Trade Studies can be downloaded from http://eros.usgs.gov/government/records/tools.php.

2.11 *Commercial Perspectives on Earth Observation Data Preservation*

In 1966, Lunar Orbiter-1 recorded the first image of Earth from space. The grainy image in Figure 6-20 was printed on posters and given out as gifts to US government officials and visiting dignitaries. The tapes containing the original data were originally stored in Maryland, and were only recently recovered and restored, through a heroic decades-long effort that involved not only preserving the original tapes, but also rebuilding the drives necessary to read the tapes. During the process of recovering the data, the team discovered that the actual data on the tape had up to twice the resolution of the original printed images. The data are now digitized, and stored in NASA's planetary data storage system, which is using electronic data storage as its standard storage medium.

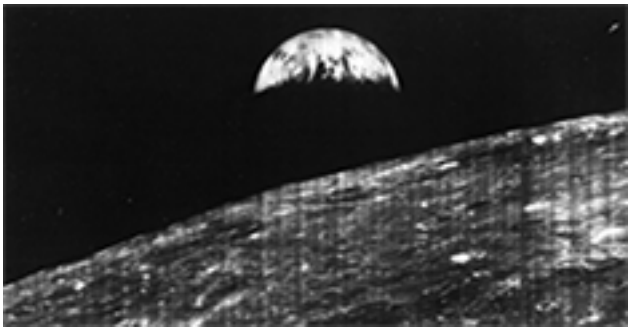


Figure 6-20. Earthrise: The first image of the Earth rising over the moon's surface taken on 23 August 1966 by NASA's Lunar Orbiter-1.

The importance of preserving Earth observation data is unquestioned. Having access to historical archives, such as the multi-decade Landsat archive, allows scientists to measure, analyze, and visualize longitudinal changes in the Earth's environment in ways that otherwise would be impossible. The benefit of storing the data increases when the data are not only preserved, but also made easily available, interoperable, and discoverable. These characteristics are best achieved by providing highly available, high fidelity, digital records of both source data and derived data.

In the past, Earth observation data were recorded on film and stored in vaults. With the advent of digital technology, bits are now recorded to tapes, which are then stored in vaults. Preserving those vault-stored data is the responsibility of the digital data preservation field, a topic too vast to cover in this section. However, advances in data availability and processing open possibilities for data preservation that have not existed until recently.

Among the important advances are the ever-decreasing expense of disk space, the even more rapid decrease in computational costs, the increased availability of super-computer like processing environments, and the public availability of software that can process petabytes of data in hours instead of years. Along with these new capabilities come additional requirements. Key to these challenges is the increased capacity to create and share large derived datasets which, themselves, become source data for further analyses. This process creates a need to track data provenance, maintain information about data integrity, and maintain multiple reprocessed versions of the data.

2.11.1 *Decreasing Costs*

Data storage costs have decreased by half every 14 years. A gigabyte of hard drive storage which cost $437,500 in 1980 costs $0.13 in 2014. As prices drop, storage density has increased from 0.1 gigabytes per square inch to a terabyte per square inch, decreasing the cost of the infrastructure's power, and physical space necessary to support it.

Processing power is becoming less expensive at an even faster rate. While the cost per microprocessor for consumer computers has remained roughly constant, performance has increased many times over. For example the first Intel 8086 chips released in 1978 had a clock speed of around 4*MHz*. Its modern counterpart, the Intel Core i7, released in 2014, has a single core clock speed of close to 4*GHz*. Over time, the cost of processing is cut in half every two years.

2.11.2 *Planetary-scale Processing*

Past analyses of EO data were largely restricted to operating on small areas and single scenes. As computer capabilities have increased, so has the ability to analyze larger areas and more scenes. Hansen *et al* (2012), for example, performed a global analysis of tree cover that leveraged the entire Landsat 7 archive to show forest change from 2000 to 2012. This involved analyzing over 600,000 Landsat scenes retrieved from a repository stored in Google's infrastructure. Although Google was not the original collector of the data, and is not responsible for their preservation, its repository is acting as a kind of backup system for the USGS data - a backup system with the additional feature of providing instant access to all the data. Being able to instantaneously access the entire Landsat-7 catalog, and having the data near where the computation was occurring, as well as being able to process the data in parallel over thousands of machines reduced the time to compute the global dataset in a few days rather than years.

2.11.3 *Compute It or Store It?*

The possibility of planetary-scale processing, coupled with the rapidly decreasing cost of CPU relative to storage, points in the direction of storing as little as possible when results can be generated dynamically instead. Raw EO data must be stored, of course. However derived datasets, especially those derived using a straightforward, well documented, method should only be stored in cases where the derived datasets are frequently used in analyses.

Consider the calculation of the normalized difference vegetation index (NDVI). This index, which is a measure of photosynthetic activity is the amount of red light (from $0.4\mu m$ to $0.7\mu m$) subtracted from the amount of near-infrared light (from $0.7\mu m$ to $1.1\mu m$) divided by the sum of these values. Imagine that we have a global land dataset at one meter pixel resolution. To generate NDVI for the globe would require terabytes of storage. There may well be places on Earth where nobody wants to know the NDVI at one-meter scale (for example, regions continually covered in ice,). The cost of computing the NDVI values is so low compared to the cost of physically storing and retrieving it that there's no point in storing the dataset. This is true for not only the dozens of commonly used indices but also for any well documented method that creates computable data easily. In general, there are so many datasets that could be useful that the task of generating and storing them all is impossible. Instead, providing a framework for on-demand calculation leads to what amounts to virtual exabytes of data (one exabyte is one million terabytes).

Further complications of storing data arise from the need to replicate the data for security and high availability. Data security can be ensured by maintaining multiple copies of the data. If the medium storing one copy of the data fails for some reason (e.g. hard drive failure, power failure at a data center) having other copies of the data ensure that they will always be available. The speed at which data may be accessed is limited by the speed at which data can be sent through a network (the upper limit of the speed of light being rarely achieved). Data access can be increased by housing data close to where they are accessed. Each replication of the data, for security or access speed, increases the amount of storage needed for each dataset. Dynamically generating the data, on the other hand, can be done by the same servers that are used to deliver the data, requiring no additional cost.

All that said, when a dataset is used frequently, and the computation time to create it is complex, storage makes a good deal of sense. In theory, geo-registration and ortho-rectification could be computed dynamically. However, a geo-registered and ortho-rectified version of an image is so often used as input to further steps, that it makes sense to store that version as well as the source data from which they are derived. The trade-off between dynamic computation of data and storage of data largely depends on usage. The rapidly decreasing cost of CPU, however, continues to move the threshold between storing and generating further into the generating side.

2.11.4 *Side Benefit of Hardware Progress*

Preserving data on tape in a vault increases the likelihood that they will become inaccessible. The data need to be actively migrated to new media before the lifespan of the storage media expire, and hardware that reads those media must be preserved as well. Maintaining an always-available and active data repository, on the other hand, requires that data providers move the data forward as hardware fails, or is replaced. Online systems can be put in place to automatically check the integrity of the data. Maintaining data on RAID servers helps ensure that the data are easily replaced with hard drive failure.

2.11.5 *Disk vs. Tape*

Tape storage is both cheaper and denser than disk, and the trends in cost reduction and density increase are similar (although some argue that hard drive density is starting to plateau). Although tape has advantages over disk in terms of price, energy consumption, and media stability, its latency makes it less useful for maintaining active databases. Medium reading latency is a combination of the time it takes to reach the data one seeks (seek time), and then to read the data once they have been found (streaming rate). Currently, hard drive and tape have similar streaming rates, but their seek times are quite different and will probably remain so. While a disk read head can move to any position in milliseconds, a tape must be physically wound to reach the data. This makes tape less useful in the context where data access is a primary component of data preservation. This is not an argument against using tape as a storage system, but rather, an argument for relying on living, accessible data storage as a primary means of data preservation.

2.11.6 *Challenges*

Preserving data in an always-available, active repository, and storing only raw and heavily used datasets presents some challenges, especially in the case where supporting reproducible analyses is desired. It may be necessary to maintain multiple versions of large derived datasets. For example, the process of converting

raw Landsat data to radiometrically, geometrically and terrain corrected (L1T) datasets is too complex to compute on demand easily. However, the United States Geological Survey, which is responsible for storing and maintaining the Landsat data, updates their calibrations frequently, requiring the maintenance of multiple versions of the data. Perhaps old versions could be thrown out, but this reduces the reproducibility benefits of having an active repository. Maintaining multiple versions of any given scene increases storage costs. This increase in the number of scenes that must be maintained can be quite dramatic in cases where an entire dataset is revised. For example, in 2014 the USGS renamed the metadata and band names for the entire set of Landsat-5 and Landsat-7 data. This affected millions of scenes. An active repository would either choose to store one version of the dataset and rewrite the metadata and band names on demand, or would make an entire second copy of the data repository. In this case, given the frequency with which the data are used, Google chose to replicate the entire dataset with the new naming scheme.

2.11.7 *Conclusion*

An infinite number of datasets may be derived from any source data, making the physical preservation of every derived dataset impossible. Which data will be valued cannot be predicted beforehand. Maintaining an active data repository where derived datasets are generated on the fly rather than stored solves the problem of having to store exabytes of data that might never be used. In addition, keeping the repository active helps guarantee that source data are preserved. As storage and computation costs decrease, we can expect a general move to maintaining data in high-availability, active production databases.

2.12 *Refresh Cycle*

Media have a viable life and need to be updated periodically. Analog materials such as aerial photography imaged on polyester safety film may last over 100 years with the proper temperature and relative humidity conditions. Aerial photography imaged on earlier media such as acetate base or nitrate would require an immediate transfer to the newer and safer polyester medium, as both of those possess serious degenerative properties. Optical and magnetic media have their own nuances. Optical CD-ROMs were thought to last generations, but due to optical reader changes, CD-ROMs can have a relatively short lifespan. A general rule today follows how fast technology changes. Often these changes far out-pace the stability of the physical media, which might last tens of years. Today, hardware, software, firmware, operating systems, formats, and compression routines, along with the physical media, all have to be considered. Below is a definition for long-term in regards to media (ISO 14721, 2012):

> "A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future."

This definition underscores the technologically fragile state of electronic media. Consider how many elements change within a three- to five-year period. Beyond that, electronic records may be in jeopardy.

3 DISCOVERY

Discovering the existence of a dataset, and learning more about it, is often the initial and essential activity for many researchers. This section contains guidance on metadata needed for discovery, how brokering activities can enhance search results, and how GeoSemantics can, and do, influence queries. Relevant examples exist from the NASA Global Change Master Directory, the Discovery Convergence and Coordination effort, OpenSearch approaches, and the Data Democracy 'movement,' which has gained momentum since inception in the 2000s.

3.1 *Metadata*

Science is becoming more visible. The scientific community is increasingly embracing a culture of data sharing, re-use, and integration (Michener, 2015). Open data access initiatives across research sectors are giving rise to a data sharing culture that emphasizes data release and data publication, and benefits data producers, users, and the science community (Porter and Duke, 2013). Foundational to this cultural shift is the recognition that metadata are a crucial part of the scientific workflow. To make their research methods and processes more transparent, scientists must ensure that their data are well understood and reproducible.

What are metadata? Metadata document the content and the context for a dataset. Often captured in formalized standards, a metadata record describes the story of a dataset: who created it, for what purpose, locations studied, dates, methodologies, processing steps, and the meaning of the data. Metadata records contain links to the data they describe, often in the form of persistent identifiers; thus, metadata records are often shared as a means of data discovery and access. A challenge for metadata is to permanently package it with the data they describe; metadata embedded in data is the wave of the future (Michener, 2015).

3.1.1 *Value of Metadata for Preservation of Data*

Data are reused in ways that increasingly extend beyond their original purpose. For example, a smaller dataset might be derived from a much larger source dataset and then integrated with other data in a computational model that, in turn, drives a scientific visualization of this new dataset. Metadata play a critical role at each stage of this scientific workflow. Scientific integrity is based on the ability to reproduce results accurately from science research, which requires an understanding of the assumptions, project design, methodology, processes, and data structures at each input stage. Metadata play a key role in understanding how a dataset was produced and are a fundamental requirement for potential future integration with other data.

In addition to sharing data, individuals may need to access the history of their *own* dataset. Years may pass between research projects, and these individuals may want to reuse data collected earlier in their careers. Over time, specific details (e.g., a particular processing step, a code or uniquely defined value in a data table) may be forgotten, preventing the individual from reusing or explaining their data reliably. Careful documentation of important details as they unfold in the course of a research project helps to prevent future misunderstanding and to ensure that data remain viable.

Without metadata, there is a risk that data will become obsolete or unusable. Date often represent a discrete 'snapshot' in time that can be vital to a broader understanding of the long-term processes of Earth systems. Metadata ensure that this snapshot can be reviewed, understood, and potentially integrated with

other data. Lastly, metadata records make data defensible. Science is often used in decision-making processes that affect public lands and interests. A metadata record provides critical information about how data were collected, processed, and analyzed. This documentation makes transparent the decisions made by scientists at each stage of the research process to support an informed conclusion.

3.1.2 *Standards for Describing Federal Government Data*

The Federal Geographic Data Committee (FGDC) oversees geospatial metadata standards used by the federal government. The organization is comprised of representation from nine federal bureaus and agencies. As part of its mandate, the FGDC is charged to develop, support, endorse and promote metadata standards for geospatial data. Currently, the FGDC supports and endorses the Content Standard for Geospatial Metadata (CSDGM) and the International Organization for Standardization (ISO 19115) suite of standards. Various data types such as remote sensing data can be described with either standard; detailed descriptions of instrumentation and other similar details may best be captured in the ISO suite of standards, as they are updated for more current technologies. The FGDC website (www.fgdc.gov) contains information about these metadata standards.

3.1.3 *Best Practices for Data Documentation*

It is extremely important to create a comprehensive metadata record that captures the key information about a dataset.

- Organize any pre-existing information about the dataset, such as project proposals and data management plans, as text can be reused. A project proposal abstract and purpose can be copied into a metadata record, and similarly, a data management plan might contain information about various datasets that were integrated into the new dataset. These are valuable records about the data that are already written or recorded and that can be reused in the standardized record.
- Craft the title of the dataset carefully. Create a title that includes information about the data – what, where, when, who and scale are examples of information to include in an informative title. An example: Greater Yellowstone (*where*) Rivers (*what*) from 1:126,700 (*scale*) U.S. Forest Service (*who*) Visitor Maps (1961-1983) (*when*).
- Use established, published thesauri or controlled vocabularies supported by your community of practice for keyword choices. Keywords are critical for data discovery. A thesaurus contains vocabulary that is broader, narrower, and related to the original terms selected. Inclusion of these terms in the record enhances discovery potential for the record. If a single vocabulary is not comprehensive enough to describe the scope of the data, one can use and cite additional resources and keywords to ensure that the full scope is described. For example, one might use a theme keyword from the ISO 19115 Topic Categories, as well as theme keywords from the AGROVOC Thesaurus and theme keywords from the USGS Bio-complexity Thesaurus. Examples of various keyword authorities include:
    - for theme keywords, the CENDI index of terminology resources (http://www.cendi.gov/projects/proj_terminology.html) provides a list from a wide range of science and technology fields;

  - for place keywords, the Geographic Names Information System (http://nhd.usgs.gov/gnis.html);
  - for stratum keywords, the Oregon Geospatial Enterprise Office Stratum Keyword list (http://www.oregon.gov/DAS/CIO/GEO/Pages/fit/thesaurus/stratumkeywords.aspx); and
  - for temporal keywords, the Oregon Geospatial Enterprise Office Temporal Keyword list (http://www.oregon.gov/DAS/CIO/GEO/Pages/fit/thesaurus/temporalkeywords.aspx).
- Be specific and quantify whenever possible. The goal of a metadata record is to allow a user to understand the data without the need to contact the dataset owner.
- Do not skip describing the entities and attributes in the data. These definitions are the most critical parts of a metadata record. They provide an understanding of the columns and rows of data and the units of measure.
- Provide detailed information about processing steps, methodologies, and data quality. These are important materials for the successful reuse of data, detailing how they were collected and processed, steps taken to ensure their quality, and why the dataset may reflect potential gaps in data.
- Keep the metadata packaged with the dataset. This is extremely important, since the main purpose of a metadata record is to understand the data;
- Use persistent identifiers such as Digital Object Identifiers to point to online public access points for the data and data products, and ensure that these DOIs are managed;
- If changes are made to the data after publication, make sure that the metadata record is updated to reflect these corrections or additions.

Other resources are available for additional details and information, including:

  - USGS Data Management website (www.usgs.gov/datamanagement); and
  - DataONE (https://www.dataone.org/best-practices).

3.1.4 *Tools Available to Document a Dataset*

Many metadata tools exist for creating records in a variety of standards. The choice of metadata tool depends largely on the type of data being described. As a longstanding and stable metadata standard, CSDGM is more broadly supported by tools than is the ISO suite of standards, which is newer and still evolving. Available geospatial metadata tools include:

- Online Metadata Editor (USGS) supports FGDC CSDGM and the Biological Data Profile;
- Morpho (UCSB-NCEAS) supports Ecological Metadata Language;
- Metadata Wizard (USGS) supports FGDC CSDGM (used with ArcGIS- ESRI);
- ESRI Tools (ESRI) supports FGDC CSDGM; ESRI ISO;
- TKME (USGS) supports FGDC CSDGM and its profiles and extensions;
- Metavist (USDA Forest Service) supports FGDC CSDGM and the Biological Data Profile;
- XML Editors (commercial) can support any standard scheme.

3.1.5 *Format for Metadata Records*

Metadata records are meant to be machine-readable to support ingest by online metadata catalogs; thus the most prevalent format for metadata records is eXtensible Metadata Language (XML). Tools that create metadata records generally export the record in XML format (Devarakonda, 2014)

3.1.6 *Publishing Metadata*

Once metadata are created, they can be used for many purposes. Primarily, descriptions such as these are used for discovery, access, and understanding of data. However, metadata are also used for such activities as data inventory assessments, defending science, marketing and promotion of data, and records management.

Metadata can be discovered only if they are published in catalogs, clearinghouses, and other portals. In the current environment, metadata records are "harvested," or copied, from an original source into online, searchable spaces such as the federal data.gov catalog, state-based clearinghouses, or those hosted by non-profit organizations.

Publication of metadata enables data discovery, promotes data sharing, and facilitates understanding and potential reuse of data. In an era of tight budgets and limits to science funding, it is critical that data be reused and integrated. Metadata records are absolutely critical to this process. Data that lack metadata are considered incomplete. Metadata allows data to retain value.

3.2 *Brokering Services for Remote Sensing Data*

The demand for remote sensing data and the number of sources for such data have both been increasing dramatically. Moreover, the services and attendant protocols for obtaining these data have become more diverse, as have the categories of users of remote sensing data. These trends have produced challenges to interoperability for data providers, data users and system designers.

Interoperability is the ability of two systems to interact without *a priori* knowledge of the internal operations of one system by the other. Traditionally, interoperability has been achieved through conformance to standard protocols for exchanging messages and data via exposed interfaces. However, with the increasing heterogeneity of systems and the rapidly-evolving nature of information technologies, it has become impractical for system designers to anticipate and build to all potential user demands while at the same time users, who range from scientists and engineers to citizens, decision makers and planners, are expecting to access data by means and in formats that are familiar to them.

One solution to these challenges is the deployment of middleware that mediates the interactions between different systems. Commonly called brokers, such middleware translates messages and mediates the exchange of data and information between otherwise incompatible systems. The systems involved can be as varied as a data repository's catalog services, a community web portal that aggregates data from multiple sources, a scientist's desktop analysis tool, or a user's mobile device. Brokering allows a data provider to store and serve information according to one set of standards and protocols while a broker makes it possible for users of those services to query the repository and access the data via different protocols.

Ideally, brokers are deployed as an infrastructure service, maintained and operated independently of any one data repository or data portal. In this way the burden of having to adapt to constantly evolving standards is removed from data providers and users and is instead borne by the middleware, which is a burgeoning field of research and is exploring methods of automatically detecting and adapting to changes in system behavior (Blair *et al*., 2011)

3.2.1 *Brokering in Support of Data and Service Discovery*

One of the most common applications of brokering is in support of data and data service discovery across distributed systems. Historically, a user had to query each system separately, in the manner required by that system, to find the information they were seeking. A broker can overcome this limitation by translating a user's query into the protocol used by each system, distributing the query to the various systems, receiving the responses, combining them and then returning the results to the user. The user may be either a human submitting the query via an online search interface, or a machine acting as a client. The broker can either distribute the query as it is received, or it can issue the query against a metadata store that was populated earlier via harvesting. In the latter case metadata harvesting must be performed on a periodic basis, but the advantage is that a user will be able to get results from systems that happen to be offline when the query is issued.

A broker can act as a catalog service responding to multiple query protocols. It also knows how to issue queries via different protocols to different data sources. Several functions can come into play when distributing queries and aggregating results: two-way translation of query and response protocol and content; mapping of content between data models; ranking of aggregated query results; semantic expansion of query terms.

When a broker harvests metadata from heterogeneous systems it may map incoming metadata to a common internal data model, which then is the target of the user's query. If the user requests the full metadata records for discovered resources the broker can either export the normalized metadata for those records or it can return the original heterogeneous metadata records. Since the internal queryable data model consists of fields that are common across all systems, it may represent only a subset of the fields in the original record, thus the latter should be an option to prevent any information loss.

Numerous commercial and custom software packages are available with a range of capabilities, including mining of unstructured information, but many of these are aimed at enterprise-level distributed systems.

3.2.1.1 Challenges

Among the most important metadata elements for data discovery are those that relate to data quality, lineage and fitness for use. The manner in which information is expressed in different system and scientific domains can vary widely, making it especially difficult for users who must interpret heterogeneous metadata records. The domain knowledge required for this interpretation can be captured, to some degree, in the broker module that is developed to map a particular kind of data source to the common model. For this reason it is best if domain experts are involved in the development of such modules. Other challenges in maintaining a broker are variations in implementation of declared standards, and managing changes in brokered resources.

3.2.2 *Brokering in Support of Data Access*

Data discovery is typically followed by a user ordering data of interest. A broker makes it possible for data stored in one format to be provided to a user in another format more suitable to them. Access can be enabled by either passing an order to a data providers' ordering interface, or via a click-through by which a user is given immediate access to the data via a web browser. The broker, if so designed, can also perform other operations on the data in addition to format conversion, such as re-gridding, re-projection or extraction

of spatial, temporal or data layer subsets. In fact, advanced brokers are capable of managing workflows for processing data as specified using protocols such as *Business Process Modeling Language* (BPML).

3.2.2.1 Challenges

When data are modified in any manner by a broker before delivery to a customer, a record of those transformations must be included in the supplied metadata. Moreover, it is vital that the original metadata indicating source and attribution also be delivered to the customer. Many systems require user authentication before delivery of data if for no other reason than to track usage of the data. When access to data is restricted, or there is a fee associated with data delivery, the broker can be bypassed so that the customer interacts directly with the provider. When a results set involves many different providers, it is desirable to have the broker pass a user's credentials to the various provider systems. The mediation of authentication and authorization are areas of active research. Favored solutions are based on open source standards such as OAuth, OpenID, SAML and shibboleth.

3.2.3 *Towards Brokering as a Common Good Service*

While brokering provides a valuable service when employed by a data repository, community portal, or business enterprise, its benefits are multiplied when the service becomes universally accessible as a common good. Rather than having each broker implement, learn, and develop modules for the various protocols and standards of the systems it serves, a common broker, by translating all standards to a common model, enables many-to-many interactions and reduces overall effort. What is necessary then is a sustainable governance model. The most likely model for broker support is a combination of community support through contributions to an open-source code base and foundation or government support for infrastructure maintenance. The sociological factors of infrastructure development must also be considered and this is an area of active research.

3.2.4 *Summary and Conclusions*

Data discovery, access and usage would be vastly simplified if there were only a few universal standards that everyone used. While in some cases technology standards tend to converge through market forces, it is likely that the heterogeneity in the standards and protocols used for remote sensing data and services will persist and may even increase. For this reason the brokering solution is likely to continue to play an important role. Through brokering it is possible to provide data to users in a way that suits their needs without placing extra burdens on data providers to satisfy diverse user needs. Furthermore, by translating between legacy standards and new version of standards both provider and users can be shielded from shifts in technology. This makes it possible to maintain a current standards baseline while simultaneously supporting new standards that accommodate evolving and potentially disruptive technologies, community needs, and market trends.

3.3 *Global Change Master Directory and the International Directory Network*

The Global Change Master Directory (GCMD) is a repository of information about Earth science data sets and services. Developed and maintained by NASA, the GCMD holds over 34,000 descriptions of data

sets and services as of January 2015, covering a broad range of subject areas within Earth and environmental sciences. The descriptions and the links provided by GCMD help researchers, policy makers, and the public discover data and related services, and related information relevant to global change and Earth science research. The GCMD also provides the tools necessary for data and service providers to make their products discoverable by the user community.

The origin of GCMD was in the prototype NASA Master Directory (NMD) as a part of the National Space Science Data Center (NSSDC) at NASA/Goddard Space Flight Center (GSFC) to promote exchange of scientific data sets. The first version of the NMD was released in 1987 through the Catalog Interoperability (CI) project based on the type of information and the level of detail defined by the CI Working Group (consisting of several U.S. Federal and international agencies). In 1990, the NMD was adopted by the Interagency Working Group on Data Management for Global Change (IWGDMGC) as a prototype to facilitate global change research. Thereafter, the part of NMD addressing Earth sciences data was renamed the Global Change Master Directory (GCMD). Organizationally, the GCMD project became a part of the Global Change Data Center within the Earth Sciences Directorate at NASA/GSFC in 1994. The GCMD is a part of the core capabilities of NASA's Earth Science Data System Program and the Earth Observing System Data and Information System (EOSDIS), providing collection level discovery and access that complements the granule level discovery and access provided by the EOS Clearing House (ECHO).

In 1989, the Committee on Earth Observation Satellites (CEOS) Data Working Group (DWG) established the CEOS International Directory Network (IDN) to foster the exchange of information among international agencies. The GCMD became the underlying "engine" of the IDN, in that many of the capabilities and content of GCMD are carried over into IDN. The IDN provides many portals to expose subsets of the content specific to individual organizations or projects. In this section, we present the structure, organization, features and governance of GCMD and its relationship to IDN.

3.3.1 *Participating Organizations*

While the primary responsibility of GCMD is to maintain a complete catalog of all of NASA's Earth science datasets and services, there are many other organizations within the United States and around the world that contribute to the GCMD. In fact, of the over 34,000 datasets maintained in the GCMD today, the number of datasets from NASA is approximately 5,500. Other providers include U.S. federal agencies, U.S. state and local agencies, non-U.S. government organizations, academic institutions, non-government/non-profit organizations, commercial entities, consortia, and multi-national organizations. The proportions of datasets in the GCMD catalog from these various types of organizations are shown in Figure 6-21.

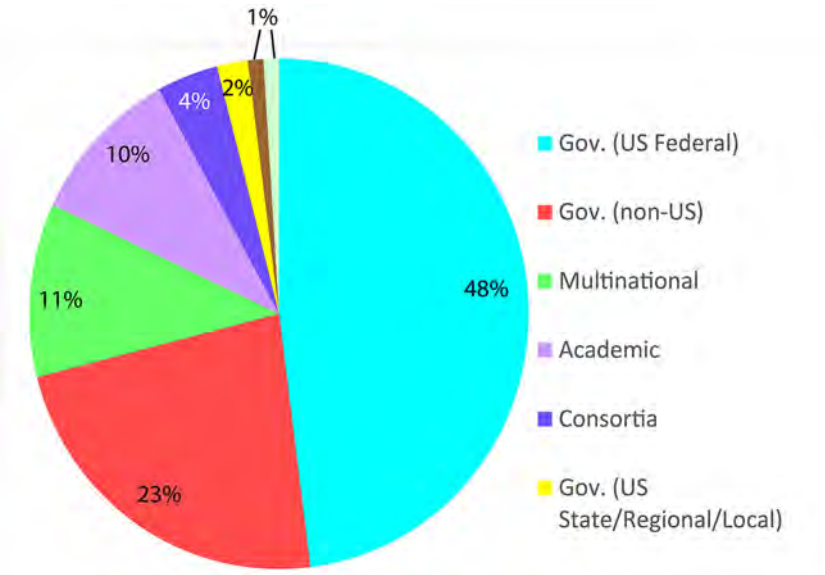


Figure 6-21. Proportions of datasets in the GCMD Catalog from different organizations.

3.3.2 *Entering Information into GCMD*

There are several types of information or metadata that a data or service provider can enter into GCMD. These are descriptions of:

- Data sets
- Data services
- Ancillary items
  - Projects or Campaigns
  - Instruments
  - Platforms
  - Data Centers
- Climate Diagnostic Visualizations

Each of these types of information is entered into data fields using a predefined template. The template for data set description is called the Directory Interchange Format (DIF). The template for data service description is called Services Entry Resources Format (SERF). Descriptions of climate visualizations in the directory are tagged with one or more of the potentially significant Societal Benefit Areas (Disasters, Health, Energy, Climate, Water, Weather, Ecosystems, Agriculture, and Biodiversity). Detailed guides for developing each of these types of descriptions are available at http://gcmd.gsfc.nasa.gov/collaborate/docbuilder.html. To help contributors to GCMD, the instructions include identification and definitions of “Required”, “Highly Recommended” and “Recommended” fields. Table 6-5 shows the required and highly recommended fields for the various types of descriptions.

Table 6-5. Required and Highly Recommended Fields in GCMD/IDN Descriptions.

Description	Required Fields	Highly Recommended Fields
Datasets	Entry ID, Entry Title, Science Keywords, ISO Topic Category, Data Center, and Summary	Temporal Coverage, Spatial Coverage Paleo-Temporal Coverage, Location Data Resolution, Dataset Citation Personnel, Project, Quality, Access Constraints, Use Constraints, Distribution Related URL, Instrument, Platform Dataset Progress, and Dataset Language
Services	Entry ID, Entry Title, Science and Service Keywords, Service Provider, and Summary	Distribution, Instrument, Platform, and Project
Projects or Campaigns	ID, Textual Description	
Instruments	ID	
Platforms	ID	
Data Centers	ID, Textual Description	
Climate Diagnostic Visualizations	Entry ID, Entry Title, Science Keyword, Visualization Provider, Visualization Description, and File Attributes	Science Keywords, Temporal Coverage Spatial Coverage, Paleo-Temporal Coverage, Location, Data Resolution Visualization Provider, Visualization Citation, Publications/References, Personnel, Quality, Use Constraints, Related URL, and Instrument, Platform

The DIF was one of the first instantiations of a way to enter metadata about a science data set, and it became a *de facto* standard for NASA Earth science data. Other standards have been developed. In order to interoperate with these new standards, the GCMD developed mappings between the DIF and each new standard. These standards include: the ISO 19115/TC211 Geomatics Metadata Standard, the Federal Geographic Data Committee (FGDC) Metadata Content Standard for Digital Geospatial Metadata, the ESRI Profile, the Dublin Core Metadata, and the Australia and New Zealand Land Information Council (ANZLIC) Metadata Standard.

3.3.3 *Discovery*

Based on the descriptions entered by the providers discussed above, users can discover datasets and services relevant to their needs. A user can also get information about project, campaigns, instruments, platforms and data centers by looking at the appropriate ancillary descriptions. Figure 6-22a shows the starting point (http://gcmd.gsfc.nasa.gov/index.html) to search for types of information from the GCMD. There are distinct areas to click on for datasets, services and ancillary descriptions. Also, free text searches can be used to find datasets and services. Datasets can be searched by various criteria (Figure 6-22b), that is, Science Keywords, Instruments, Platforms, Locations, Providers, Projects, and Map/Date. Each of these criteria leads to a hierarchy of screens, indicating the numbers of datasets available at each level for all the options available at that level. At the bottommost level of the hierarchy, the user gets a list of all the datasets meeting the selected criteria and links to the complete metadata records, which have information describing the datasets as well as from where to obtain them (Figure 6-22c). Searching for services works similarly, with the following criteria - Science Keywords, Instruments, Platforms, Providers, and Projects.

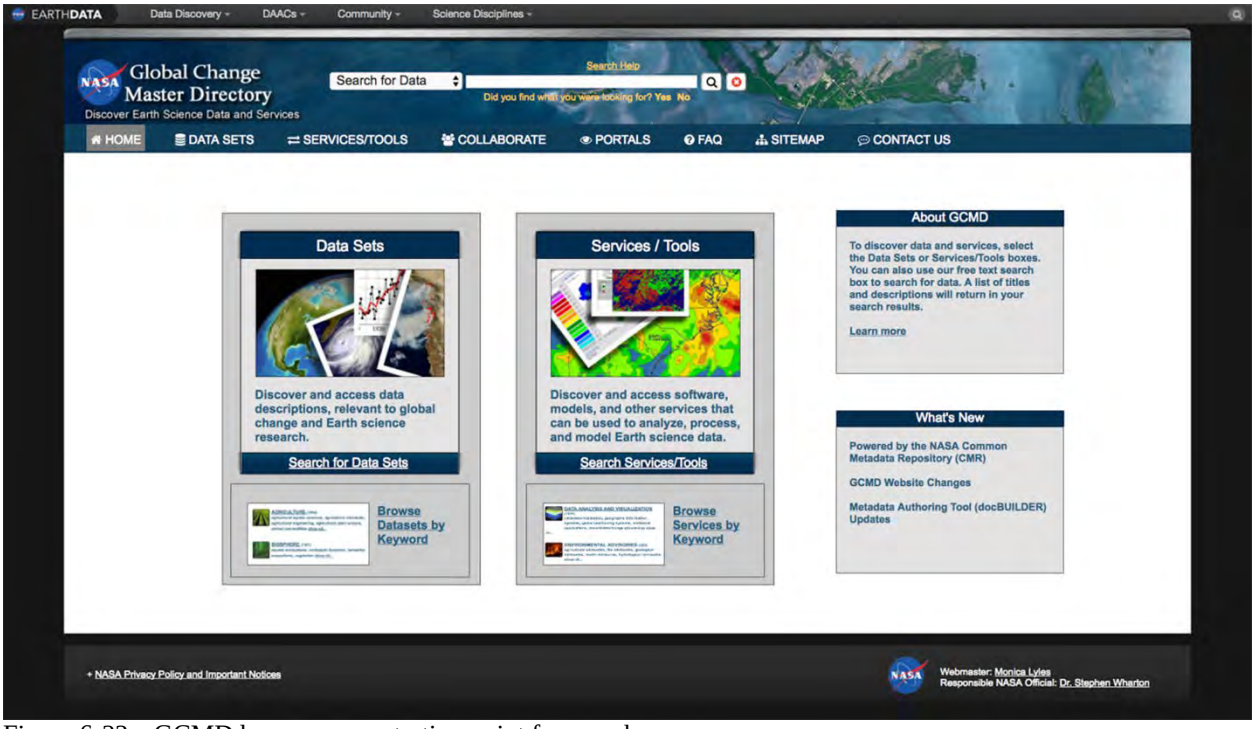


Figure 6-22a. GCMD home page – starting point for searches.

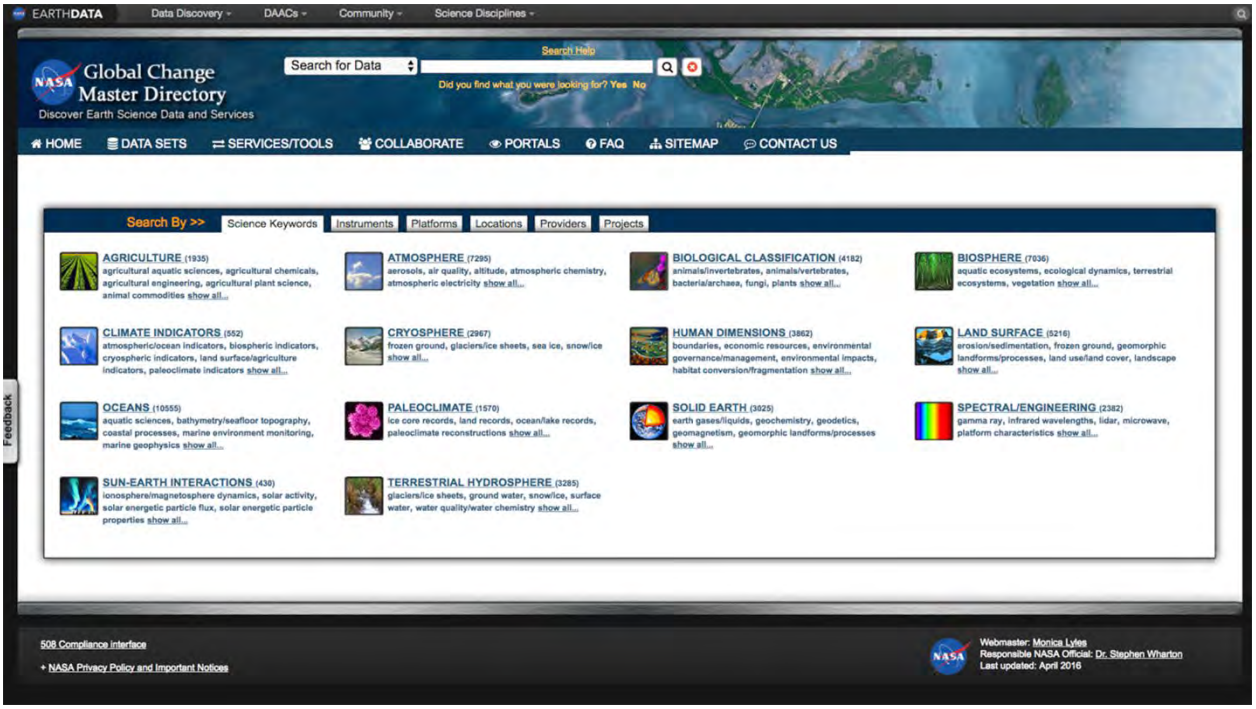


Figure 6-22b. Searching for datasets.

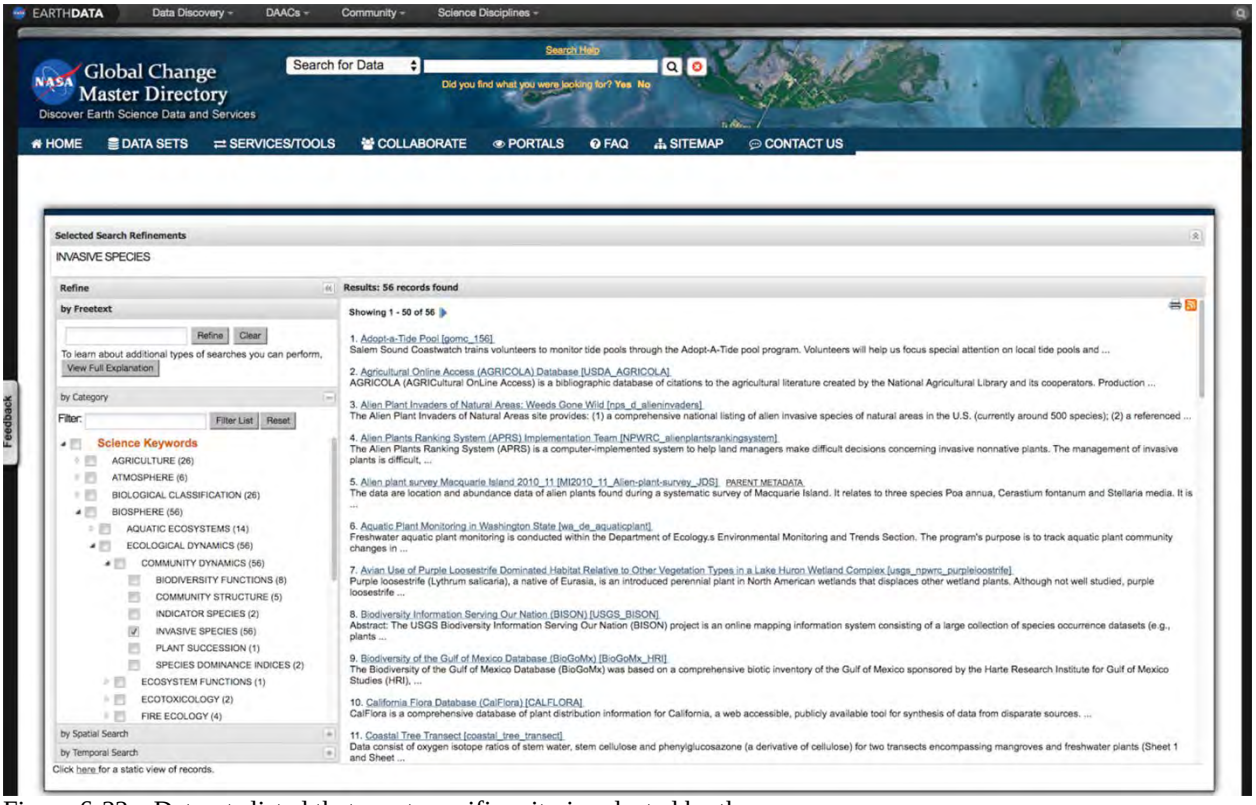


Figure 6-22c. Datasets listed that meet specific criteria selected by the user.

Additionally, the GCMD/IDN provides a number of portals. Portals are views into subsets of the complete directory, which have been customized for various partner organizations. Portals provide focused views, which organizations can maintain within GCMD/IDN, without needing to create a separate on-line directory. Portals may be trademarked with organizational logos and provide full functionality of the GCMD/IDN search engine and tools. When data providers add metadata to a portal, even though the portal is limited to a particular organization's view, those metadata are visible to anybody who accesses GCMD since those inputs become part of the overall GCMD metadata database. More than 60 different organizational or discipline oriented portals are in use in 2015(http://gcmd.nasa.gov/add/portals.html).

One such portal is the Climate Diagnostics Directory, which provides over 850 visualizations of climate diagnostic images in the following disciplines: Oceans, Land Surface, Atmosphere, Terrestrial Hydrosphere, Cryosphere, Biosphere, Climate Indicators, and Paleoclimate. An example of a visualization page from the Climate Diagnostic Directory is shown in Figure 6-23.

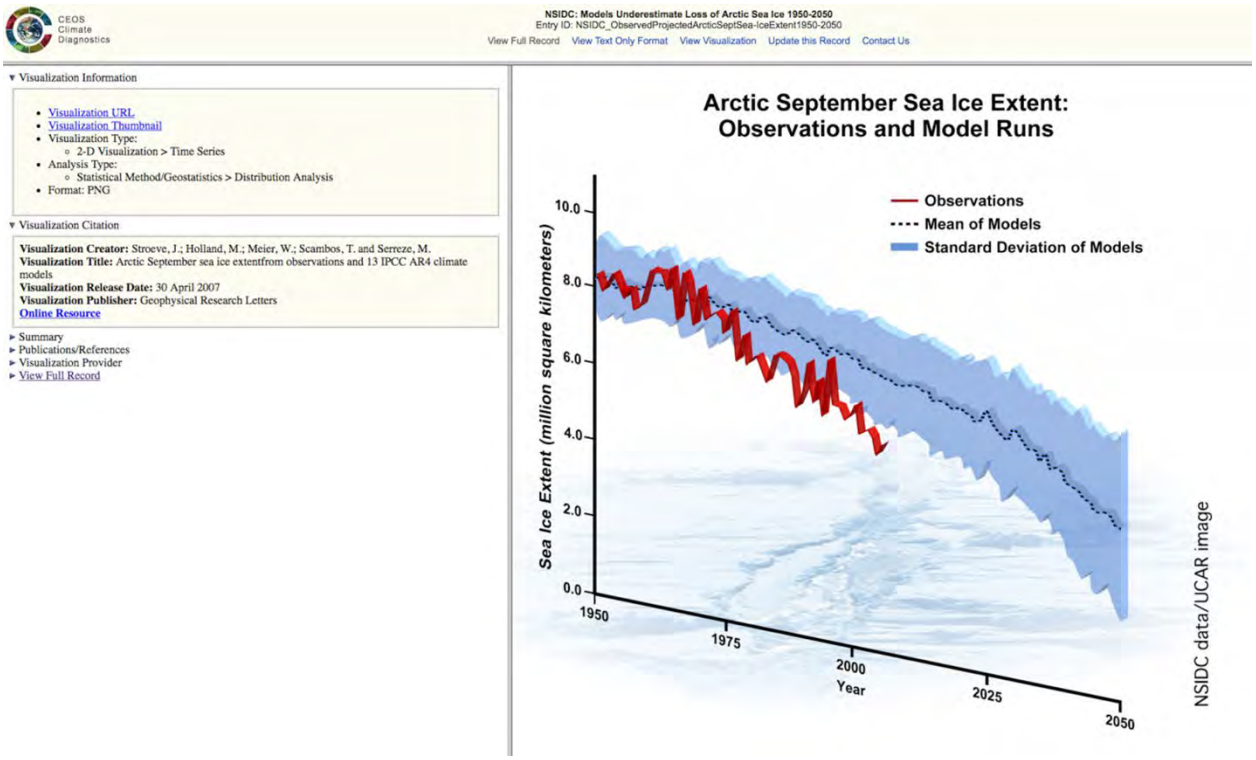


Figure 6-23. Example of the visualization page from the Climate Diagnostic Directory (Stroeve *et al*., 2007).

3.3.4 *Keyword Management*

To facilitate search and access, GCMD science coordinators have developed four sets of controlled keyword vocabularies (Earth Science, Data Services, Ancillary, and Enumerated), and integrated them into GCMD's user interface. The use of controlled keyword vocabularies ensures that Earth Science search terms and categories are described in a consistent manner to allow precise searching of metadata records and subsequent retrieval of Earth Science data and services. Today, meaningful keywords are the basis for most science data discovery. GCMD has been leading the way for developing a controlled vocabulary for NASA Earth Science data. It has become the 'gold standard' in keyword vocabularies. Numerous U.S. and international agencies, research universities, and scientific institutions use the GCMD controlled vocabularies as an authoritative vocabulary, taxonomy, and thesaurus: (1) to describe Earth Science data, services, and visualizations; (2) to search and access Earth Science data within a metadata catalogue and/or website; and (3) to create ontologies for the semantic web.

Established over twenty years ago, these controlled vocabularies maintain their relevance by continuing to be refined and expanded based on established processes, keyword principles, and rules. GCMD science coordinators, with input from users, metadata providers, and the NASA Earth Science Data and Information System (ESDIS) Project's Standards Office, review recommendations regularly for keyword additions, modifications, or deprecations. Once reviewed, additions and revisions are made following guidelines based on pre-defined and controlled vocabulary rules and principles that delineate what constitutes a well-curated keyword list.

After internal review, submitters are invited to review the keyword changes and upon agreement, the authors of all affected metadata are notified to indicate that the changes are acceptable. After the changes

are implemented by the science coordinators, a new keyword release announcement is posted on the GCMD web site and sent to user communities and metadata authors by subscription email. GCMD controlled keyword vocabularies are available to the public at http://gcmd.nasa.gov/learn/keyword_list.html.

3.3.5 *Evolution of Metadata Systems*

It is clear that metadata are the basic driving force for EOSDIS and its components, and most especially in GCMD. At the start of the development and deployment GCMD data were held on tapes and only the metadata were accessible on-line. The GCMD was developed as a repository of metadata at the "directory" level, providing information about collections. The number of records in GCMD ranges in tens of thousands. In the mid-1990's NASA developed "inventory" level metadata, providing information about granules (instances of files of instrument sensor data and/or derived higher-level products within collections) to facilitate search and access to specific data within a collection identified by spatial, temporal and other constraints. The EOSDIS Clearing HOuse (ECHO) is a metadata repository at the "inventory" level. The number of records in ECHO ranges in the hundreds of millions, and increases with every measurement taken by a satellite or aircraft. ECHO is the *middleware* over which clients are built for accessing granules of interest across multiple data archive centers. Data archive centers provide metadata to both GCMD and ECHO. Users need both collection and granule level information for various purposes. For example, the Reverb client, developed and operated by the ESDIS Project, provides cross-data center search and access capabilities, and uses both collection and granule level metadata. In combination, GCMD, ECHO and Reverb have evolved along with data storage and system technologies to expedite access to vast quantities of on-line data at the data file level in EOSDIS. Subsetting and other operations on granules can be performed as a user accesses data through various clients. This is quite a change from the original intent and capabilities of GCMD twenty-five years ago.

Future evolution of EOSDIS will require that GCMD and ECHO metadata be consistent. Consistency between the two has been difficult to achieve since the systems were managed separately. Given this situation, the ESDIS Project initiated an effort in late 2013 to develop a Common Metadata Repository (CMR). The CMR uses a Unified Metadata Model, and takes advantage of the efforts in GCMD and ECHO, to provide a unified and authoritative repository for NASA's Earth science metadata. The CMR will accommodate not only the collection and granule inventory metadata, but will allow new metadata concepts to augment data systems including: Variable Parameters, Documentation, Services, Visualizations, and products. CMR provides a flexible ingest system with pluggable adapters to handle multiple metadata record formats, multiple metadata record concepts, and relationships and validations between them. The CMR is designed to handle hundreds of millions of metadata records. These records will be available through a high-performance, standards-compliant, temporal, spatial, and faceted search. Quality assurance of the metadata in CMR is handled through human and machine assessment, with automated metadata scoring rubrics giving data providers some insight into how to make their data more discoverable and usable. In addition to automated metadata scoring, manual review of metadata is conducted and data providers are given recommendations on correcting the identified errors or omissions. Also, the CMR will be able to take multiple metadata records associated with a common core concept, such as a Collection, and merge the disparate information into a robust and standards-compliant (ISO-19115, 2003) representation for interested

clients. As of this writing, significant progress is being made in the implementation of CMR. The first phase, which had a primary objective to improve search and access performance, has been completed.

As CMR evolves, processes are evolving to ensure consistency and the rules of interaction between the metadata providers and the system. Two of the governing documents behind this evolution are:

- GCMD Keyword Community Guide (NASA GSFC, 2014), which records the process for submitting, reviewing and accepting GCMD keywords; and
- CMR Life-Cycle Document (NASA GSFC, 2015), which specifies how all CMR element requirements are managed, updated, modified, reviewed and approved for implementation through a process involving all the stakeholders.

3.3.6 *Conclusion*

This section has provided a brief description of the Global Change Master Directory (GCMD) and the International Directory Network (IDN). Developed and maintained by NASA, GCMD holds over 34,000 descriptions of data sets and services as of January 2015. It covers a broad range of subject areas within Earth and environmental sciences. The descriptions and the links provided by GCMD help researchers, policy makers, and the public discover data and related services; and, related information relevant to global change and Earth science research. The GCMD also provides the tools necessary for data and service providers to make their products discoverable by other user communities. Numerous U.S. and international agencies, research universities, and scientific institutions use the GCMD controlled vocabularies as an authoritative vocabulary, taxonomy, and thesaurus to describe Earth Science data, services, and visualizations; to search and access Earth Science data within a metadata catalogue and/or website; and to create ontologies for the semantic web. The GCMD is also a part of the International Directory Network (IDN), organized by the Committee on Earth Observation Satellites (CEOS). The GCMD is the underlying "engine" of the IDN, in that many of the capabilities and content of GCMD are carried over into IDN. Focused views are provided to organizationally specific subsets of GCMD, which can be maintained within GCMD/IDN without needing to create a separate on-line directory.

To meet future demands for new data discovery tools not only from GCMD's user interface, but also for a broad commercial audience, older metadata structures need to be changed in order to facilitate future modifications , hence the development of the EOSDIS Common Metadata Repository (CMR). It once used to take years to modify the GCMD database structures. With the CMR and features under development, enhancements to the underlying metadata model will be achieved quickly enabling new capabilities for data discovery. It is natural for users to want to find data quickly, but sifting through hundreds of Google links is not the best way to find data that users need. Science users will need the ability to sift through collections of data using tools that have not yet been invented. In future, NASA's rich collection of Earth science data will be discoverable in many new ways and there will be a GCMD continuing to provide the directory tools needed to traverse the complex repositories of Earth Science data.

3.3.7 *Acknowledgement*

Section 3.3 is a summary of material from the web pages under https://earthdata.nasa.gov maintained by NASA. The authors are grateful for the review of the initial draft provided by Stephen Wharton and Scott Ritz (NASA Goddard Space Flight Center). H.K. Ramapriyan acknowledges the support provided by

NASA under the contract NNG15HQ01C. N.L. James and J. Behnke performed this work as a part of their duties as employees of the United States Government.

3.4 *EO Data Discovery Convergence / Coordination*

Earth observation through remote sensing is the most commonly used method to collect Earth science data. The platforms that carry remote sensors can be ground- or sea- based, airborne, or space borne. Because of the advantages of continuous observation and global coverage, space-borne (e.g., satellite) remote sensing has become the most popular method of Earth observations, especially when the purpose of such observation is to solve regional, country, continental, or global issues. Because of the importance of Earth observations in socio-economic activities and national defense, many governmental and private organizations around the world are actively conducting remote sensing-based Earth observations. For example, the Committee on Earth Observation Satellites (CEOS), an international organization of space-agencies that coordinate and harmonize civilian satellite Earth observations to make them easier for user communities to access and utilize data, has 55 member agencies operating 135 EO satellites. Several commercial companies, such as Digital Globe, also operate remote sensing satellites. Through those organizations, huge volumes of EO data have been collected. At the end of September 2014 the NASA EOS program alone had collected over 10 petabytes of EO data.

To facilitate data discovery, data centers of those agencies have created catalog systems, which are located in individual agencies worldwide and are mostly Web based. Each system may have its own interface protocols, metadata information model, and clients. Therefore, if a user wants to find data they have to know where those data are located, and how to find them through the catalog system. However, there are so many data centers in the world that it is not easy to know where needed data are located. To solve this problem, a two-step data discovery approach has been developed. The first step is to find where the data that users want are located, and the second step is to find data granules at that location. The first step is called the *directory-level* discovery and the second step, the *inventory-level* discovery.

The system that facilitates the directory-level discovery is the registry that contains information about *types*, which are defined by well-known vocabularies. Registry normally implements a registry interface that allows data and service providers to register their data and services; and a search interface that returns metadata or the names of *types* and the catalog, which contains information about instances of the types.

One of the important registries for Earth observations is the Global Change Master Directory (GCMD). It also serves as the International Directory Network (IDN) Master Directory of CEOS. The GCMD has also developed a vocabulary, called GCMD Science Keywords, which cover the disciplines of Earth observations and related societal benefit areas. The vocabulary has been widely used in the Earth science community. The details of GCMD can be found in Section 3.3.

Another important registry is the Component and Service Registry (CSR) of the Global Earth Observation System of Systems (GEOSS). Coordinated by the Group on Earth Observation (GEO), GEOSS, being built by the contributions from over 90 member countries and more than 20 participating organizations, will provide decision-support data and tools to a wide variety of world-wide users addressing nine societal benefit areas of Earth observation. These are: Agriculture, Biodiversity, Climate, Disasters, Ecosystems, Energy, Health, Water, and Weather. GEOSS consists of the GEOSS Common Infrastructure (GCI) and

members-contributed EO resources interoperable through recommended standards and best practices. As one of the components of the GCI, CSR provides two interfaces: one for the resource contributors to register their resources to GEOSS; and another for resource users (both machine and human) to discover those resources. Resources include the datasets, the data catalogs, services, computing facilities, and EO infrastructure. CSR is mainly used by another GCI component, the GEO Discovery and Access Broker (DAB), which provides brokered inventory-level data discovery and access.

Inventory-level data discovery is facilitated by data catalog systems. Currently, almost every EO organization in the world has a web-based data catalog containing information about their datasets (aka granules). Therefore, the catalog systems for inventory-level searches are numerous. Some large EO catalogs contain millions of records, each for a granule.

Because individual data catalogs have been developed by individual agencies at different times and with different technologies, they are very diverse in their use of metadata models, search interfaces, request and response schema, and search behaviors. The diversity makes the development of general catalog clients difficult. Users need to deal with multiple catalog systems to find all the data they need. To address this issue, three approaches: standardization, metadata harvesting, and catalog federation, have been developed. The standardization approach attempts to develop catalog standards and promote their use in developing new catalog systems. Multiple catalog standards have been developed by standard setting organizations, most notably, the Open Geospatial Consortium (OGC) and Technical Committee 211 of International Standardization Organization (ISO/TC 211, 2003). The most commonly used interface standards are the OGC Catalog Service for Web (CSW) and its profiles, and OpenSearch. The commonly used metadata information models include ISO 19115 and OASIS ebXML Registry Information Model (ebRIM). The catalog and registry systems mentioned in this section are all following these standards.

Both metadata harvesting and federation approaches provide a single point of entry (portal) with a standardized information model and interfaces for users to discover the data cataloged in diverse cataloging systems produced by multiple organizations; i.e., the one-stop shopping approach. The metadata harvesting approach attempts to harvest metadata in all related catalog systems to build a centralized clearinghouse for users to find the data they need. Since all metadata records are stored locally, the advantage of the harvesting approach lies in performance and reliability. However, there are significant disadvantages in this approach, including catalogs too big to be harvested, outdated records because of infrequency of harvesting, and unavailability for harvesting due to either organizational or technical reasons. The federation approach provides a single point of entry for users to submit their query, and internally it performs a distributed search to its member catalog systems, assembles the search results from these systems into a single result, and returns the result to the client. Advantage of the federation approach is that all are associated with no metadata harvesting; but the disadvantages are also obvious. These include relying heavily on network speed and reliability, overall performance heavily impacted by the performances of individual member catalogs, and low reliability.

For the federation approach, it is relatively easy to build a federation if the member catalogs all use the same standards for their metadata information model and interface protocol. Such a federation is called the homogeneous federation, within which the member catalog systems use the same interface protocol and information model; hence, no protocol translation and information model mapping are needed. However,

in world-wide consortia such as CEOS and GEO, the catalog protocols and information models in member organizations are diverse. In order to provide a one-stop-shopping capability for inventory search in such a consortium, a heterogeneous catalog federation within which the member catalog systems may use different interface protocols and information models, has to be built.

In a heterogeneous federation, the protocol translation and information model mapping have to be constructed to mediate heterogeneity among the member catalog systems. To realize the integrated search of heterogeneous, autonomous data catalogs, heterogeneous federations commonly use the mediator-wrapper architecture (MWA) shown in Figure 6-24 Catalogs of individual data sources provide search capabilities to users through the Internet. The wrapper on top of each catalog provides a standard-compliant universal query interface to the outside world by encapsulating heterogeneity of metadata information models, query protocols, and access methods. The mediator accepts a query from a user/client, distributes the query to relevant catalog systems of the individual data repositories through the corresponding wrappers, assembles the query results from individual data repositories, and returns assembled results to user/clients. The MWA is sometime called brokered architecture since the core function of the heterogeneous federation is to broker queries to the member catalog systems.

Currently, multiple heterogeneous catalog federations are being built with the mediator-wrapper architecture. The most significant ones are the CEOS WGISS Integrated Catalog (CWIC), the Federated Earth Observation catalog (FedEO) of European Space Agency (ESA), and the GEOS Data Access Broker (DAB) catalog. CWIC is being built by international partners of space agencies and coordinated by the CEOS Working Group on Information System and Services (WGISS). As of the end of March 2015, CWIC had eight space agencies as its partners with 80 million searchable data records (granules). The interfaces to users are both CSW/ISO 19115 and OpenSearch. FedEO is mainly an ESA initiative to build a heterogeneous catalog federation with member space agencies, mainly from Europe. FedEO provides an OpenSearch interface to its clients. As of the end of 2014, more than six million granules of EO data were searchable through FedEO. Since both FedEO and CWIC use the OpenSearch interface to outside world and internally to their member catalogs, both CWIC and FedEO treat each other as its member catalog system. Therefore, through CWIC clients, all FedEO records are searchable and discoverable, and vise-versa. The GEOSS DAB is another brokered federation, which provides one-stop shopping of discovery and access to data sources registered at GEOSS CSR, including both CSW and FedEO.

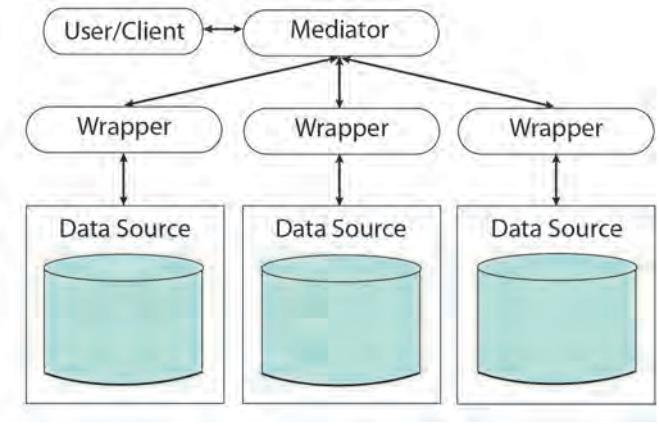


Figure 6-24. The Mediator-Wrapper architecture.

3.5 *Dark Data and Distributed Data Discovery Challenges for Remote Sensing Data*

With increasing numbers of Earth observing satellites and advances in remote sensing technology, the variety of remote sensing data products has exploded. While this is a boon to the science, applications, and education communities, the rise in available data has sharpened the need for effective data discovery. Through the early 1990's, much remote sensing data was in the hands of the scientists producing them. Data in this area were often discovered through publications about them, presentations at scientific meetings, or person-to-person networking. This began to change with the creation of institutional data centers, such as the Distributed Active Archive Centers (DAACs) making up the Earth Observing System Data and Information System (EOSDIS) (Asrar and Ramapriyan, 1995). These institutional data centers typically provide custom search tools for their archives customized for their communities. Improvement since the early days of data discovery, mainly in the distributed nature of data collections and the need to learn and become proficient with each data center's custom search tool have complicated the data discovery process. At the same time, increasing availability of affordable computing resources and increased use of remotely acquired data have enabled many more scientists to develop additional value-added data collections, further complicating the distributed discovery problem.

3.5.1 *Early Frameworks for Distributed Discovery*

In the early 1990's, NASA constructed an EOSDIS search capability that provided both data-collection and file-level search at its constituent DAACs. This was based on a custom socket-based protocol, enforced and enabled through a common library. The system was a joint development with the DAACs, each of which developed its own server-side search engine to respond to the EOSDIS client. The interface used a special-built socket protocol, which was implemented in a library provided to the DAACs for incorporation into their search engines. EOSDIS also used a message syntax and vocabulary that was developed jointly among the DAACs and the EOSDIS client developers (Yang and Johnson, 2000).

At about the same time, a cross-site search capability called the Wide Area Information Service, or WAIS (Kahle and Medlar, 1991) was developed, based on a communications protocol from the library science community named Z39.50 (Z39, 1988). Although this was primarily targeted at keyword searches of documents, it also provided fielded search capabilities. At one point, it was popular enough to be included as a supported protocol in the early versions of the Mosaic web browser. Ironically, the explosion of the World Wide Web itself eventually rendered WAIS obsolete: distributed search was unable to scale to the immense number of sites that burgeoned in the late 1990's as it requires that each site maintain a local search engine. Instead, the model of harvesting documents by crawling the web and providing a centralized (or apparently centralized) search capability became the dominant model, particularly with the rise of Google around the start of the century.

While general-purpose search engines such as Bing and Google have prevailed in text-based searches of web pages and documents, they were difficult to apply to the search for science data. Though they can be (and are) used for data collection searches, they work better with keywords than with spatial constraints (e.g., bounding boxes) or time ranges. As a result, work continued on special-purpose distributed search engines for Earth science data. The Open Geospatial Consortium (OGC) developed a distributed search mechanism named Catalogue Services for the Web (CSW). CSW uses HTTP as the transport mechanism

and encodes the search parameters in XML documents that are posted to the search engine. It offers primarily fielded search, including temporal and spatial constraints and offers different profiles such as ISO 19115 and ebRIM, ISO 19115 and even Z39.50.

3.5.2 *OpenSearch Distributed Discovery*

In 2003, Amazon created a subsidiary named A9 to specialize in search technology. Among the developments from A9 was the OpenSearch mechanism for executing a federated (distributed) search of multiple sites at once (McCallum, 2006). In 2005, A9 released OpenSearch 1.0, made available to the community at large under a Creative Commons license. OpenSearch is currently maintained by the community at the website http://www.opensearch.org.

Although described as a collection of simple formats for sharing search results, taken together, it constitutes a convention for clients to search multiple sites at once, without developing site-specific code for each search engine. The key to enabling client development is the OpenSearch Description Document (OSDD), an XML document that describes the capabilities of a given search engine. Figure 6-25 shows a simple OpenSearch Description Document.

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <os:OpenSearchDescription xmlns:os="http://a9.com/-/spec/opensearch/1.1/"
3 <xmlns:echo="http://www.echo.nasa.gov/esip">
4 <os:ShortName>NASA ECHO Dataset Search</os:ShortName>
5 <os:Description>Open Search compliant NASA ECHO Dataset search That responds in the ATOM
  format with Open Search, GeoRSS
6 Open Search Time and ECHO extensions.</os:Description>
7 <os:Tags>ESIP ECHO NASA</os:Tags>
8 <os:Contact>testbedops@echo.nasa.gov</os:Contact>
9 <os:Url type="application/atom+xml"
10 template="http://testbed.echo.nasa.gov:80/echo-
11 esip/search/dataset.atom?keyword={os:searchTerms?}&amp;
12 instrument={echo:instrument?}&amp;satellite={echo:satellite?}&amp;
13 cursor={os:startIndex?}&amp;numberOfResults={os:count?}>
14 </os:Url>
15 </os:OpenSearchDescription>
```

Figure 6-25. OpenSearch description document.

The OSDD lists the formats of responses it can support (in this case Atom on line 11), and, most crucially, includes a template describing how to execute a search (lines 12-14). Curly braces indicate sections where the search client is expected to substitute values. A question mark indicates that a parameter is optional. In this case, OpenSearch search terms are in the 'os' namespace (xmlns:os="http://a9.com/-/spec/opensearch/1.1/") and are understood as standard by OpenSearch adherents. Additional, optional parameters may be defined (in this case in the xmlns:echo="http://www.echo.nasa.gov/esip" namespace), but are not recognized standards and are perforce optional for search clients. In order for a client to invoke a search, the client need simply replace the curly braces and their contents, e.g., "{os.searchTerms}", with the client-supplied search term values, e.g. http://example1.gov/ /search/
dataset.atom?keyword=ozone

This formulation allows a client to interact with search engines responding to very different syntaxes, so long as the searches can be expressed as a URL; another search engine might have a URL template like: http://example2.gov/cgi-bin/search/collectionlist.pl?freetext={os:searchTerms}with the same search as the previous example would look like: http://example2.gov/cgi-bin/search/collectionlist.pl?freetext=ozone.

This formulation has limitations, particularly with respect to structured searches. However, given a choice between structured searches (which are often complex) and free text, users have, by and large, opted for the latter, likely due to the simplicity and ubiquity in the wider world web. A study of usage of the EOS Clearinghouse (ECHO) indicated that over 90 percent of users used free text searches.

The formulation is not only simple for users to utilize, but also simple for client developers. Indeed, it is simple enough that users with minimal knowledge of scripting languages are able to script machine-level queries to OpenSearch engines. This is a boon to many scientists now beginning to tackle much larger datasets than in the past, i.e., Big Data.

The returned response from an OpenSearch engine leverages several official and *de facto* standards. Several formats are possible, including Atom (Nottingham and Sayre, 2005), an XML format for syndication, shown in the example (Figure 6-26).

```
<?xml version="1.0" encoding="UTF-8"?>
<os:OpenSearchDescription xmlns:os="http://a9.com/-/spec/opensearch/1.1/"
  xmlns:echo="http://www.echo.nasa.gov/esip">
  <os:ShortName>NASA ECHO Dataset Search</os:ShortName>
  <os:Description>Open Search compliant NASA ECHO Dataset search That responds in the ATOM format
    with Open Search, GeoRSS, Open Search Time and ECHO extensions.</os:Description>
  <os:Tags>ESIP ECHO NASA</os:Tags>
  <os:Contact>testbedops@echo.nasa.gov</os:Contact>
  <os:Url type="application/atom+xml"
    templage="http://testbed.echo.nasa.gov:80/echo-esip/search/dataset.atom?keyword={os:searchTerms?}&amp;
    instrument={echo:instrument?}&amp;satellite={echo:satellite?}&amp;
    cursor={os:startIndex?}&amp;numberOfResults={os:count?}"
  ></os:Url>
</os:OpenSearchDescription>
```

Figure 6-26. Example of Atom XML format for syndication.

Within the Atom format, each result is represented as an <entry> item, with a unique <id> child node and one or more <link> child nodes. This allows the inclusion of links to different aspects of the returned <entry>, such as icons, metadata and data. The *rel* attribute in each link describes what role that link serves in describing or comprising the <entry>, based on the standard relation types standardized in the IANA relation registry as specified in the IETF RFC 5988 (Nottingham, 2010). Links also can include a *type* attribute that further describes what lies at the end of the URL in the <link> element. The type attribute is based on the IANA Media Type registry as provided for by RFC 6838 (Freed *et al*. 2013). This can provide clients with a rich description of complex entities.

3.5.3 *Application of OpenSearch to Remote Sensing Data Discovery*

However, searching for remote sensing data is not quite the same as searching for web pages and documents. At this point, it is helpful to consider the organization of remote sensing data. Typically such data are organized into data collections of related data files. While the web pages describing the data are text-

rich, individual data files tend to be text-poor and often stored in proprietary or API-based binary form. When textual descriptions at the file level are readily accessible, they tend to be nearly identical throughout the collection. Rather, the key distinguishing characteristic for data files within a data collection is the position in time and/or space of the data themselves. Space and particularly time are also applicable criteria for the data collection level of dataset discovery. Fortunately, members of the OpenSearch community had proposed draft extensions to OpenSearch for both time and space (location).

3.5.3.1 Draft Geo Extension

The draft of the OpenSearch Geo extension proposed conventions for specifying location in both the query and the response. For the query, the extension proposed several types of location specifications:

- *name* for named locations
- *lat* and *lon* for point locations
- *lat*, *lon* and *radius* for circular regions
- *box* (as lon, lat, lon, lat) for bounding box regions
- *geometry* (following the Well Known Text convention) for: POINT; LINESTRING; POLYGON; MULTIPOINT; MULTILINESTRING; MULTIPOLYGON.

The draft also proposes a convention for inclusion in the Atom response, namely to use the GeoRSS standard (e.g., Reed *et al*., 2006) for description. Note that this does lead to the unfortunate situation of expressing the bounding box as *lon,lat,lon,lat* for the query, but *lat,lon,lat,lon* in the response, a continual source of confusion for implementers.

3.5.3.2 Adapting and Adopting OpenSearch Conventions for Remote Sensing Data Discovery

Although OpenSearch and its draft Geo and Time extensions provide the basics for a data discovery framework, the conventions are still not well defined enough. A human could probably work within the basic OpenSearch conventions, but client applications tend to be less tolerant of small variations in behavior from one server to the next. Accordingly, a Data Discovery "Cluster" (informal working group) was formed in the Federation of Earth Science Information Partners (ESIP) to extend the convention to support Earth science data discovery.

The first major "extension" goes back to the difference between discovering data collections vs. data files. The main discriminator for data collections is the semantic content of the data, that is, what measurements it contains, and various constraints including the satellite, instrument and data creator, while the main discriminator for data files within a data collection boils down to time and space covered by each file. The nature of the data collection search is that it usually has a low precision. For example, to take a common search term such as "ozone", the EOSDIS system contains several hundred datasets related to "ozone". Yet, researchers usually work with only one or a few datasets at most; they simply cannot "scale up" to use all of them, or even most of them. Thus search precision at the data collection level for "ozone" is thus on the order of a few percent. On the other hand, when scientists specify a spatio-temporal region for data search, they often expect to use most or all of the files available for that region in the datasets they are working with. This number can be enormous: there are over 1 million files for *each* dataset of 6-minute satellite scene from the Atmospheric Infrared Sounder (AIRS) for example. If one tried to search for all of the data files in combination with datasets in a single search, the low precision of the dataset search results in an

overall low precision for the search. That is, useless data files are returned for each dataset that a user may not be interested in using.

In order to solve this conundrum, most remote sensing data search interfaces break the search into two steps. In the first step, the user searches for data based on semantic information, plus space and time coverage. The user then selects specific datasets to perform the largely spatio-temporal search for data files.

As it happens, the basic OpenSearch convention included a search recursion mechanism that can support this two-step search (data collections and then data files). Specifically, the convention allows for a link with a *type* of application/opensearchdescription+xml**.** This allows a search recursion of the type shown in Figure 6-27.

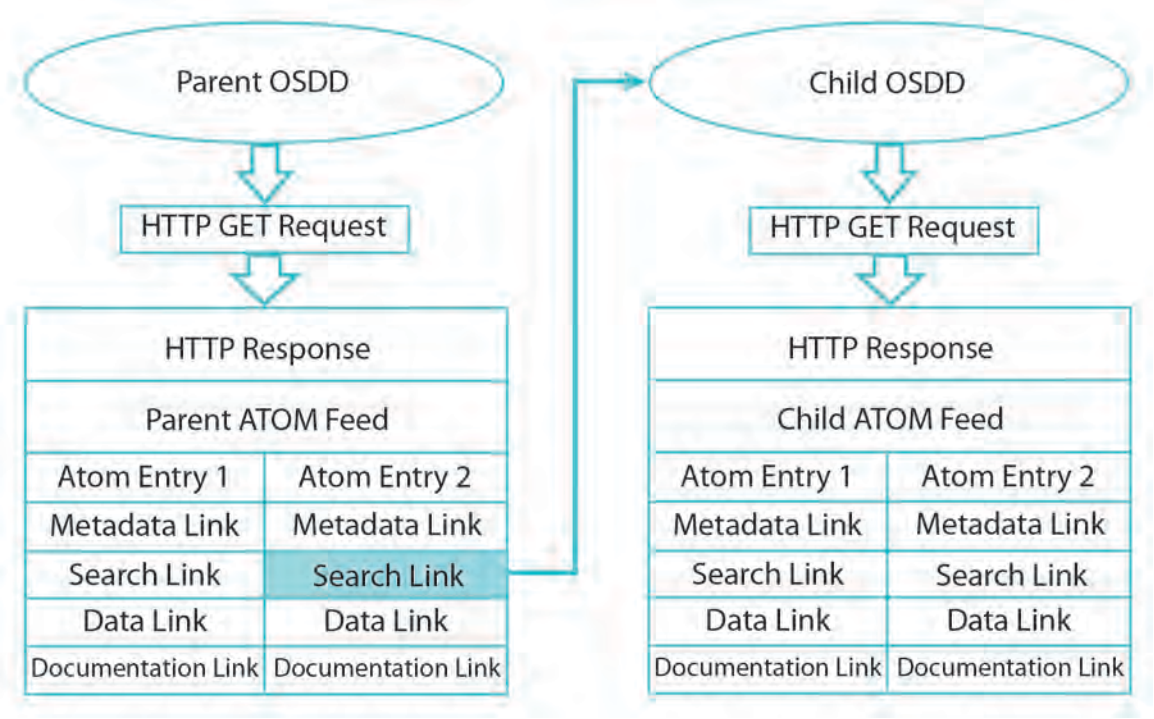


Figure 6-27. Executing a two-step recursive search in OpenSearch.

The two-step search scenario begins with the acquisition of the top level Open Search Description Document (OSDD) and searching based on whatever criteria. Search engines that offer the two-step search would include an Atom <entry> with a <link> element identified as type=”application/opensearchdescription+xml**”,** which points to a child OpenSearch Description Document. The client then fetches the child OSDD and repeats the search process.

In applying this to data collections vs. data files, the first search of the archive returns data collection entries, each entry of which includes a link to an OpenSearch Description Document for searching that particular data collection for constituent data files. This also resolves another difficulty with the data file search. Most data centers and search providers serving many remote sensing data collections provide a file-level search that involves specifying which data collection to search. However, the syntax for expressing the data collection constraint varies widely from data center to data center. The recursive search allows the search provider to return a child OSDD whose search “template” already includes the hard-coded data

collection syntax and identifier. For example, one search provider might return a child OSDD with a URL template beginning:

<os:Urltemplate="http://mirador.gsfc.nasa.gov/cgi-bin/mirador/granlist.pl?
dataSet=AIRX2RET.006&osLocation={geo:box?}&endTime={time:end}and
startTime={time:start}/>

whereas another might return

<os:Url template=”https://api.echo.nasa.gov:443/opensearch/granules.atom?
shortName=AIRX2RET&versionId=006&boundingBox={geo:box?}&
startTime={time:start?}&endTime={time:end?}/>

Despite the use of quite different syntaxes for specifying which data collection to search for files, the client software can execute the query for granules in the AIRX2RET v. 6 dataset without any code or foreknowledge of the difference in data collection identifiers in either the syntax or the underlying database of each search engine. All the client need do is to replace the placeholders with actual search values.

Thus the basic algorithm for conducting a recursive search against a server is as follows:

- Obtain the root OSDD for the search engine;
- Replace the URL template’s placeholders with the user’s search criteria;
- Fetch the resulting OpenSearch URL;
- Parse the Atom response for data collection entries;
- (Obtain the user’s selections from these data collections);
- For each selected data collection entry:
    - Locate the <link> with *rel*=”search” and *type*=“application/opensearchdescription+xml”;
    - Dereference the link’s URL to obtain the data collection level OSDD;
    - Replace the URL template’s placeholders with the user’s search criteria;
    - Fetch the resulting OpenSearch URL;
    - Parse the Atom response for data collection entries.

Although OpenSearch’s recursive search is ideal for solving the common data-collection-followed-by-data-file search model used by many search providers for remote sensing data, the recursion is not limited to this two-step hierarchical case. Another possibility is a multi-step search, looking first for search providers, and then executing the data collection search based on a search provider selected by the client or user. The use of a simple OSDD file to represent the syntax and semantics of search also opens other possibilities, such as the publication of data collections via OpenSearch-response-style Atom files, which in turn provides the URL template for further searching for data files, sometimes called “collection casting”. Links to OSDDs can also be easily embedded in web pages.

Note that the use of recursive search to solve the data-collection vs. data-file dilemma required no change to the OpenSearch conventions per se, simply an informal agreement by search providers to provide OpenSearch Description Documents for searching individual data collections.

3.5.3.3 Extending the OpenSearch Draft Time Extension

In contrast to adopting the recursive search practice, there were some areas in OpenSearch that were insufficiently expressive for robust usage in searching Remote Sensing data. Chief among these was the

lack of a common understanding of how to describe a data start and stop time in the Atom response. These can be critical parameters in deciding which data collections or data files are usable in a given context. Initially, the ESIP Discovery Cluster promoted the use of XML elements that closely mirrored the data search terms, namely <time:start> and <time:end>, where "time" refers to the OpenSearch Draft Time extension namespace. However, it was pointed out that the Dublin Core had a long-established standard for Date, which follows the ISO 8601 as a best practice. This was later adopted by the ESIP Discovery Cluster, though the lack of symmetry in the query vs. the response is unsatisfying.

3.5.3.4 Describing Link Types

The key for any client to taking action on a link returned in the search response is an accurate understanding of what the link represents. Accordingly, the ESIP Discovery Cluster defined the most common link types for data and related objects, as represented in the combination of the *rel* and *type* attributes shown in Table 6-6.

Table 6-6. Recognizing Link Types by *rel* and *mime-type* Values.

Type of Link	Definition	*rel* value	*mime-type* (type value)
data	link representing a data file or other science data resource; may be large in size	enclosure	application/x-netcdf, application/x-hdf, text/csv
browse	image of the data typically used for making data request decisions	icon	image/jpeg image/png image/pdf image/gif
metadata	file with (usually) structured information about corresponding data files	describedBy	text/xml
OSDID	link to an OpenSearch Description Document; useful for recursive searching	search	application/ opensearchdescription+xml

Mapping link types relevant to remote sensing data to the standard *rel* and *mime-type* values used for the diversity of Internet resources is not completely straightforward. For example, "browse" is often quite large, larger than typically signified by the term "icon"; however, as a visual representation of the item, it is the best candidate, short of inventing a new term. Likewise, a metadata link might be equally described by the "via" relationship, which was specified in the Open Geospatial Consortium (OGC) version of the OpenSearch standard.

3.5.3.5 OpenSearch Conventions and Best Practices in Use for Remote Sensing Data Discovery

Contemporaneously with the ESIP development of OpenSearch for data discovery, other organizations were pursuing similar directions. The Open Geospatial Consortium (OGC) published a standard in 2014 entitled "OGC® OpenSearch Geo and Time Extensions". This standard has many aspects in common with the ESIP OpenSearch standard, such as the use of Atom as a response format and the adoption of the OpenSearch Time and Geo proposed extensions. However, it does not codify the recursive (or two-step) search aspect for dataset and file-level searches. In addition, several applications utilizing OpenSearch for real-life use cases began to expose areas where additional consistency and precision would be desirable to

simplify the client development process. These include the European Commission's GENESI-DEC (Ground European Network for Earth Science Interoperations - Digital Earth Communities), and the CEOS Water Portal, led by the Japanese Aerospace Exploration Agency. Accordingly, an internationally-staffed project was formed with CEOS to develop a Best Practices document in order to recommend conventions to increase consistency across search engines. As of 2014, 15 Best Practices have been drafted, covering:

- Two-Step Search
- Use of Dublin Core's dc:identifier
- Support of OpenSearch Parameter Extension
- Specification of the *rel* attribute in the OSDD
- Supported search parameters
- Multi-words for searchTerms
- Use of startPage over startIndex
- Search with geo:name
- Output encoding format in search URL
- Output encoding format support
- Metadata representation in search response
- atom:summary
- GeoRSS
- Browse Image
- Data access

As search engines come into compliance with these practices, client development will become easier, producing more robust and consistent behavior across search engines.

3.5.3.6 Client Usage of OpenSearch

OpenSearch-enabled clients run the full range from extremely simple to complex distributed systems. On the simple end of the scale, most browsers themselves can act as clients, because the OpenSearch specification is used to describe search engines for the browsers that support customization of which search engine to use. Unfortunately, these do not support the Geo and Time extensions, so they are useful primarily for data collection search. It is also possible to create an OpenSearch client using simply Cascading Style Sheets (CSS). By parsing the URL template, a simple web form can be constructed to execute a space-time search of data collections and or data files.

Another simple application for OpenSearch is data search and access scripts written by scientists who are looking for scripted access to the data centers. For some users, existing search tools do not provide the precise set of selection criteria to find the particular data for their study areas. Alternatively, they may wish to run a script to download the most recent data for a given data collection and other space-time criteria.

At the other end of the spectrum is the Giovanni application for online visualization and analysis of science data (Lynnes *et al.*, 2013). Although developed at the Goddard Earth Sciences Data and Information Services Center (GES DISC), the system also provides services for data hosted at other data centers, particularly within EOSDIS. In order to search and acquire data from the distributed community, Giovanni

relies on an application called the Simple Subset Wizard, which in turn uses OpenSearch to query remote data centers and ECHO.

3.5.3.7 Future Directions of OpenSearch for Remote Sensing Data Discovery

Despite the successful adoption of OpenSearch in both the client and server realms, the community continues to work on improving the convention to be more useful for federated data discovery. One area, ironically, lies in the free-text nature of the search, the very bedrock of the OpenSearch convention. Although this is very easy for users to understand and use, the dataset-level results are often voluminous and difficult to weed through. Depending on the community in question, three basic approaches are either under discussion or in active development. The first approach is to leverage the support for Facets that is found in many off-the-shelf search engines these days, such as the Apache SOLR search engine. This allows users to narrow results by checking off which values of dataset attributes are desirable, filtering the results accordingly. In a similar vein, these attributes might be added to the search itself, using the OpenSearch Parameter Extension to advertise which attributes are supported and how they are to be queried.

An alternate approach is to improve relevancy ranking of the results, hopefully obviating the need for a user to learn the vocabulary of either the search engine or the client. In this scenario, the search engine would attempt to do a good enough job in relevancy ranking that the user could expect to see the optimal dataset results show up at the top of the list. Here, the differences between data and document searching could be exploited. For instance, the context in which a keyword appears can be leveraged. Consider a user typing in only “ozone” in the search blank. One might expect the user to be more interested in data collections where ozone appears in the description of the measurements, as opposed to, say, sulfur dioxide retrieved from the “Ozone Monitoring Instrument”. Another data-specific ranking criterion would be to leverage how completely the data time range covers user’s search box. Even some basic user intent modeling may be possible by identifying referral web pages or by looking at how much jargon is used in the search terms.

3.6 *The Open Data Movement*

There are many good reasons for keeping data and information secret or proprietary. Military and intelligence services must protect national security by keeping some information secret. Government officials and business managers need to maintain some decision-making processes and advice confidential. Businesses broadly protect their private investments in data and information through intellectual property laws and business methods. The privacy of individuals needs to be maintained and there are niche concerns such as protecting the location of cultural artifacts and endangered species, or the rights of indigenous peoples.

All of these interests and concerns are protected legally by international treaties and executive agreements and at the national level by legislation, regulation, and public policy. Restrictions on information access and use may even extend to state and local levels.

At the same time, there are equally good reasons to provide data and information freely and openly, especially if they are created by government and posted on digital networks. This section of the chapter looks specifically at publicly funded Earth Observations (EO) data from the perspective of “open data” and

traces some key developments in that area in recent decades, first in the United States and then internationally.

3.6.1 *Open Data for Government Earth Observation Systems*

Chapter 11 of the Manual of Remote Sensing fourth edition, describes in more detail the United States Section 105 of the 1976 Copyright Act. It expressly waives copyright protection of any work created by the U.S. Government (USG). This means that any information created in the course of government business is in the public domain and is free to use once lawfully accessed or expressly protected by another law. The Paperwork Reduction Act of 1989, and its amendments and subsequent implementations by the White House Office of Management and Budget (OMB) through Circular A-130 beginning in 1985 added rationales and implementation details for the "Management of Federal Information Resources," generally; and specifically to the policy governing federal public domain information, (Office of Management and Budget 2000).

Another important point is that international copyright law and its implementation in national law does not protect the facts or values—the data themselves—contained in a database, unless the selection or arrangement of the data is creative or original. Although there are other intellectual property laws that may protect databases in the United States. to some extent, even in the private sector, such protection is considered "thin" and not broadly applicable to all the contents of a database.

With regard to EO satellite and other sensor data collected by the USG, all such data and resulting higher levels of information that are created as official federal government business by government employees are in the public domain. The open policy for public EO data was recently reinforced by the White House *National Strategy for Civil Earth Observations* (NSTC, 2013).

The open data policy for civil EO data and other USG information is justified from many different perspectives. No intellectual property incentives or protections are needed for activities already funded by taxpayers. Open data stimulate the economy, provide information for value-added firms, support broad social uses by the citizens, provide the factual material for improved research and education, and generally enhance transparency in governance and better decision-making (CODATA, 2015).

Other nations now provide open data from many public government activities and even from some EO systems, although they are frequently protected by intellectual property laws and licensed for a fee, with restrictions on (re)use. Although a discussion of the laws of each country is beyond the scope of this discussion, there have been some notable developments regarding EO data access and use in several intergovernmental organizations.

Most prominent among these has been the Group on Earth Observations (GEO). GEO currently has 103 government Members and 87 non-governmental Participating Organizations, which together represent all the major EO data collection systems in the world (https://www.earthobservations.org/index.php).

It is useful to see the evolution of GEO's original Data Sharing Principles to those proposed for adoption in the 2015 Ministerial Plenary. That progress toward greater openness reflects the growing understanding of governments worldwide in the benefits and value of open public EO data, especially on digital networks. The Principles in 2005 were as follows (Group on Earth Observations, 2005):

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

The draft new Data Sharing Principles advocate a more open position for data made available through GEO (DSWG, 2014):

- Data, metadata and products will be shared through GEOSS as Open Data by default, by making them available as part of the GEOSS Data-CORE without charge and without restrictions on reuse, subject to the conditions of registration and attribution when the data are reused.
- Where international instruments, national policies, or legislation preclude sharing data as Open Data they should be made available through GEOSS with minimal restrictions on use and at no more than the cost of reproduction and distribution.
- All shared data, products and metadata will be made available through GEOSS with minimum time delay.

Finally, it is also worth noting the "data democracy" principles developed by the Committee on Earth Observation Satellites—CEOS, another inter-governmental organization of EO satellite operators founded in 1984 to help coordinate the space segments and various technical and management concerns. According to CEOS, there are there are "four pillars" of the data democracy initiative in the Earth observation area:

- Providing wider and easier access to Earth Observation data.
- Increasing the sharing of software tools such as the use of open source software and open system interfaces.
- Increasing data dissemination capabilities, transferring relevant technologies to end users, and discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work." In considering the application of copyright law to data in databases in 1991, the U.S. Supreme Court exclude all factual data from such protection "unless the selection and arrangement of the data are creative and original."
- Providing intensive capacity building, education, and training (including awareness and outreach) for enabling end users to gather the information they need and for increasing communication on achieved results.
- (http://www.ceos.org/index.php?option=com_content&view=category&layout=blog&id=104&Itemid=153).

3.6.2 *Open EO Data in the Academic Sector*

Academics have also been at the forefront of adopting open access policies and practices in many forms, including in their use and re-dissemination of EO data, especially for research purposes. Perhaps the earliest example of an open data policy was articulated by the White House Advisor for Science and Technology (D. Allen Bromley) in 1991. Known as the "Bromley Principles" Regarding the Full and Open Access to

"Global Change" Data, the key provisions regarding "full and open access," that have been cited and emulated by many other data policies since, included:

1. Full and open sharing of the full suite of global data sets for all global change researchers is a fundamental objective.
2. Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost, as a first principle, should be no more than fulfilling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers.

For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as they become widely useful. In each case, the funding agency should define explicitly the duration of any exclusive use period (Bromley, 1991).

Principles, statements, and declarations have been made by the academic community for over the past quarter century, covering research data across all disciplines. These have included, e.g., the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Berlin Declaration, 2003); the Panton Principles for Open Data in Science (Murray-Rust *et al*., 2010); the Nairobi Principles on Data Sharing for Science and Development in Developing Countries (CODATA, 2014); the Liber Statement on Enabling Open Science; and the Policy Recommendations for Open Access to Research Data in Europe (RECODE Policy Recommendations, 2015).

A list of many other such statements, covering different research disciplines and contexts—including especially open access to the online, publicly-funded research literature—is available from the Open Access Directory. They mostly call for open access to publicly funded research data with few or no restrictions on reuse and re-dissemination. When combined with the policies and practices of government EO ministries and agencies, as well as public research funding organizations, the trend in the Internet era has been to make public EO data more openly available and reusable.

4 VISUALIZE AND ACCESS

4.1 *The Australian Geoscience Data Cube*

There has been an explosive growth in the volume, variety and velocity of geospatial and geoscientific data sources, along with a greater understanding of the societal benefits that can be derived from the information extracted from these data (Oxera Consulting Ltd., 2013). Un-like other types of Big Data, which are often comprised of vast amounts of small pieces of semi- or unstructured data, these geoscientific data are often highly structured and stored as large, to very large, binary data files. In short, data users and application developers face huge challenges over the next 15 years to realize the value of current Earth observation and geoscientific data archives while preparing for the upcoming deluge of highly structured Big Data. There is a clear need to shift from sensor specific data formats to standardized high performance data formats suitable for analysis within high performance computing environments.

Utilizing high performance data architectures and high performance computing facilities, Geoscience Australia has developed a data infrastructure that is helping to address these 'big data' challenges. The vision for the Australian Geoscience Data Cube (AGDC) is to establish a *common analytical framework* for analysing and modeling the geophysical properties of Earth, in multiple dimensions (x,y,z,t *and* λ). The AGDC enables interoperability of very large data in a high performance computing environment by adopting common data structures suited to large-scale computation. Although in its nascent stages, the Australian Geoscience Data Cube (AGDC) is already allowing analysis of Earth observation data almost 10,000 times faster than was previously possible. The AGDC presents an opportunity to test and implement international standards for Discrete Global Grid Systems being developed by the Open Geospatial Consortium.

4.1.1 *Addressing the Big Data Challenge*

The Earth and its systems are complex and interconnected. Gaining a comprehensive understanding of these systems, and how we interact with them, is of critical national and international interest, but poses an enormous technological challenge (e.g. Neill and Hewson, 2013; Coleman, 2013; Hart and Saunders, 1997; Boyd and Crawford, 2012; and the Australian Public Service Big Data Strategy (Australian Government Information Management Office, 2013). As our knowledge and understanding of the Earth increases so do the volume and variety of the data, and data sources, used to build that understanding; and, in particular, Earth Observation (EO) data. We have moved into the world of 'Big Data'.

This section describes the new data structures and analytical approaches being developed to more effectively derive benefit from existing EO data archives; and to prepare for the influx of many more new, and probably even larger EO data streams. As fundamental background, one must review the established and new image processing, spatial positioning and quality control methods that underpin AGDC. Examples of early applications of the AGDC to land management problems illustrate this point. Looking forward to the eventual development of a global network of interoperable, 'data cube' one can expect hubs supported and linked by open standards and Discrete Global Grid Systems.

4.1.1.1 The Deluge of Big Data

'Big Data' is defined by the Oxford Dictionary as *"Data sets that are too large and complex to manipulate or interrogate with standard methods or tools"*. The term was first described by Laney (2001) as the three V's: Volume, Velocity and Variety. Since then many others have extended this list to include additional words such as Validity, Veracity, Value, and Visibility, among others.

Often when we think of the term 'Big Data', we conjure up images of the massive databases of financial transactions, health and government records, or of crowd sourcing and data mining from social media to gain insights that can be utilized by advertising companies that deliver targeted marketing to individuals. However, there is another type of 'Big Data'; one that is truly massive, diverse and growing at an exponential rate (Overpeck, *et.al.*, 2011). These data represent our Earth and unlike other types of 'Big Data', that are often comprised of vast amounts of small pieces of semi- or unstructured data (e.g. Twitter feeds - bytes to kilobytes of data per record), these geoscientific data are often highly structured and stored as large to very large binary data files (e.g. Satellite Earth Observation data - gigabytes to terabytes of data per file).

Of the many sensors that are used to acquire information about the Earth and its systems, the constellations of Earth Observation Satellites (EOS) are the single most important sources of information; as

highlighted in the Australian Strategic Plan for Earth Observations from Space (Australian Academy of Science, 2009) and Australia's Space Utilisation Policy (The Commonwealth of Australia, 2013). These satellites are orbiting the Earth, routinely and repeatedly, to observe diverse physical phenomena over the entire globe. These phenomena include crustal deformation, soil mineralogy, vegetation and surface water conditions, sea surface temperature, magnetic and gravimetric field intensities, to name a few. Of these constellations, the Landsat Program, built by NASA and operated by the United States Geological Survey (USGS), represents the longest, most continuous, openly available Earth observation program in the world (Arvidson *et al.*, 2001). The Landsat constellation has acquired moderate resolution multi-spectral data since the launch of Landsat-1 in 1972, and is still acquiring data with the current Landsat-8 Mission. In February 2013, the Landsat-5 satellite was awarded the Guinness World Record for the 'Longest-Operating Earth Observation Satellite' – outliving its original 3-year design life, it orbited the Earth over 150,000 times and acquired more than 2.5 million images of the Earth's surface in the 29 years it operated from its launch in 1984 until its decommissioning on 05 June, 2013.

The longevity of existing satellite missions, coupled with the increasing number of new EOS missions (each acquiring more and more data) pose a great challenge for many government agencies and other organizations concerned with curating and making available the vast and rapidly increasing volumes of EOS data (Australian Academy of Science, 2009). The number of these satellites has grown from 12 in 1980 to over 69 operational EOS missions today, with an additional 137 new missions being planned for launch in the next 15 years (CEOS and ESA, 2014). If we only consider a handful of these current and new EOS missions (Landsat-8, Sentinel-1,-2, -3, and Himawari-8/9) we can expect at least a 20x increase in the global volumes of 'raw' data acquired over this period – producing a global archive of 'raw' data of at least 20 Petabytes of data (Figure 6-28). If we also include the various processed data products derived from this 'raw' data the expected data volumes will be at least 3-5 times greater – this is consistent with the assessment of Overpeck *et al.* (2011).

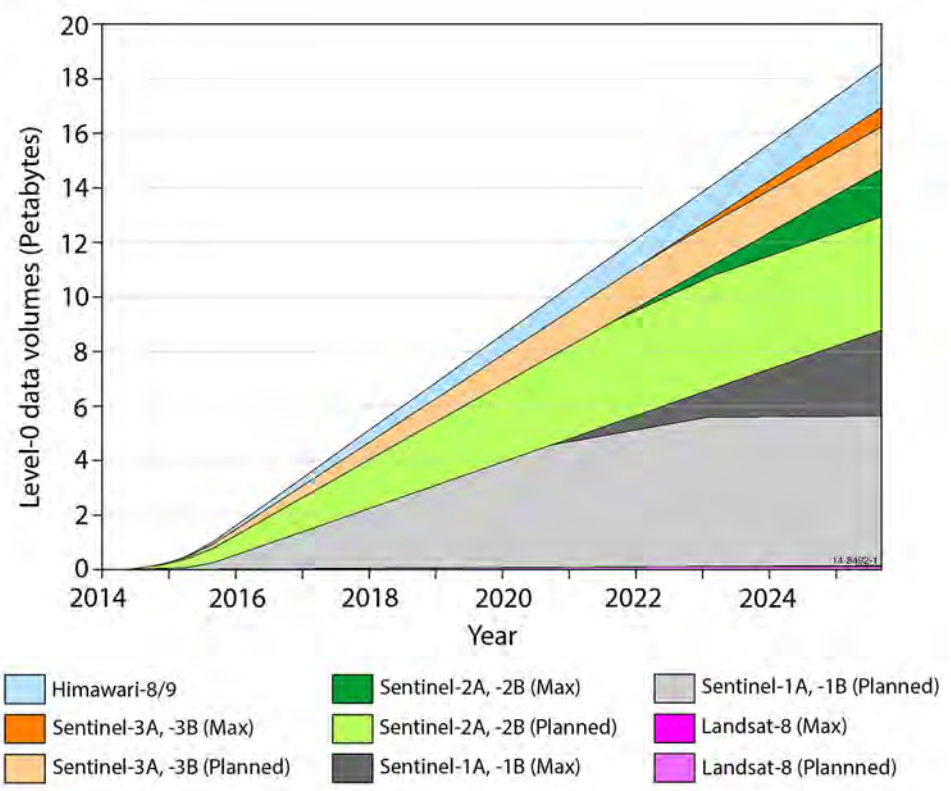


Figure 6-28. Estimated growth of global raw (Level-0) EOS data acquired from the Landsat 8, Sentinel-1, -2, -3 and Himawari 8/9 missions from 2014-2015.

In short, data user communities face a huge challenge over the next 15 years to realize the value of the current EOS and geoscientific data archives, and to prepare for the upcoming deluge of 'Big Data' there is a clear need to shift from sensor specific data formats to high performance data (HPD) formats suitable for analysis within high performance computing environments (Mattmann, 2013). The deluge has already begun!

4.1.1.2 EOS Data are More Than 'Photographs'

Although many Earth Observing Satellites include optical sensors, along with infrared and short-wave infrared sensors that produce images that look like 'photographs' taken from space, these sensors actually acquire multiple geophysical measurements of the electromagnetic spectrum from solar radiation reflected or emitted from the surface of the Earth. Consequently, these satellites provide a wealth of information that, through statistical and scientific analyses, allow one to identify features and trends in the data that are not discernable with the 'naked eye' (Figure 6-29 and Figure 6-30).

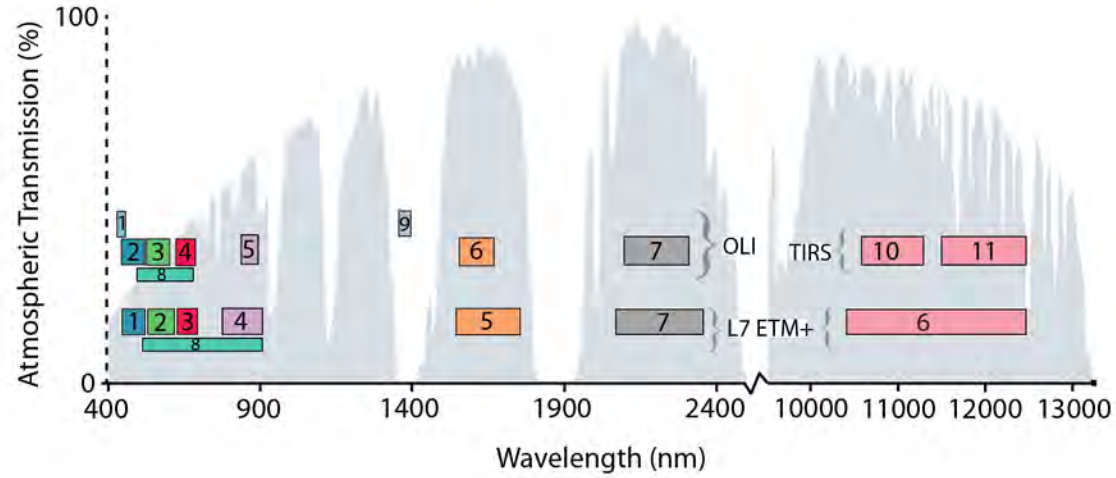


Figure 6-29. Band passes of the Landsat-8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) instruments compared with the Landsat-7 Enhanced Thematic Mapper Plus (ETM+) Sensor (USGS, 2013).

Figure 6-30. Animation showing Landsat-8 spectral bands (with associated wavelengths [μm]) and how they are combined to create True Color and False Color images of the data.

What's in a scene? There are two main types of Earth Observation satellites based on their orbits:

- Geostationary satellites (e.g. Himarwari-8, GOES, GOES-R etc.), and,
- Polar Orbiting Satellites (e.g. Landsat and Sentinel 1-3)

Geostationary satellites orbit the Earth at relatively high altitudes (e.g. 35,800*km*) above the equator and they are positioned so that their orbital speed matches the rotation of the Earth. This enables them to observe the full hemisphere (also referred to as a full disk) of the Earth in one image. These types of satellites are mainly focused on weather and climate applications and while they tend to monitor the same region of the Earth at only a medium to coarse spatial resolution, they do so at a relatively frequent time interval (e.g. 10-30 min per full disk image).

Polar Orbital Satellites orbit the Earth around the poles at much lower altitudes than geostationary satellites (e.g. around 700-800*km*). Consequently, they observe a much narrower field of view than geostationary satellites. The Earth constantly rotates underneath the path of the satellite resulting in strips of data being recorded by the satellite; this can be thought of in terms of film strips of data (Figure 6-31). The width of this strip (or swath) will vary depending on the particular sensor and satellite configuration (e.g. Landsat-8

= 185*km*, Sentinel-2 = 290*km*). One orbit typically takes about 99 minutes resulting in approximately 15 orbits within a 24 hour period. The orbit is maintained such that the entire surface of the Earth is observed by the satellite within a given period of time; for Landsat satellites this is every 16 days.

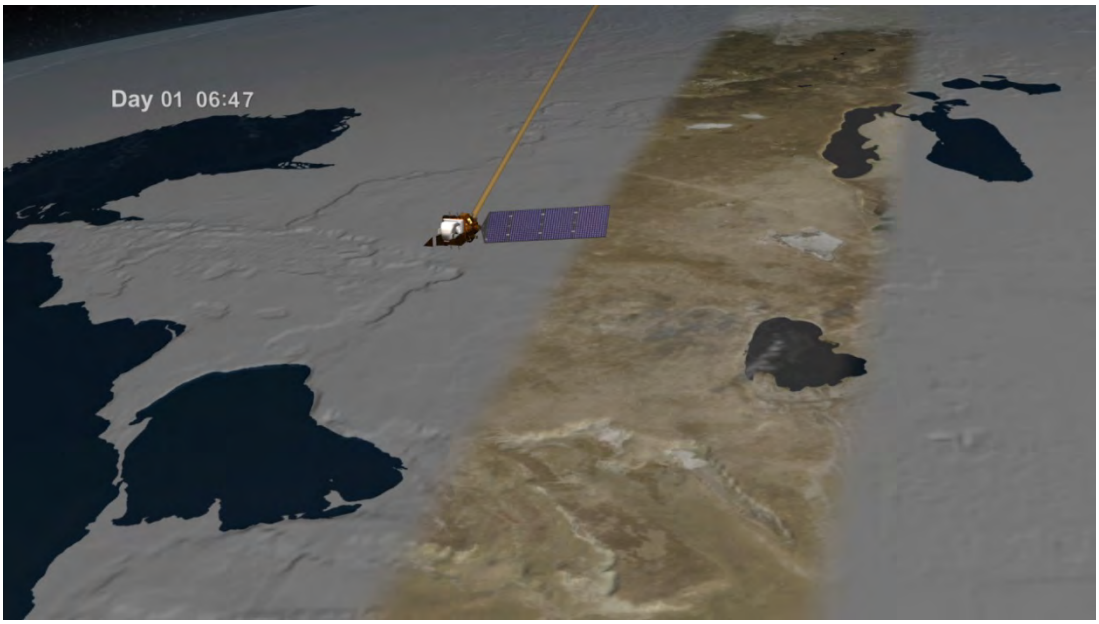


Figure 6-31. Visualization of the orbit of Landsat-8 by NASA's Goddard Space Flight Center.

Each orbital track of a polar orbiting satellite is termed a pass (or sometimes path) and depending on the application focus of the particular satellite mission the sensors will be turned on and off to preserve and extend the life of the satellite (e.g. Landsat mission satellites typically only acquire data over land). Historically the data volumes acquired in a satellite pass have been too large to process and analyze in their entirety. Consequently, there have been a number of schemes developed to break the data up into smaller, more manageable bits; much like the frames of a film strip. These 'frames' are usually referred to as 'scenes' of data.

The Worldwide Reference System (NASA, 1998) is a global notation system for Landsat data. It enables a user to identify and reference satellite imagery over any portion of the Earth, between 81° North and South of the equator, by specifying a nominal scene center designated by 'Path' and 'Row' numbers (e.g. a scene with a Path-Row identifier of 127-043 relates to Path number 127 and Row number 043). The framing is uniform for each orbit with adjacent East-West scenes having scene centers at the same nominal latitude (Figure 6-32).

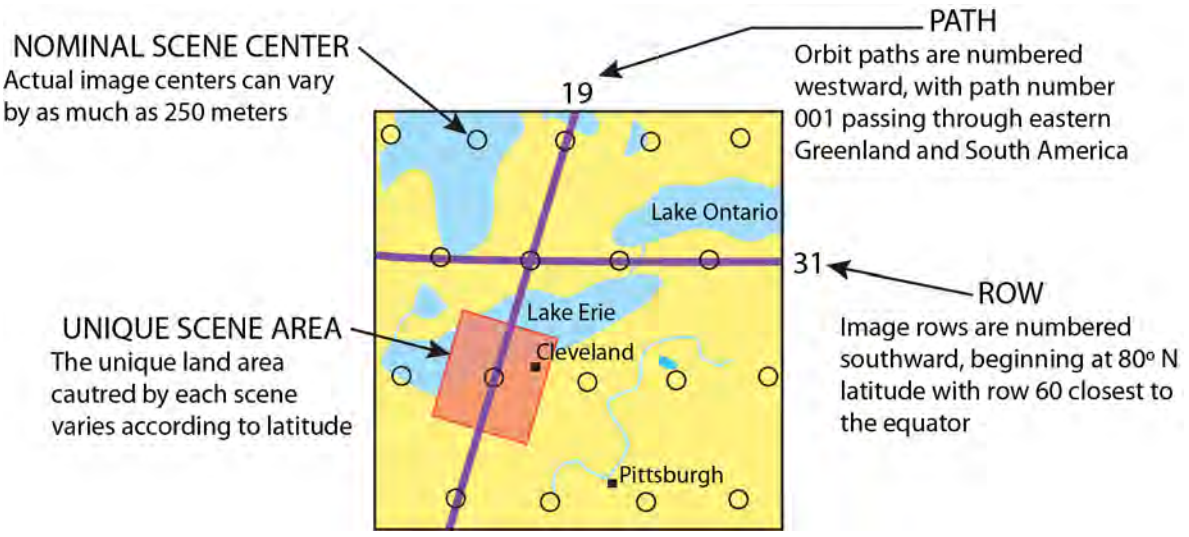


Figure 6-32. Worldwide Reference System Path/Row numbering scheme from the Landsat-7 Science Data Users Handbook (NASA, 1998).

4.1.1.3 Traditional Approaches to Data – Drivers for a 'Data Cube'

Traditional "On-Request" data delivery models where data are only processed and prepared when a user requests them, creates a significant barrier to entry for those businesses, agencies or researchers thinking of a new initiative or application of the data. The cost involved in the data preparation process can be prohibitive and often results in the data being discarded because:

- Too many of the data requested clouds or contain other undesirable artifacts. Using traditional data storage models it can be difficult and time consuming to identify suitable observations prior to retrieving, processing, downloading and interrogating the data;
- The dataset(s) requested are often too large to process, so initiatives are scaled back to focus on smaller areas or smaller time periods;
- Data from just one sensor are often requested because it is too difficult for a particular group to justify the technical work required to make data from different sensors comparable; or,
- Data aggregation approaches, such as 'mosaicking', are applied when data are merged from multiple satellite over-passes to produce best observed pixel products that are more manageable and usable on commodity ICT infrastructure. However, this can result in a significant loss of valid observations and degrades the quality of temporal change detection analyses using these aggregate snapshots of the data record. Such approaches often result in products that only provide a snapshot perspective of the data holdings, losing the richness of the time-series. Frequently, the outcome does not meet the important challenges of providing dynamic change detection information to business and government. Furthermore, the data are often stored on hierarchical tape archive infrastructures. While this ensures that data are preserved robustly and economically for the long term, the data retrieval latencies associated with bringing these off-line data on-line so they can be utilized does not support the increasing expectation of value added information products on-demand and in near real-time.

Australia's Earth Observation Program has downlinked and archived satellite data for the Australian Government since the establishment of the Australian Landsat Station in 1979. Geoscience Australia

maintains this archive and produces image and other value added products to aid the delivery of government policy objectives. The entire archive of raw and processed EOS data held by Geoscience Australia is now approaching a petabyte. The archive of calibrated Landsat data products currently comprises close to 50 per cent of the total data volume (excluding ALOS data holdings which have restricted usage and distribution license conditions) and continues to grow (Figure 6-33).

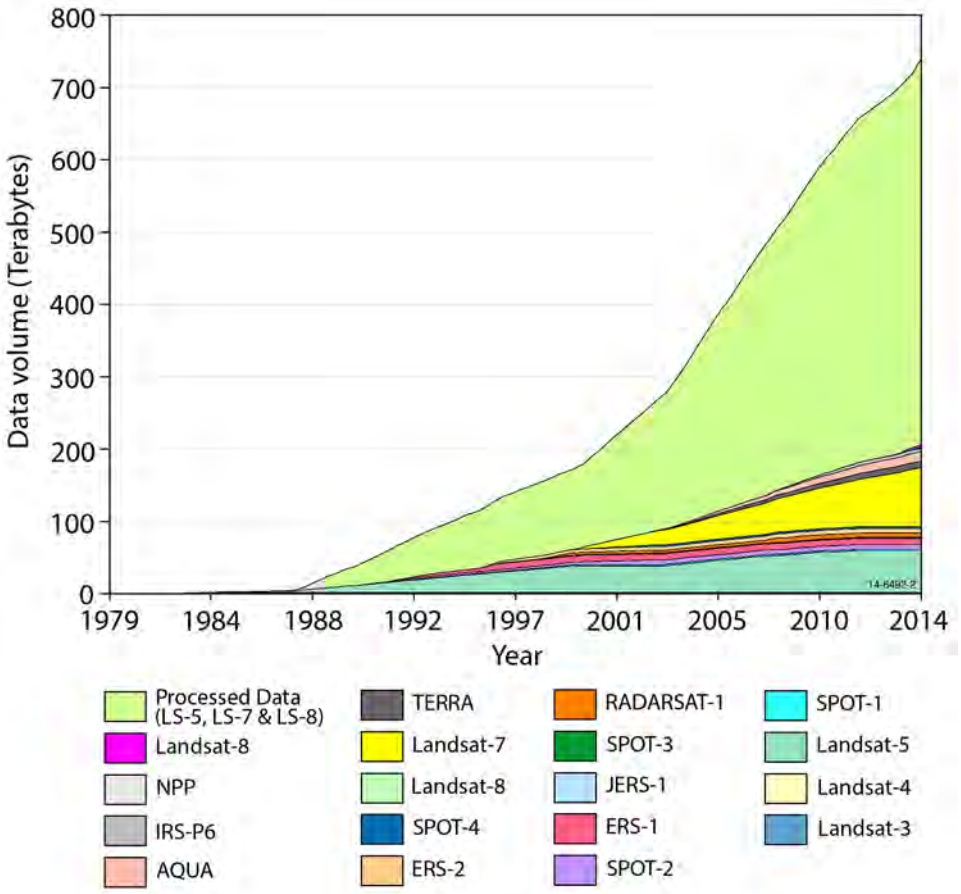


Figure 6-33. Data holdings of EOS data (Level 0 + processed Landsat data) held by Geoscience Australia (excluding data from ALOS which have restricted usage and distribution license conditions).

Looking to the future, Geoscience Australia faces an enormous challenge to both maintain existing capacity to provide geoscientific advice to support stakeholders and to enhance our capability to effectively deal with the impacts of the new EOS Missions that will be launched over the next decade. The enormity of this challenge is demonstrated by considering the expected increase in 'raw' (or Level 0) data volumes over the Australian continent from Landsat, Sentinel and Himawari EOS missions shown in Figure 6-28 and Figure 6-34, below.

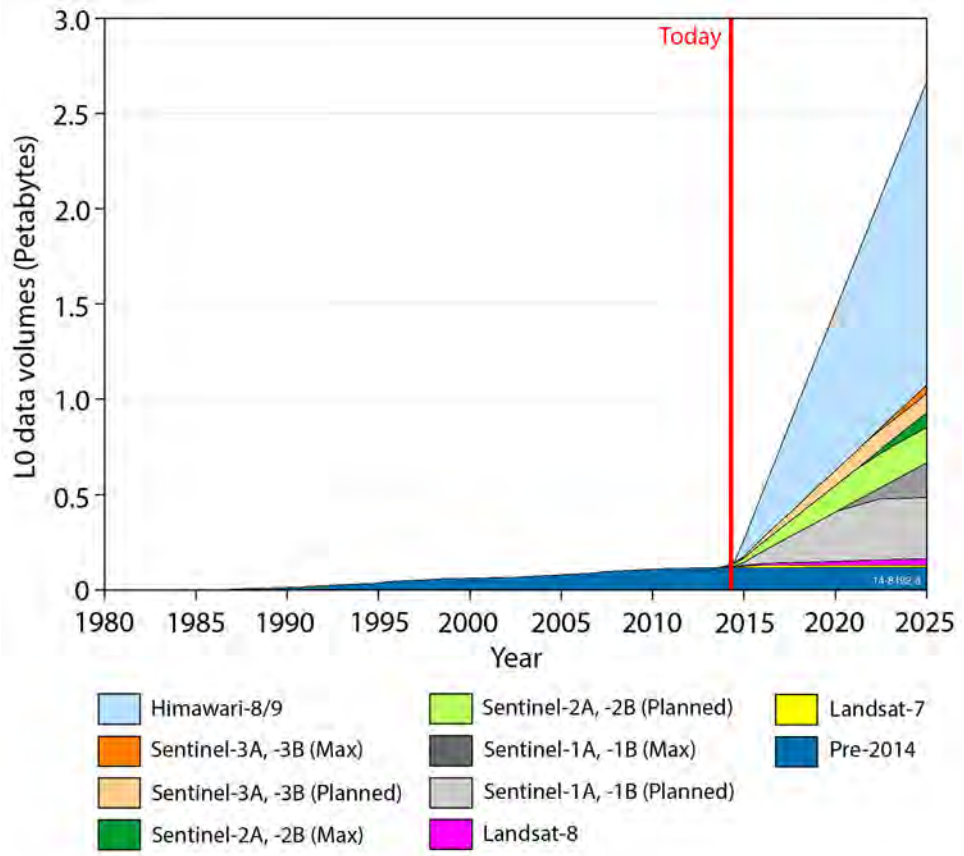


Figure 6-34. Estimated data volumes over Australia through 2025 from the Landsat, Sentinel and Himawari EOS programs.

Due to the labor-intensive nature of the processing requirements for EOS data, there have been few national-scale datasets created. As data volumes have increased over time, and the demand for the processed data and derived information products has also grown, it has become increasingly difficult to produce these products rapidly and achieve satisfactory policy outcomes using traditional processing methods. The result is that scientists, policy makers and the public in Australia have been "drowning in a sea of non-calibrated data" and have not been able to realize the full potential of the Australian Landsat Archive.

This issue is demonstrated in Figure 6-35, which shows a traditional 'on-request' data storage, processing and delivery chain that is used commonly by data curators to deliver data and derived products to users. It describes the typical workflow employed to deliver calibrated and other value- added data products to users upon a request. A typical scenario of how this process operates is:

- The client/user makes a request for a particular product;
- The client/user defines the desired footprint (in space and time) for the request;
- The client/user executes the search for available data (either through a search interface provided by the curator of the data or with the assistance of the data manager);
- The search is implemented by the robotic tape storage infrastructure – (note:) depending on the size and nature of the request this could be a very laborious process as the robot traverses up and down the banks of data tapes to locate the correct tapes from which to extract the data;

- The extracted data are placed in a temporary 'landing space' (a fast access 'spinning-disk' hard drive) where they can be downloaded to another computer for further processing or analysis. If the 'landing space' is not large enough for the volume of requested data, retrieval may fail, or will need to be completed in smaller stages, thus delaying the data request;
- Often only the 'raw' (or Level 0) data are archived, so once the data have been extracted from the archive they need to be processed to a calibrated level before analyses of physical observations can be conducted. In the case of EOS data, calibration includes: (1) Ortho-rectification to align each pixel/observation properly to its correct location on the surface of the Earth). (2) Spectral corrections to remove atmospheric effects from the observations to produce a calibrated measure of surface reflectance; and, (3) pixel quality assessments to identify the 'best' pixels to use in subsequent analyses. Because of the large file sizes involved, this process is expensive computationally with a high I/O overhead;
- Only after the data have been calibrated can they be used to derive other value-added interpretation products necessary to fulfil the original Client/user request;
- Once all calibrated and derived data products have been generated, they then need to be packaged and delivered. In many cases the data volumes are too large to transfer efficiently across a network. It is often more expedient to copy the data onto large capacity external hard drives and ship them to the client. Once the request has been fulfilled the data curator discards all of the processed data to save disk space.

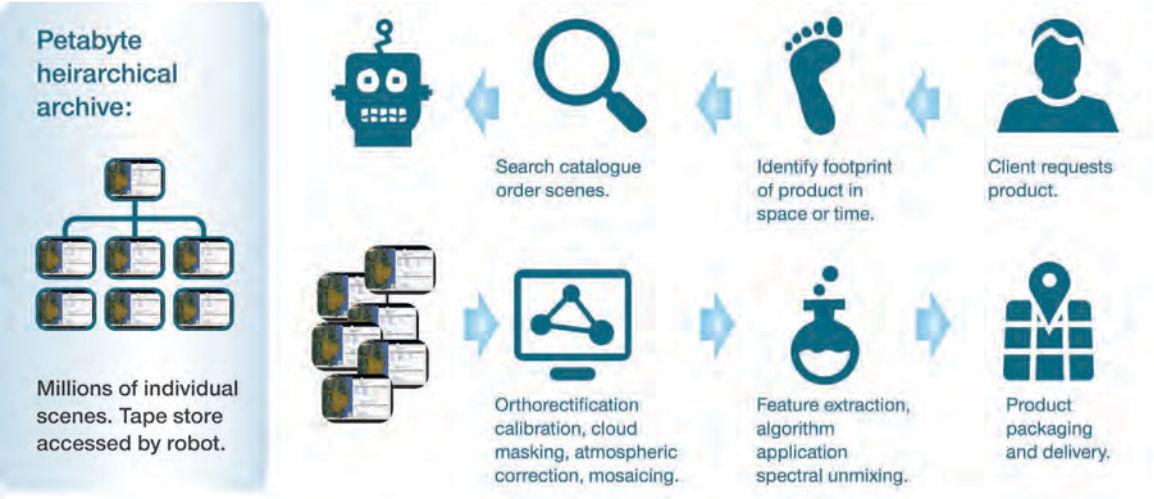


Figure 6-35. Traditional 'On-Request' data processing and delivery chain for EOS data.

Having completed this laborious process to locate, calibrate and retrieve the desired data for the original request the client/user may still face a number of challenges in making use of the data. Some examples include:

- They may not have the suitable software on their systems to properly use the data, or the data formats are not compatible with their analysis software;
- They may find that the original request was not adequate to answer the question(s) they were originally attempting to solve;

- They may find that there was an error in their original request; or,
- They may not have sufficient disk space to accommodate and use the requested data.

Similarly, the user may wish to do another query on an adjacent footprint or another client/user may wish to do a similar study over the same or another area of interest to them. In all of these cases it will be necessary to begin a new data request from scratch. This repetition and duplication of effort let alone the material and intellectual costs involved with completing these bespoke data requests will make it impossible for data custodians to support research, businesses innovation and government policy decisions into the future; particularly when one considers the impending 'data deluge'.

To manage the expected 'data deluge' effectively, it is imperative that we embrace a new paradigm where we access data instead of moving them around. By calibrating, organizing, and storing data in HPD infrastructures it may be possible to reduce the duplication by up to 80%. This has profound implications for the use and re-use of data and the ability for both government and industry to innovate and benefit from the emerging geo-services market. Figure 6-36 represents the new data access and analysis chain made possible by the Australian Geoscience Data Cube (AGDC).

Figure 6-36. Data access and analysis chain using the AGDC. New 'raw' data from sensors (e.g. EO Satellites) are acquired, processed, calibrated, and ingested into the AGDC. From there, multiple users are able to access and query the data via the AGDC without requiring a geomatics specialist to generate data products repeatedly from 'raw' uncalibrated data. Current applications of the AGDC are focused on meeting the needs of the Australian Government through the delivery of regional to continental scale information products (upper right – shown in blue). Future development of the AGDC will focus increasingly on providing information products and services that are of interest to the general public and private industry with access mechanisms spanning multiple platforms; including mobile devices (lower right – shown in grey).

4.1.1.4 From images to calibrated observations

Comparing Earth observation measurements through space and time requires a shift from 'raw' data collected by the sensor (Level 0) through to calibrated comparable measurements of the Earth's surface (Level 3) (Hilker *et al*., 2009). Transformation of data into a consistently structured, inter-relatable form, and the provision of the necessary supporting ICT infrastructure, are essential to this realization.

Figure 6-37 provides a high level view of the processing workflow necessary to transform 'raw' Level 0 Landsat data into calibrated EOS products. It shows the processing steps required to produce:

- Level 1 – Ortho-rectified data;
- Level 2 – Surface Reflectance Products using the Nadir BRDF (Bidirectional Reflectance Distribution Function) Adjusted Reflectance (NBAR) correction;
- Level 2 – Pixel Quality Assessment products; and,
- Level 2 – Fractional Cover spectral classification products.

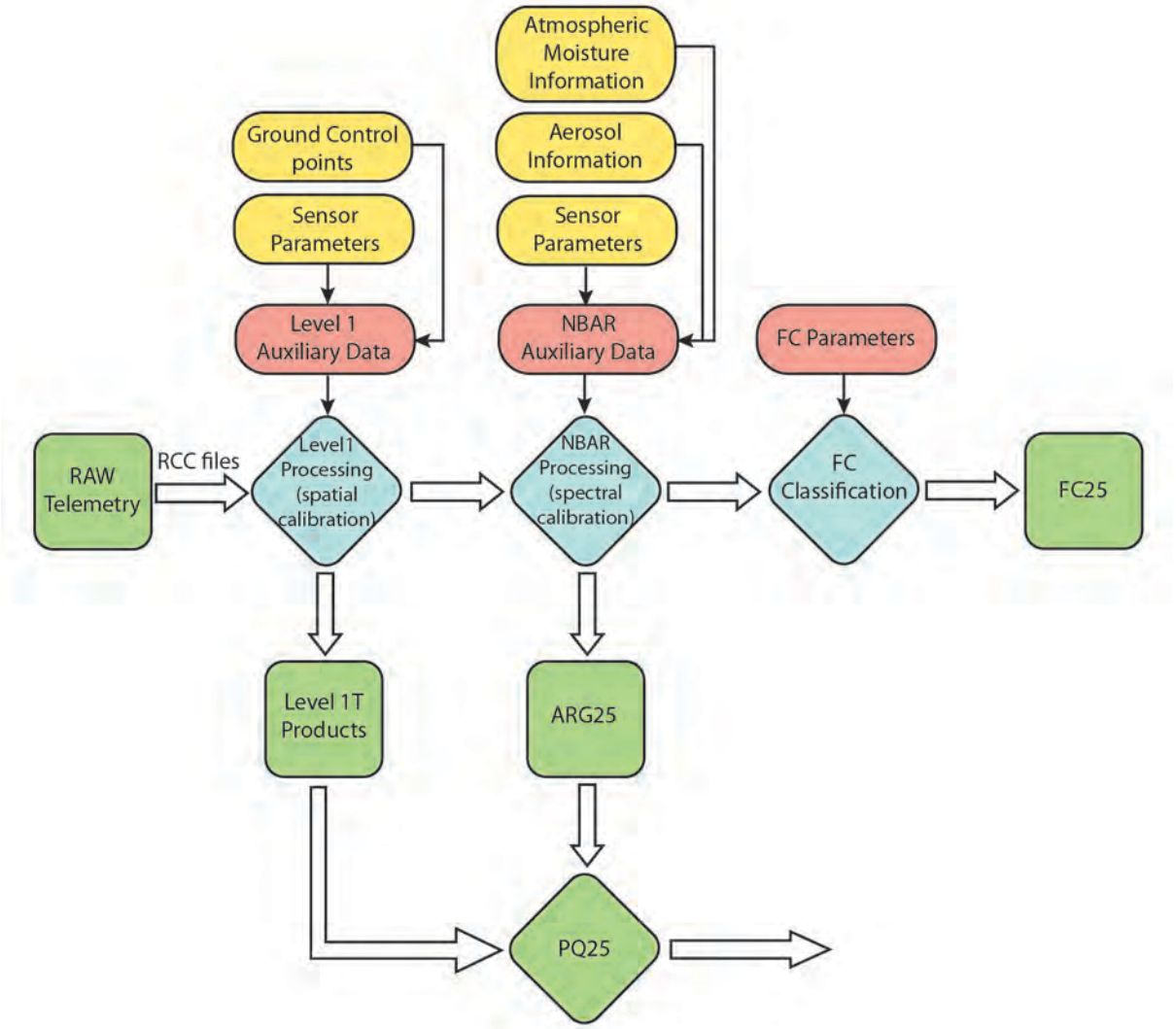


Figure 6-37. High-level workflow for Landsat Data Processing. Green boxes refer to data products, blue boxes refer to data calibration processes, orange and red boxes refer to sources of ancillary data used as inputs to the respective data processing stages. RCC files refer to files containing raw telemetry data. L1T data products refer to Level 1 Standard Terrain Corrected data. ARG25 refers to 25*m* Australian Reflectance Grid products. PQ25 refers to the Pixel Quality Assessment product and FC25 refers to Fractional Cover (FC) products derived from ARG25. Level 1 – Ortho-rectified Data Products.

The first step in the processing workflow is to generate Level 1 Ortho-rectified data products. This step translates raw telemetry data files (pass based) into Ortho-rectified Level 1 (scene based) products. This is achieved using the Level 1 Product Generation System (LPGS) software packages provided by USGS for Landsat International Cooperators. Depending on the quality of the available auxiliary information and the degree of cloud cover at the time of acquisition, the output products are reprocessed using one of the following levels of correction:

- Standard Terrain Correction (L1T)
    - Precision systematic radiometric, geometric and terrain correction using ground control points for geometric accuracy and a Digital Elevation Model for topographic accuracy.

    - Ground control points and DEM used for Level 1T correction of Landsat data over Australia come from the GLS2000 data set and SRTM DEM
- Systematic Terrain Correction (L1Gt)
    - Only applied to Landsat-7 data over Antarctica and Landsat-8 data where insufficient ground control points are available to process L1T data.
    - Systematic radiometric, geometric and terrain correction using the Ramp V2 DEM for elevation correction.
- Systematic Correction (L1G)
    - Systematic radiometric and geometric correction using data collected by the sensor and spacecraft – provides a geometric accuracy of 250m (1σ) for low-relief areas at sea level.

For the processed Level 1 data to progress to the next processing stage, they must be processed using the L1T correction.

(a) Positional Accuracy of Level 1 Products

There are many applications of medium spatial resolution satellite data that require accurate spatial registration of the imagery to a range of other spatial datasets. For multi-temporal analyses of Landsat data it is often necessary for them to be spatially registered to the sub-pixel level to minimize the effects of positional uncertainty from the analytical results. Figure 6-38 shows an example of sub-pixel registration of ground features across eight Landsat-7 ETM+ scenes.

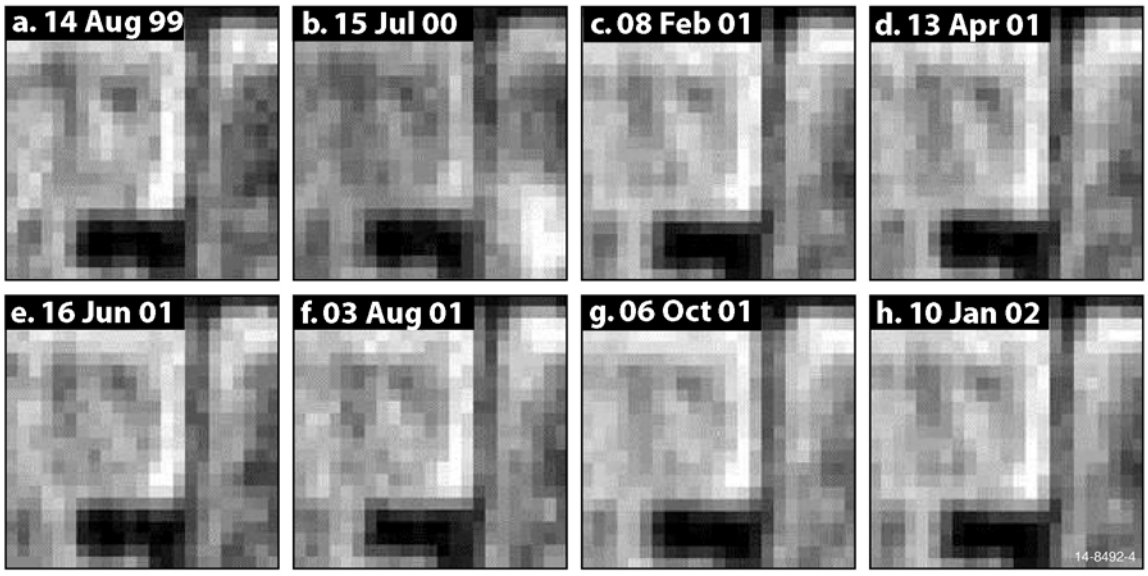


Figure 6-38. Example of sub-pixel registration when displaying the eight ETM+ scenes together through geo-linking.

Modeling spacecraft sensor position and attitude is the method commonly accepted for removing positional errors and distortions in satellite image data (Poli *et al.* (2004), Gruen and Zhang (2002), Fritsch and Stallmann (2000) and Hattori *et al.* (2000)). The Level 1 processing algorithm produces orthorectified Landsat products which are generated using Ground Control Points (GCPs) to refine the spacecraft model, and a Digital Elevation Model (DEM) to remove terrain distortions. Geoscience Australia currently uses

the Global Land Surveys (GLS) dataset (Gutman *et al.*, 2008) as the control source for Landsat processing using the LPGS software package.

Like all land surveys, the quality of GLS is limited by the availability of GCPs. In Australia there are large areas of the continent where GCPs are sparse. This introduces uncertainty in the positional accuracy of spatial products produced using the GLS. Recognizing this limitation, Geoscience Australia developed the Australian Geographic Reference Image (AGRI) to address the need for a higher resolution reference image, of known accuracy, over the entire Australian continent (Ravanbakhsh *et al.*, 2012). AGRI is a national mosaic that provides a spatially correct reference image at a 2.5*m* resolution across Australia. Figure 6-39 is a comparison of AGRI and GLS datasets with Differential GPS measurements across a road intersection, highlighting the issue of positional uncertainty in the reference image used to spatially register EOS datasets.

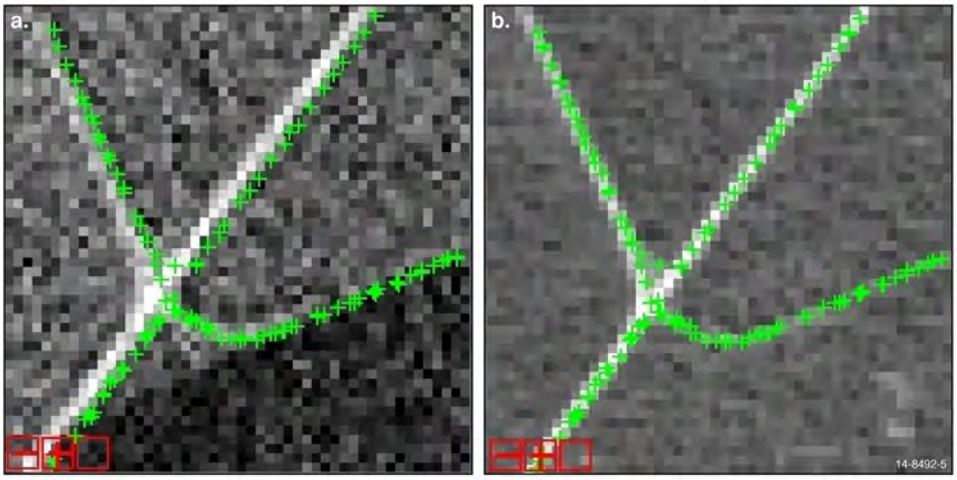


Figure 6-39. Comparison of (left) the GLS panchromatic band and (right) the AGRI panchromatic band (resampled to 15*m*) with differential geographic positioning system (DGPS) measurements across a road intersection.

To assess the accuracy of Level 1 data products produced using the GLS, a comparison between the AGRI and GLS reference datasets was conducted where the base AGRI dataset was resampled to 15*m* resolution and compared against the 15*m* panchromatic band from the GLS dataset. A digital image correlation technique (normalized cross correlation) was used to identify the quality assurance points from the reference AGRI image and its corresponding location on the GLS image. Figure 6-40 shows the assessment result between the AGRI and GLS reference datasets. This highlights the regions of Australia where there is a higher positional uncertainty of pixels in the GLS reference dataset compared to the AGRI dataset. Geoscience Australia is coordinating with the USGS to improve the GLS control source over Australia.

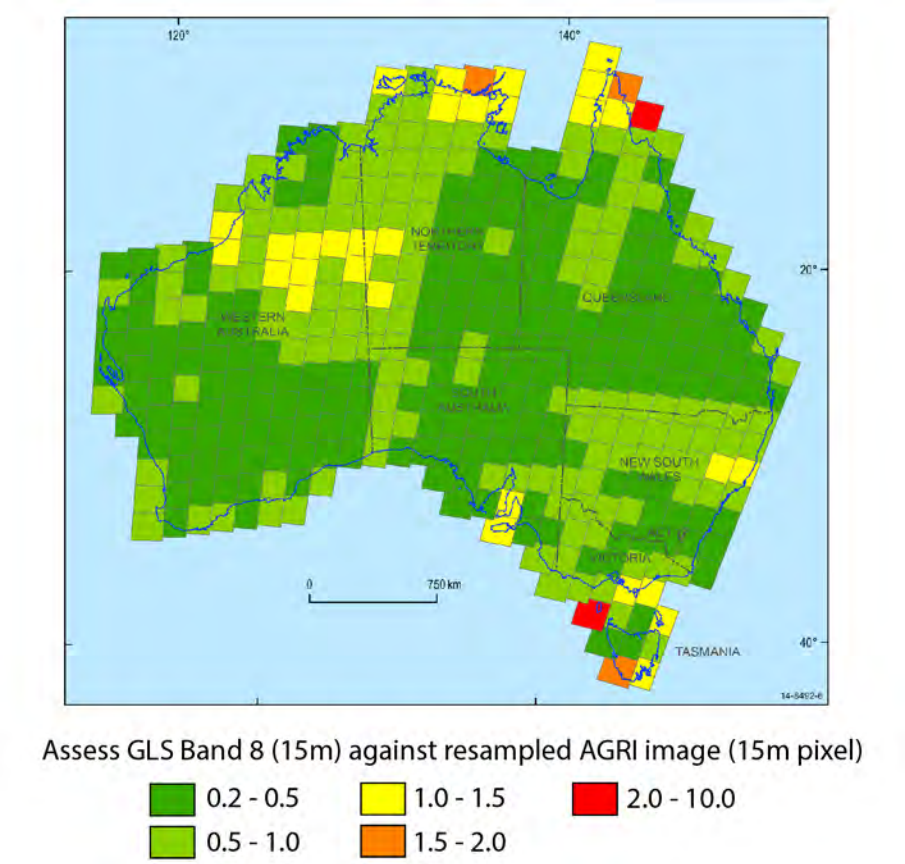


Figure 6-40. Map of Australia showing the spatial variation of positional accuracy of the GLS reference dataset compared to AGRI on a scene-by-scene basis. Green represents scenes where the difference between the location of pixels in GLS and AGRI is less than 15*m*. Red represents scenes where the difference between the location of pixels in GLS and AGRI is greater than 30*m*; this may indicate difficulties in obtaining clear observations of the ground at these locations.

- Level 2 – Surface Reflectance Data Products

The LPGS correction results in a calibrated dataset that represents the radiance at the top of the atmosphere (at the sensor). However, in order to compare and analyze data from multiple sensors across multiple scenes (or even the same scene acquired at different times), it is necessary to correct for the combined variations in sun and satellite view angles, aerosols and atmospheric moisture content. This requires a number of steps:

- Atmospheric correction
    - Using the MODTRAN4 radiative transfer modeling software (Berk *et al.*, 1999) along with the best available aerosol data from AERONET, AATSR, MISR, MODIS and Climatology ancillary data sources.
- Bidirectional Reflectance Distribution Function (BRDF) correction
- Terrain correction

Figure 6-41. Animation of the 25*m* Australian Reflectance Grid (ARG25) across the Australian continent showing seasonal variations throughout the year. Continental mosaics were produced using temporal stacks of ARG25 data from 2000 to 2011.

There are several methods for correcting BRDF effects (both empirical and physics based), although most empirical methods are difficult to apply in an automated workflow and add a significant overhead to the management and curation of the data. Geoscience Australia has developed a physics based coupled BRDF and atmospheric approach to correct both atmospheric and BRDF effects using BRDF shape functions derived from the MODerate resolution Imaging Spectro-radiometer (MODIS) launched by NASA on the TERRA Satellite on 20 December 1999 and currently being flown on both the AQUA and TERRA Satellite Missions by NASA observations (Li *et al*., 2010). Because this correction requires no human intervention to adjust empirical factors, it is possible to apply it as part of an automated parallel workflow on a High Performance Computer facility. Geoscience Australia has applied the NBAR correction to the Australian Landsat Archive (Landsat-5, -7 and -8) to produce a calibrated surface reflectance product which is sensor and scene independent (Figure 6-41). This product is available as the 25*m* Australian Reflectance Grid (ARG25) from both Geoscience Australia (http://www.ga.gov.au/search/index.html#/) and the Research Data Storage Infrastructure (RDSI) hosted at the National Computational Infrastructure.

- Pre-MODIS BRDF Correction for Landsat Data

The BRDF correction provides an objective, physics-based method of deriving Atmospherically Corrected Surface Reflectance observations from Landsat data (i.e. ARG25). Because the determination of the BRDF surface for each Landsat data acquisition is dependent on data provided by concurrent acquisitions from the MODIS sensor, there is a large portion of the Landsat-5 archive (March 1984 to December 1999) in which the standard NBAR correction cannot be applied.

To address this shortcoming and extend the temporal coverage of comparable data, Geoscience Australia led a study into the variation of MODIS BRDF shape functions acquired over a 10-year period from 2001 to 2011 (Li *et al*., 2013). Its aim was to develop a proxy BRDF shape function that could be reasonably

applied for the BRDF correction of Landsat data at times when no MODIS or other similar observational data are available. By comparing the variation of BRDF response over a 10-year period (Li *et al.*, 2013) were able to show that while there is strong intra-annual (or monthly) variation in BRDF response, the inter-annual (yearly) variation in BRDF response was relatively stable. This relationship enabled the successful development of an average BRDF shape function (Figure 6-42) that can serve as a proxy BRDF response where MODIS data are unavailable. This has enabled the extension of the ARG25 archive back to the launch of the Landsat-5 Satellite – an archive now containing 28 years of surface reflectance observations from the Landsat program.

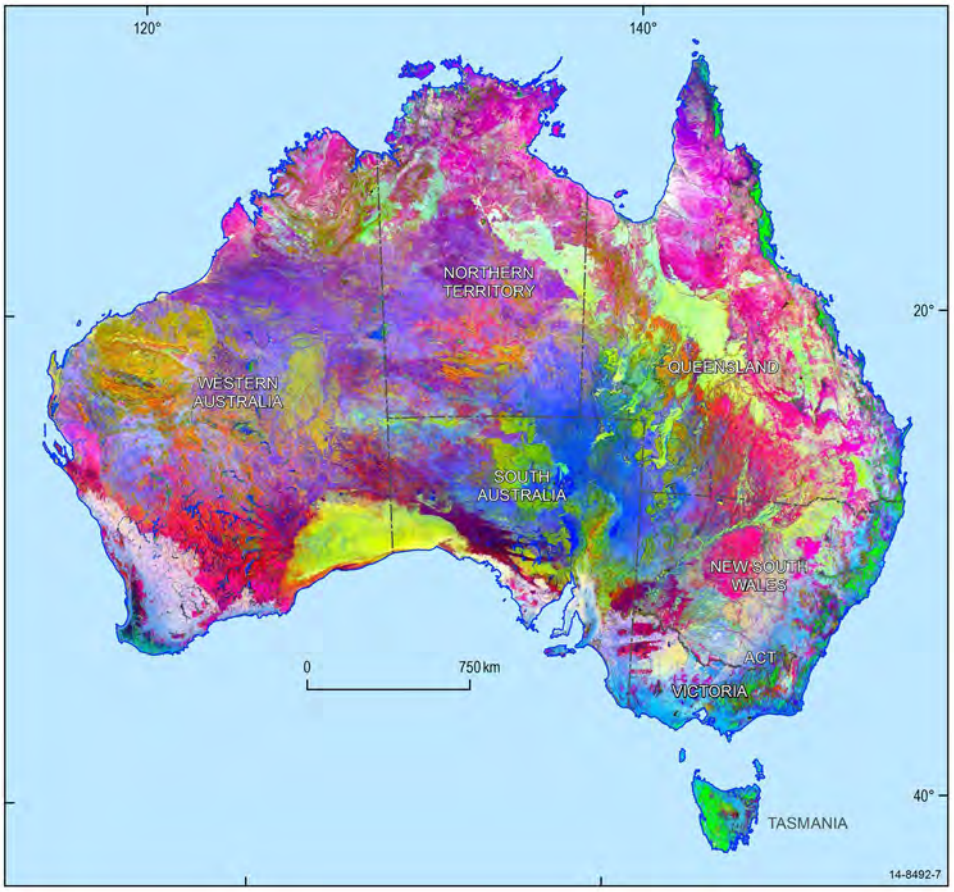


Figure 6-42. BRDF shape patterns in Australia, averaged over the period 2001 to 2011, indicated by color combination of 3 Minimum Noise Fraction (MNF) bands of shape indicators (Li *et al.*, 2013). The different colors represent variations in the way different types of land cover absorb and scatter light from the sun. Dark pink and mauve colors are representative of grasslands; light pink and pale blue colors represent agricultural crops and pastures; dark greens represent forested areas; yellow and brown colors represent semi-arid areas dominated by shrubs and sparse grasses; and, dark blue colors represent arid regions with little or no ground cover.

- Atmospheric and BRDF Correction for Mountainous Terrain

The standard NBAR correction does not apply a correction for variations in topography. This creates an issue when assessing multi-temporal land surface changes in areas of steep terrain, where the deep shadows produced by the terrain can either (a) obscure or (b) lead to systematic fluctuations in the observations of the surface reflectance. To obtain the corrected land surface reflectance and detect land surface change through time series analysis over rugged surfaces, it is necessary to remove or reduce the topographic

effects. While a number of empirical methods have been developed to correct topographic effects in surface reflectance observations (e.g. Richter *et al*., 2009; Soenen *et al*., 2005), it is difficult to apply these methods reliably in the automated processing workflows required to manage effectively very large and rapidly growing archives of EOS data.

Recognizing this issue, Li *et al*. (2012) extended the standard physics-based NBAR correction algorithm (Li *et al*., 2010) to reduce the topographic effects (Figure 6-43) on surface reflectance observations in rugged terrain. The approach taken was to improve the BRDF and atmospheric correction to account for both flat and inclined surfaces and to use them in conjunction with a 1-second SRTM (Shuttle Radar Topographic Mission) derived Digital Surface Model (DSM) product to adjust for both seasonal variations in terrain cast shadows and the shadow effects of nearby terrain.

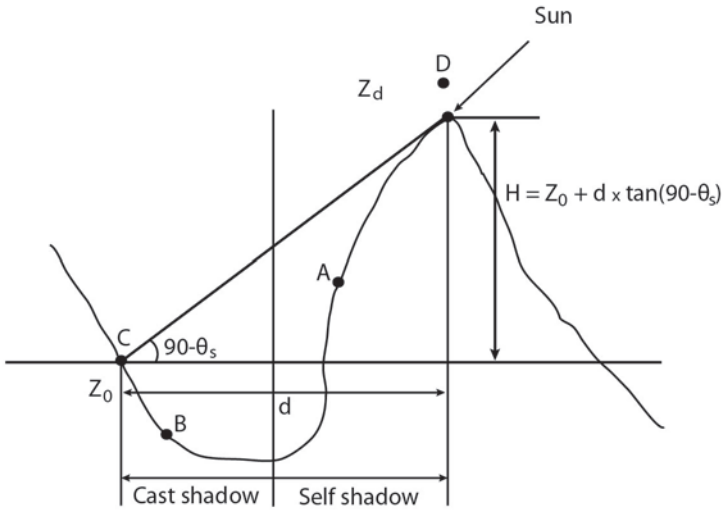


Figure 6-43. Two main causes of terrain shadow effect are Self shadow and Cast shadow (From Li *et al*., 2012). A is a location (or pixel) where the surface faces away from the sun and cannot be seen (i.e. Self-Shadow); B is a location (or pixel) that is blocked from the direction of the sun by nearby terrain, even though the surface may face the direction of sun light (i.e. Cast Shadow); C is a location (or pixel) with elevation Z_0 that is visible to the sun; D represents a body that is elevated above the surface of the Earth at Z_d (e.g. a cloud) that would occlude or cast a shadow on the location at C; and d represents the horizontal distance between point C and the highest point of adjacent terrain in the direction of the sun.

- Level 2 – Pixel Quality Assessment Products

As with most geophysical datasets it is important in analyzing EOS data to identify and remove 'noisy' or undesirable observations from the calibrated (Level 2) data. But can one remove all of the 'bad' observations from a dataset and still provide a data product that is relevant and useful to everyone? The answer is no. This is because what is considered to be 'noise' by one user could be a valid observation for another. Geoscience Australia, as a curator of geoscientific data, has an obligation to all users and so cannot provide access to authoritative data products that are only suitable for a specific stakeholder group. In the age of 'Big Data' it is not feasible to store multiple versions of the same data which have been filtered to remove different types of 'noisy', or undesirable observations.

For EOS datasets, most researchers typically use quality masking schemes to remove erroneous pixels, whether they represent clouds or simply saturated pixels. In the past, researchers would reject entire scenes

of data which are only partly cloud-affected in favor of only using completely cloud-free scenes. This has resulted in a significant waste of useful and valid observations of the Earth; particularly in regions such as Tasmania, that experience more cloud than others. Historically, these observations would have been lost and their value unrecognized and untapped.

Recognizing this issue Geoscience Australia has invested in the development of a Pixel Quality Assessment companion product for the Australian Landsat Archive (specifically Landsat-5, -7 and -8) that identifies a number of quality measures for every pixel of the associated ARG25 data product.

While there are numerous pixel quality products available for MODIS time-series data, prior to Landsat-8 there have been no equivalent products for Landsat-5 TM and Landsat-7 ETM+ datasets. However, there are a number of individual algorithms that detect clouds contained within a single scene of Landsat data. Of these, the most widely used algorithms are the Automated Cloud Cover Assessment (ACCA) (Irish *et al*., 2006) and the 'Function of mask' (Fmask) (Zhu and Woodcock, 2012) methods. Geoscience Australia developed a hybrid pixel quality product for Landsat data that can be applied in automated HPD workflows and thus provided a standard companion product to the ARG25 (Sixsmith *et al*., 2013).

The approach taken was to consider a range of possible contaminants in user driven analyses and to combine them into a single pixel quality product. These contaminants include:

- Under- and over-saturated pixels;
- Pixels without contiguity across bands;
- Sea/Ocean pixels;
- Cloud and cloud shadow affected pixels (using both ACCA and Fmask algorithms); and,
- Pixels affected by topographic shadows.

To generate a single overall quality metric of erroneous pixels, each result from the quality tests are set at specific bit positions. Valid bits ranging from 0 to 15 are combined into a single 16-bit quality value. For example, the pixel quality value 13166 is the cumulative sum of the binary value 0011001101101110 (reading from right to left). Table 6-7 shows the bit position and binary values allocated to each pixel quality test described above. Figure 6-44 shows an example of how the pixel quality product can be applied.

This pixel quality product allows a user to select which quality masks they wish to apply to the ARG25 data before undertaking their analysis; thus giving the user full control to apply customized masks or filters to remove only those pixels that are not relevant to their specific analysis. For example, a user interested in studying changes in coastal inter-tidal zones might choose not to apply a land/sea mask but to exclude all saturated, cloud and cloud shadow affected pixels from their analysis.

While the Pixel Quality product is an effective tool for enabling flexible exclusion of data that are either not useful or not appropriate for specific analyses using Landsat-5 and Landsat-7 data in their current form, it is not sufficiently tuned for application to other data sources (e.g. Landsat-8, Sentinel-2 etc.) Work is currently underway at Geoscience Australia to extend the Pixel Quality product so that it can be applied to additional EOS data sources.

Table 6-7. 16-bit Binary Pixel Quality Bit Mask for Landsat-5 (TM) and Landsat-7 (ETM+)

Pixel quality Test	Bit	Value	Cumulative Sum
Saturation Band 1	0	1	1
Saturation Band 2	1	2	3
Saturation Band 3	2	4	7
Saturation Band 4	3	8	15
Saturation Band 5	4	16	31
Saturation Band 61*	5	32	63
Saturation Band 62*	6	64	127
Saturation Band 7	7	128	255
Contiguity	8	256	511
Land/Sea	9	512	1023
Cloud (ACCA)	10	1024	2047
Cloud (Fmask)	11	2048	4095
Cloud Shadow (ACCA)	12	4096	8191
Cloud Shadow (Fmask)	13	8192	16383
Topographic Shadow*	14	16384	32767
*To Be Determined***	*15*	*32786*	*65553*

*Designed to match Landsat 7 ETM+. The thermal band for Landsat 5 TM will correspond to band 61, and the result is duplicated into band 62.
**Currently not set. A method for calculating topographic shadow has been developed, and will be added to the PQ. A final 16th test has yet to be investigated and developed.

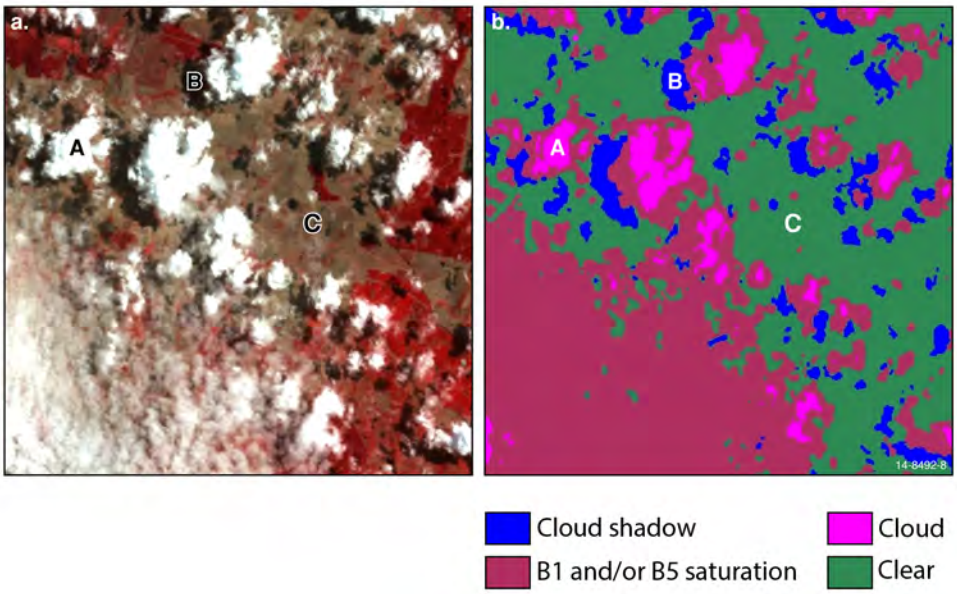


Figure 6-44. (a) False color (Band-4, -3, -2 RGB) image and (b) the corresponding Pixel Quality product. All saturated pixels are also cloud affected pixels. The binary representation and cumulative sum indicating a pass (1) or fail (0) for each quality test (reading from right to left) at points A, B and C are: A: 0011001111101110 (13294) – Bands 1 and 5 saturated; Cloud detected using both ACCA and Fmask; B: 0000111111111111 (4095) – Cloud Shadow detected; All other tests passed; and C: 0011111111111111 (16383) – Pixel is clear, all tests passed.

- Level 2 – Fractional Cover Products

Nationally consistent information about fractional cover dynamics enables policy agencies, natural and agricultural land resource managers, and scientists to monitor land conditions over large areas over long time frames. These data can be used to identify large scale patterns and trends and inform evidence- based decision making and policy on topics including wind and water erosion risk, soil carbon dynamics, land management practices and rangeland condition.

The Fractional Cover product (FC25) is derived from Geoscience Australia's ARG25 product and provides a 25*m* scale nationally consistent fractional cover representation of the proportions of green or photosynthetic vegetation, non-photosynthetic vegetation, and bare soil surface cover across the Australian continent (Figure 6-45 and Figure 6-46). It is generated using the algorithm developed by the Joint Remote Sensing Research Program and described in Scarth *et al*. (2010). A summary of the algorithm developed by the Joint Remote Sensing Centre is also available from the AusCover website: http://data.auscover.org.au/xwiki/bin/view/Product+pages/Landsat+Fractional+Cover.

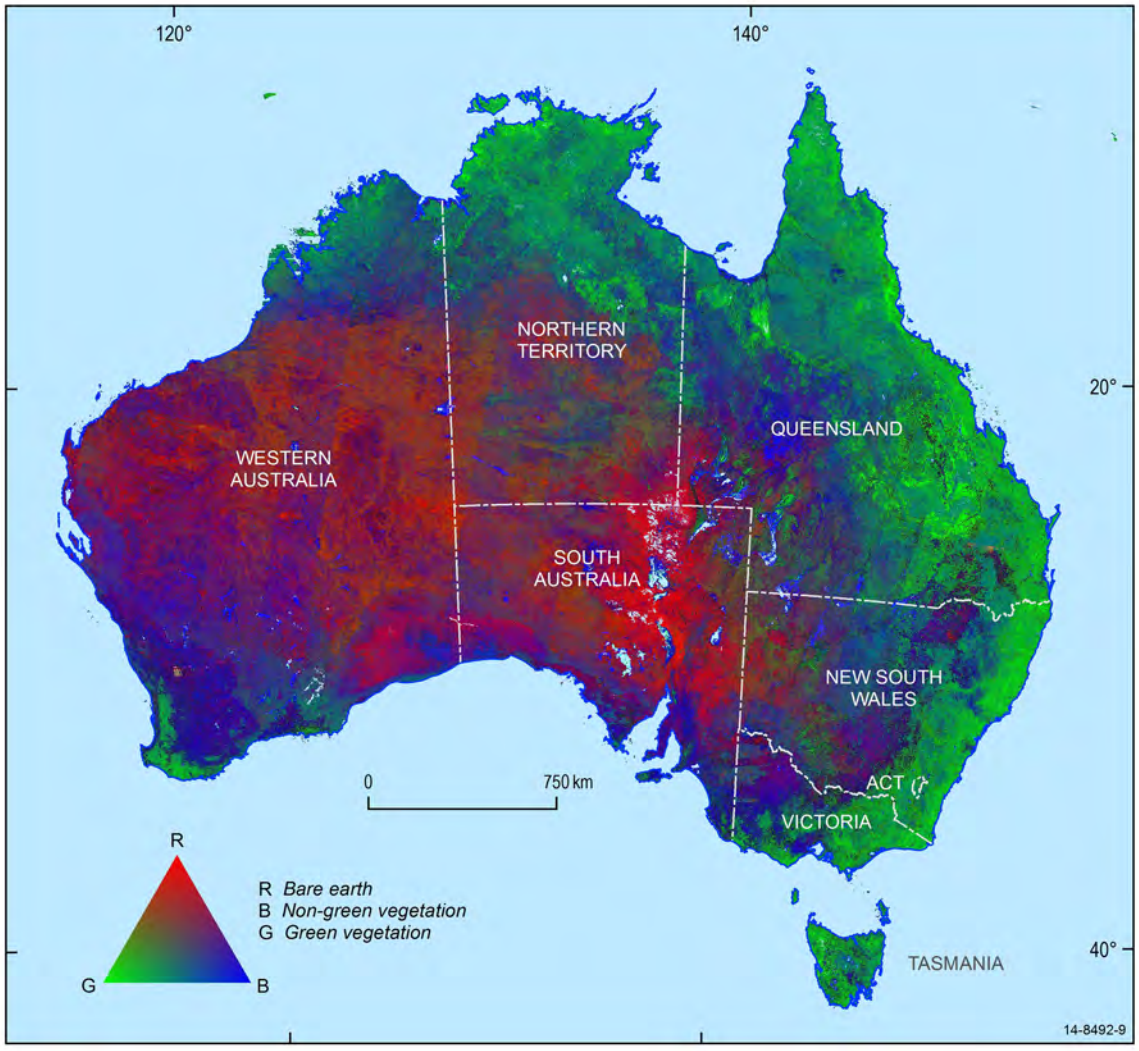


Figure 6-45. Continental scale FC25 mosaic created using data acquired from 2008-2011.

The Fractional Cover un-mixing algorithm uses the spectral signature for a pixel to break it up into three parts or fractions. This is based on field work identifying the spectral characteristics of each of the three fractions. The green fraction includes leaves and grass, the non-photosynthetic fraction includes branches, dry grass and dead leaf litter, and the bare soil fraction includes bare soil or rock. The bare soil, green vegetation and non-green vegetation endmembers are calculated using models linked to an intensive field sampling program whereby more than 600 sites covering a wide variety of vegetation, soil and climate

types were sampled to measure over storey and ground cover (Figure 6-46) following the procedure outlined in Muir *et al.* (2011).

The FC25 product suite is currently available for every scene Landsat Thematic Mapper (Landsat 5) and Enhanced Thematic Mapper (Landsat 7) scene acquired between 1986 and 2012 (for scenes that were successfully processed through to ARG25). It provides a consistent classification which will be an important foundation for fractional cover mapping and monitoring across Australia (Figure 6-47). It is a resource for natural resource managers, land surface process modellers, carbon modellers, rangeland managers, ecosystem scientists and policy makers. Future versions of the product suite will include data from the Operational Land Imager (OLI) sensor aboard Landsat 8.

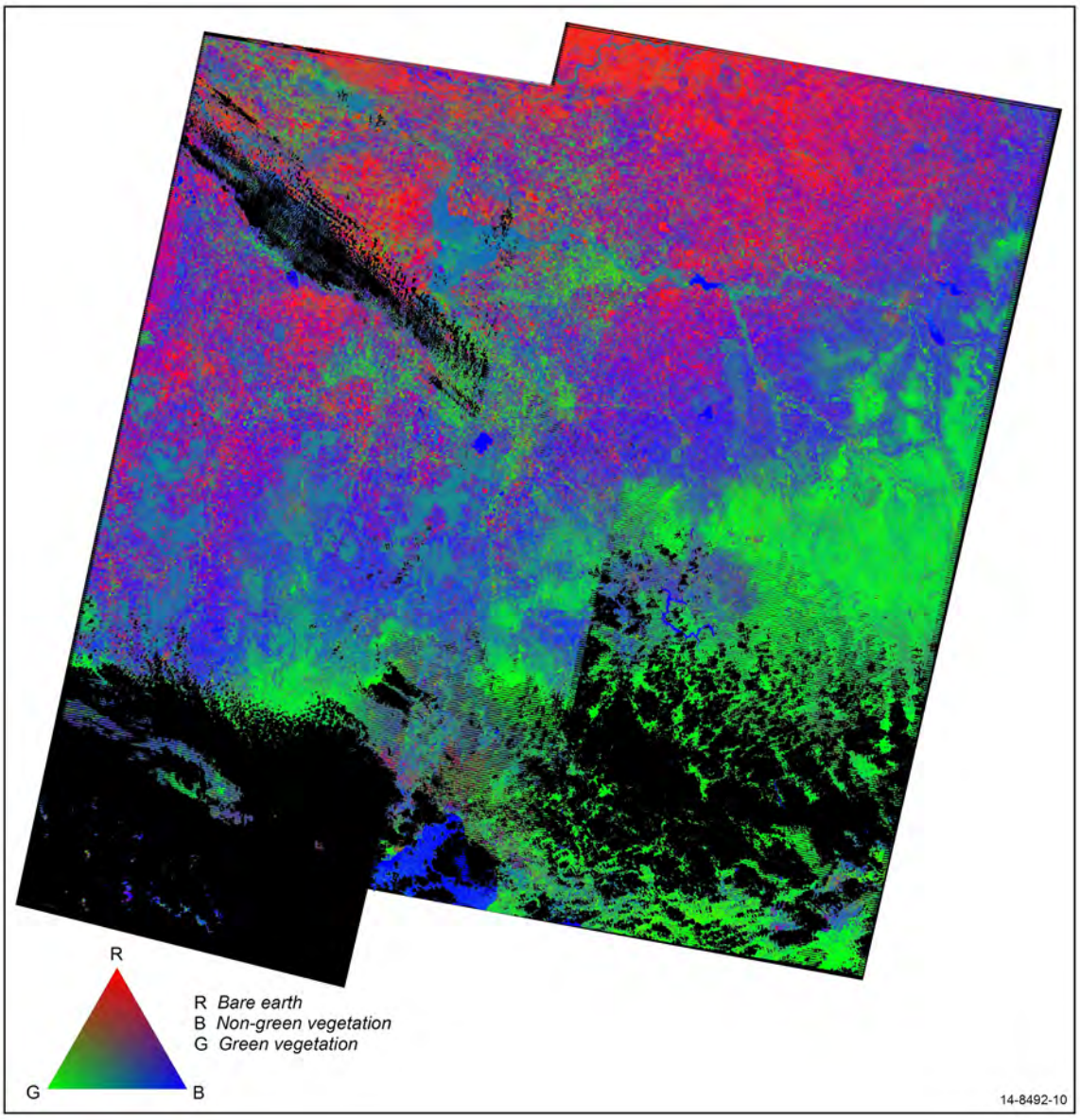


Figure 6-46. Four -scene composite RGB images of from both Landsat-5 and Landsat-7 sensors (Paths 92 and 93, Rows 85 and 86). Red = Bare Earth; Green = Green Vegetation; Blue = Non-Photosynthetic Vegetation. The black regions across the image represent pixels that have been masked out due to the effects of cloud and/or cloud shadow identified by the Pixel Quality Assessment Product.

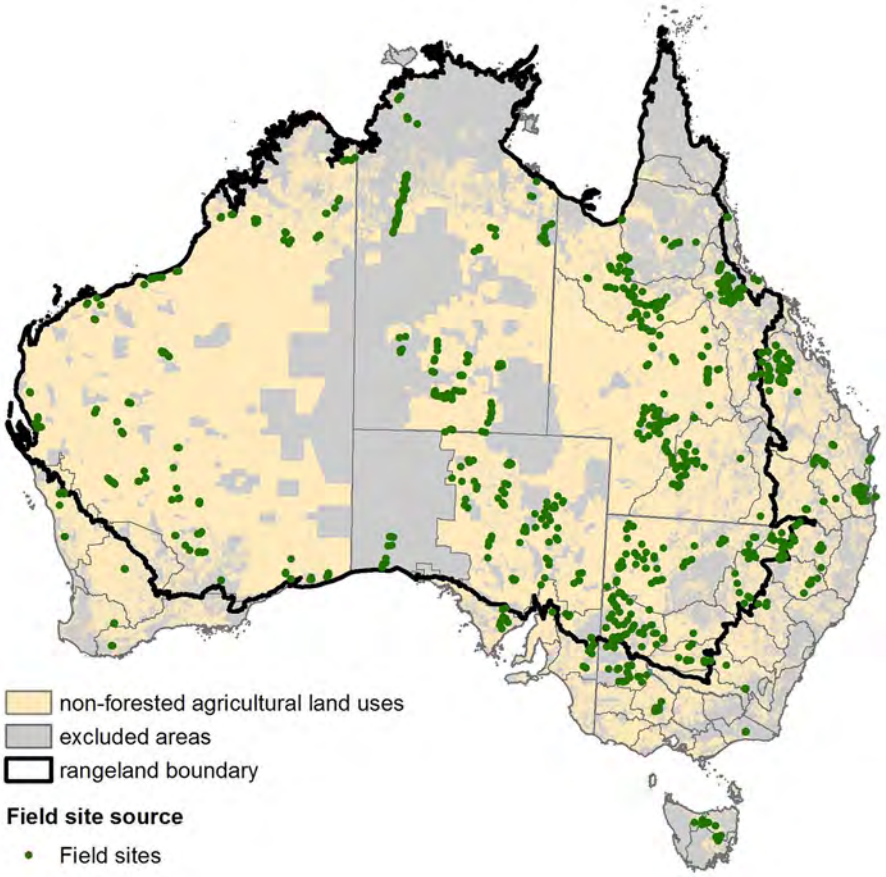


Figure 6-47. The distribution of the field sites across Australia used to train and evaluate the FC25 product (ABARES, 2014)

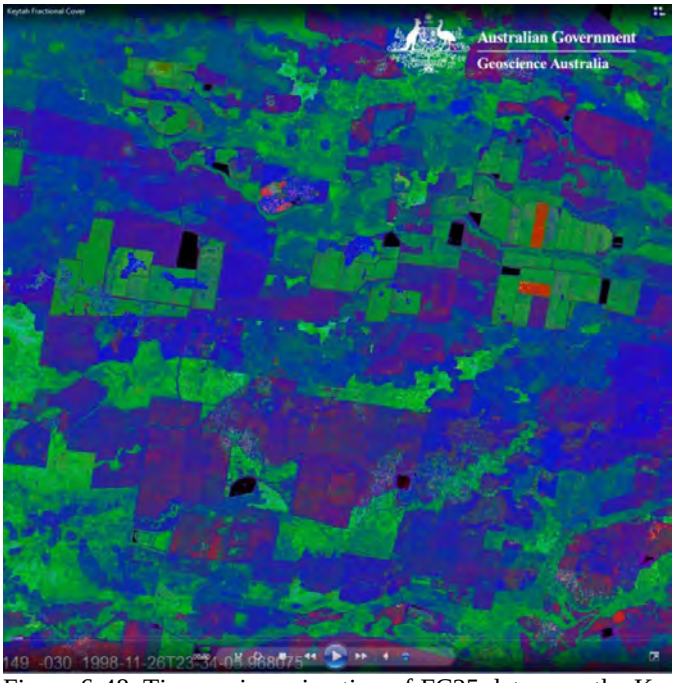


Figure 6-48. Time-series animation of FC25 data over the Keytah Station near Moree in New South Wales showing changes in ground cover between 1998 and 2012.

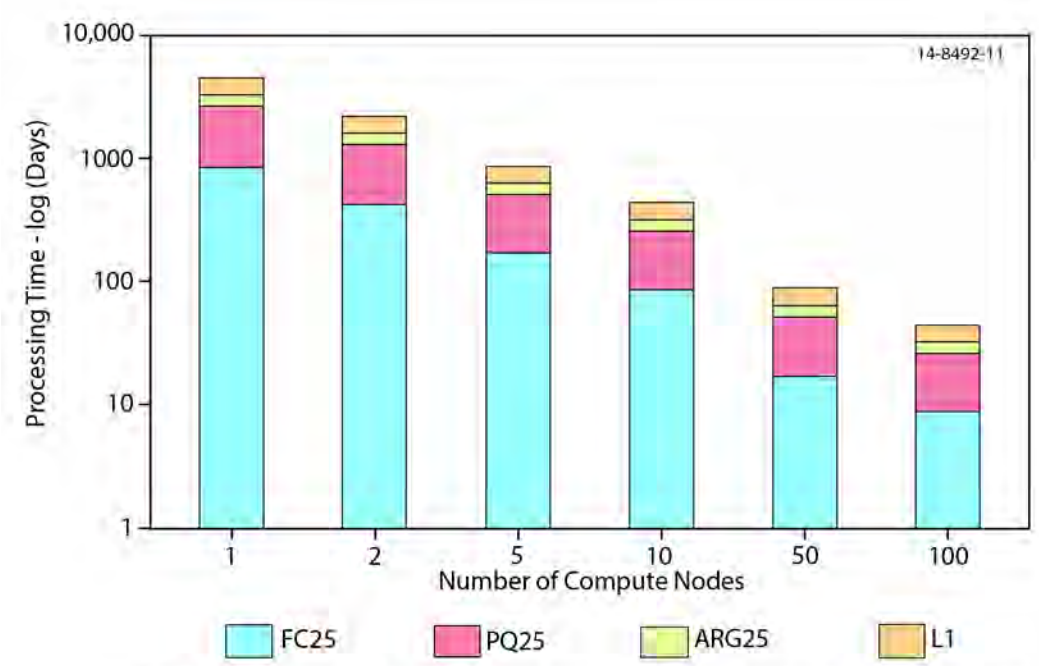
4.1.1.5 The Shift Away From Bespoke Production

Historically, the challenge of processing large sets of EOS data was too expensive computationally, and too time consuming to make it feasible. As a result, the on-request bespoke processing workflows described above have been used to provide calibrated EOS products. In recent years the increasing availability of HPC and other cloud-based ICT infrastructures have made it possible for data managers to reassess these bespoke workflows and investigate more automated solutions.

Figure 6-49 shows the comparison of end-to-end data processing times for 111,000 scenes of Landsat-5 and Landsat-7 data, representing eleven years of the Australian Landsat archive (from 2000 to 2010 inclusive). It shows an exponential decrease in processing time with an increase in the number of compute nodes used to conduct the processing run. Using a single node it would take more than 12½ years to process this volume of Landsat data. By contrast, using only 100 nodes (a relatively modest number of compute nodes compared to total numbers available on many HPC facilities) the entire 11 years of Landsat data were processed in 46 days. This demonstrates the importance and value of HPC facilities in enabling the shift away from bespoke workflows towards automated bulk processing of calibrated EOS products.

The shift away from bespoke production of EOS data products is not without its challenges. Firstly, the large volumes of the Level 0 EOS data necessitate implementing standardized automation workflows. Secondly, as shown in Figure 6-33, the combined volume of the various levels of processed data (Level 1 – Level 3) is at least three or four times the data volume of Level 0 data. This creates a significant challenge for storing and managing large archives of calibrated EOS data effectively. And, thirdly, it is not enough to simply calibrate the EOS data and store them ‘off-line’ in hierarchical robotic tape storage infrastructures. The requirements and expectations of key stakeholders for rapid generation of information products derived from time-series analyses necessitates using fast-access ‘on-line’ (or ‘spinning-disk’) storage infrastructures attached to HPC facilities to store and manage calibrated data (Level 2 and above).

Another challenge is that large collections of geoscientific data are often dynamic, with new data being added and data updated with the advent of improved processing and calibration methods. This presents a significant challenge for data curators in building and implementing suitable data infrastructures to effectively manage collections with the necessary degree of flexibility, while maintaining the required level of data provenance to enable higher level data products to be traced to their original observation. For example, many relational database systems and array based data storage optimization methods have difficulty in managing sparse, spatially and temporally irregular datasets where new data are not just being appended to the dataset.



FC25	1234	617	247	123	25	12
PQ25	617	309	123	62	12	6
ARG25	1851	926	370	185	37	19
L1	882	441	176	88	18	9
No. Compute Nodes	1	2	5	10	50	100

Figure 6-49. Comparison of end-to-end data processing times necessary to process 111,000 Landsat-5 and Landsat-7 scenes of data against the number of compute nodes used* (Purss *et al*., 2013). This analysis was conducted on the now retired VAYU HPC facility at the NCI. Results using the current NCI HPC facility (RAIJIN) would be approximately 8.6 times faster than these figures.

*The VAYU HPC facility at the NCI was retired in September 2013. Each compute node consisted of dual 4-core Intel Xeon (Nehalem) processors (i.e. 8 cores per node). Peak performance of VAYU was 140 TFlops – 8.6 times slower than the current RAIJIN HPC facility (1.2 PFlops).

4.1.1.6 International Forest Carbon Initiative

It has been estimated that deforestation and damage to forests is responsible for about 20% of global emissions of the greenhouse-enhancing gases that drive climate change. International efforts to ameliorate climate change (for example by creating a market for carbon trading) are estimated to cost several billion dollars/year and require sound and objective evidence to support them. In February 2009, the Commonwealth of Australia commenced the 3.5 year, AUD$8.4M, *"International Forest Carbon Initiative" (IFCI)* project to provide the building blocks for this evidence; namely, systematic delivery of the best available, consistently-processed satellite imagery, across the Australian continent on an ongoing annual basis.

This project provided the means for Geoscience Australia to replace aging and inadequate systems and practices, and replace them with systems providing a flexible and efficient foundation for all its satellite data processing needs. It resulted in a dramatic, step-function increase in the ability of Geoscience Australia to generate high-level satellite data products from the raw data stream acquired at ground stations. For example, the baseline for generating improved Landsat scenes at the commencement of the Project was around 4000 scenes per annum. Towards the end of the Project, in September 2012, Geoscience Australia

was generating more sophisticated satellite images than before, at a rate of about 1000 per day. These advances provided the necessary technological baseline for *Unlocking the Landsat Archive (ULA)* project (see Section 4.3.1.7) to be delivered.

The Project delivered nearly 19,000 processed Landsat images covering the entire length and breadth of Indonesia to the Government of Indonesia and CSIRO in support of Australia's International Aid Program. These images comprised the highest quality (lowest cloud, noise and haze) scenes available for each year from 1990 to 2010, inclusive. Consequently, ULA resulted in a significant expansion of Geoscience Australia's area of interest in terms of satellite data under its management, from its previous focus on the Australian continental land mass and adjacent areas, to essentially Australian coverage. This expansion of geographic coverage has already proven useful to unrelated projects such as Commonwealth Department of Foreign Affairs and Trade (DFAT) funded natural hazard projects in the Philippines; and this usefulness is expected to grow over time.

4.1.1.7 Unlocking the Landsat Archive

In 2011 the Australian Space Research Program initiated a 3½ year (AUD$3.5M) public/private consortium project called '*Unlocking the Landsat Archive*' (ULA). Its fundamental aim was to improve access to Australia's archive of Landsat data, and to provide an analysis capability for delivering environmental data and information to support government policy. The ULA Project was led by Lockheed Martin Australia (LMA) and involved technical input from Geoscience Australia (GA), the Victorian Partnership for Advanced Computing (VPAC), the National Computational Infrastructure (NCI) at the Australian National University (ANU) and the Cooperative Research Centre for Spatial Information (CRC-SI).

The technical work program under the ULA Project was run in four (4) parallel streams:

- *EO Science***.** This stream aimed to improve the fundamental processes and algorithms used to transcribe raw Landsat images into useful Earth observation products. This included improving cloud detection and removal algorithms, accounting for other atmospheric phenomena such as aerosols and atmospheric moisture content, and making the relevant local adjustments such as orthorectification, saturation evaluations and coordinate transformations.
- *National Nested Grid (NNG)***.** The adoption of a national nested grid is a significant step in reducing the complexity and efficiency of accessing raster data sets. This project undertook the development of the NNG concept and took it through the relevant processes to be adopted as a preferred data model for storing and accessing raster based data sets in the EO context. The NNG was accepted by the Australian and New Zealand Land Information Council (ANZLIC) as a Specification Guideline in 2012 (ANZLIC, 2012).
- *Workflow***.** Traditionally satellite processing has taken place in a linear workflow with standard techniques for creating relevant metadata and media distribution. This project enhanced these capabilities by implementing the NNG concepts as part of the satellite data processing workflow, and bridging the gap between the current satellite processing capabilities and the large-scale processing required for Landsat and other satellite data needs. These include adding measures for data provenance and quality at the pixel level that have not been available previously.

- *Computing.* Satellite data processing requires a significant amount of computing resources. This need has been steadily increasing with each new EO satellite launch and the ULA project explored and prototyped appropriate computational techniques and scalable resources to ensure Australia's ongoing EO science capability and capacity. The Computing stream implemented the systems developed in the three other streams on the NCI.

The ULA Project provided a solid platform that enabled the development of the AGDC to be achieved. The success of the project was recognized in the 2013 International Data rescue award in the Geosciences where the ULA Project received an honorable mention (Showstack, 2014; Purss *et al*. (2015).

4.1.1.8 The Australian Geoscience Data Cube

The ULA Project has enabled a paradigm shift toward automatically generating standardized, well calibrated, products at the national scale that incorporate both Landsat-5 and Landsat-7 data) and are scene independent. The outputs of the ULA Project, in particular the ARG25 data product, have been an important breakthrough enabling the rapid generation of continental scale mosaics and continental scale time series analyses. These products are generating much greater benefit from the Landsat Archive. However, while the observations are consistent across multiple sensors, atmospheric conditions and scene boundaries, the data produced by the automated work flow (Figure 6-37) are still stored as individual scenes.

It is possible to align individual pixels on the ground through time (Figure 6-50 and Figure 6-52), but because each scene has a slightly different spatial footprint, the geospatial processing required to align corresponding pixels is expensive computationally; particularly, when conducting a continental time-series analysis throughout the entire archive.

Added to this is the artifact known as the ‘north-south’ overlap of adjacent scenes along a single acquisition pass (Toutin, 2003). Data in this overlap region are duplicated in each adjacent scene and provide a reference frame that allows the original acquisition pass to be reconstructed by spatially aligning adjacent scenes of processed of data. However, this introduces a duplication of data that, when propagated throughout the entire collection, become significant. Furthermore, these intra-pass overlaps, coupled with the variation with latitude of scene overlaps across adjacent passes (Bindschadler, 2003), introduces complexity into the analysis processes to ensure that there is no double counting of observations and that pixel orientations are properly aligned to enable consistent comparison of observations throughout the entire collection.

By organizing and referencing the data using spatial database structures (such as ‘data cubes’) it is possible to remove the duplicated pixels from the collection and to thus remove the artifact and the complexity it creates for analyses. To explore possible methods for achieving this, Geoscience Australia conducted an assessment of ‘data cube’ technologies to enhance the effectiveness of ULA Project outputs and to ensure the Australian Landsat archive is truly “unlocked”. During this assessment Geoscience Australia developed a prototype data cube infrastructure guided by science-driven principles, to meet stakeholder requirements (see Figure 6-51).

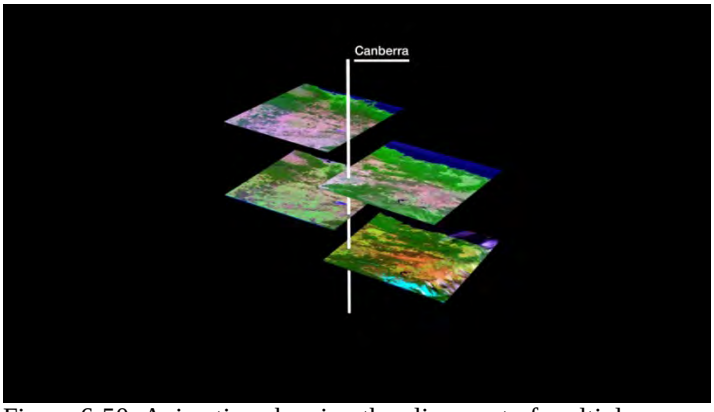


Figure 6-50. Animation showing the alignment of multiple scenes of Landsat data to enable comparison of the same point on the ground.

Figure 6-51. Video explaining the concept behind the Australian Geoscience Data Cube.

The Australian Geoscience Data Cube (AGDC) has evolved from this initial prototype and is being developed as part of a multi-agency collaborative project involving Geoscience Australia, CSIRO and the National Computational Infrastructure (NCI). The vision for the AGDC is to establish a *common analytical framework* for analysing and modeling geophysical properties of the Earth, in multiple dimensions (x,y,z,t *and* λ). The value proposition of the AGDC is massive data interoperability in a high performance computing environment by adopting common data structures suited to large scale computations. Although in its nascent stages, the AGDC is already allowing data analyses almost 10,000 times faster than was possible previously (Lewis *et al*., 2016) processes that would once have taken several years are now able to be completed in only a few hours.

4.1.1.9 Data Tiling

The success of AGDC depends on common structuring of data for computational efficiency. While raster data are easier to work with, AGDC's goal is to provide a common analytical framework for all types of data: point, vector and raster. The concept of tiling relates to segmenting data into regular and standardized units that are more efficient in analytical computations. These units can range from very simple tiling schemes that merely segment the data and produce a regularized grid, to highly sophisticated data

infrastructures such as a Discrete Global Grid System (DGGS) that enable data nesting and fusion from multiple sources through multiple resolutions.

At present the AGDC contains only raster (or gridded) observations of the Australian continent (excluding external/offshore territories) acquired by the Landsat-5, -7 and -8 missions from 1986 to 2014 (Level 3 data products). The data are gridded at a resolution of 0.00025 degrees and have been tiled using a simple tiling process that specifies tiles that are 4000 x 4000 cells in dimension, such that each tile is exactly one degree of latitude x 1 degree of longitude.

Figure 6-52 shows a tiling scheme currently implemented for the Australian Geoscience Data Cube (AGDC). While each tile location is fixed relative to the coordinate reference frame and datum used (e.g. WGS84), only tiles that contain observations are created. This results in a tile stack that is variable in space and time due to variations in spatial and temporal extent of the data, but containing tiles with constant locations (see Figure 6-53), which enables efficient queries and application of the data using standard relational database systems (the current version of the AGDC uses a PostGres relational database to index and query data tiles). The AGDC relational database also stores ISO 19115 compliant metadata records for each scene ingested into the AGDC on a per tile basis. This enables provenance tracking of each observation on a per scene basis.

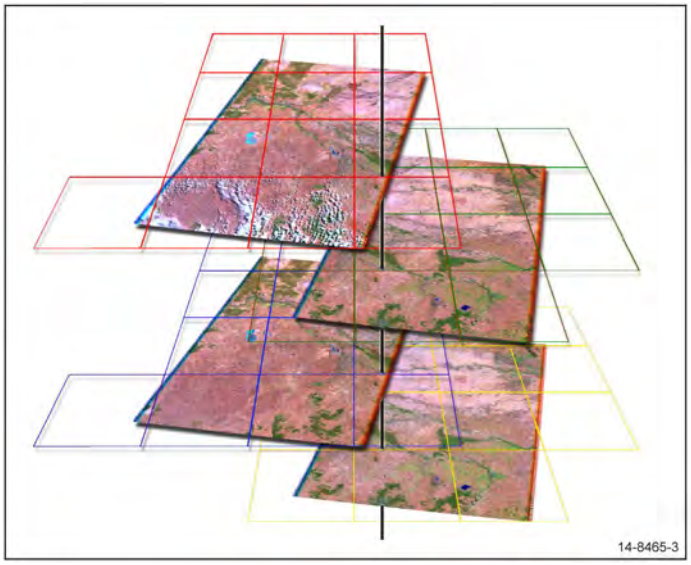


Figure 6-52. Representation of the tiling scheme implemented on the prototype.

The types of query that the system supports are, for example, 'calculate satellite-derived vegetation indices for a given "season" [arbitrary time period] over a defined geographic area using specific spectrum [band] information, and repeat for N years'.

A regular, and spatially fixed, tile grid is used to subdivide the individual Landsat scenes into separate files; each indexed spatially to the corresponding tile, and time-stamped according to the time of acquisition for the original Landsat scene. Because the data have been calibrated spatially and organized into regular tiles, it is possible to analyze the data contained within a single cell/pixel through the entire data store using the tile index as the primary key to locate specific cell/pixel (e.g. the black line intersecting the tile layers).

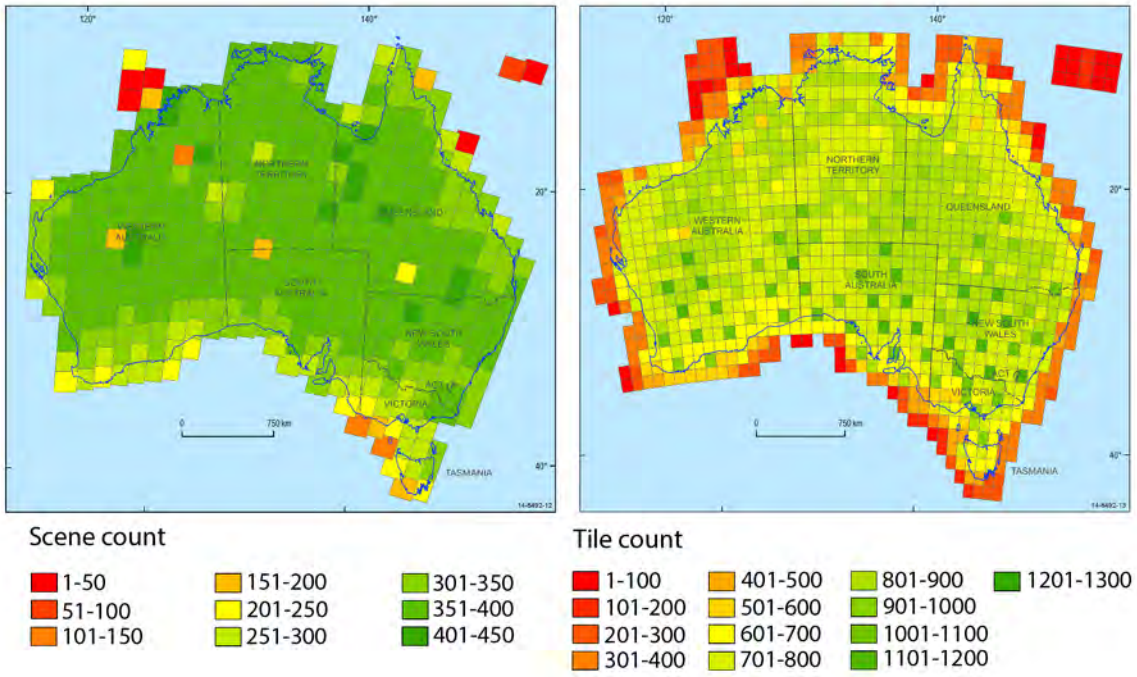


Figure 6-53 (left) Map of Australia showing number of scenes of calibrated (Level 2) Landsat data. Red=low number of observations; green=higher number of observations; (right) 'tiled' representation of the Landsat Archive over Australia. Areas in Red represent a low number of tiles; Green represents a high number of tiles.

4.1.1.10 The Australian Geoscience Data Cube in Action

In November 2012 the Murray-Darling Basin Authority (MDBA) approached Geoscience Australia to assist them in determining vegetation condition changes across the Murray-Darling Basin (Figure 6-54) on an annual basis; ultimately to monitor and evaluate the effects of the release of water as 'environmental flows', on the floodplain ecosystems. The challenges posed by this project were a catalyst for the initial development of the AGDC.

The goal of environmental watering is to protect and restore the resilience of the Basin's rivers, wetlands, floodplains, lakes and red gum forests, together with the plants and animals that depend on them. Many rivers in the Murray-Darling Basin (MDB) have been declining over the past 30 years; and thus a coordinated approach to overall water use was required by the MDBA. The Basin Plan (Murray-Darling Basin Authority, 2009) was developed under the Water Act 2007 to limit water use at environmentally sustainable levels by determining long-term average Sustainable Diversion Limits for both surface water and groundwater resources. The Environmental Watering Plan (Murry-Darling Basin Authority, 2011) is a central part of the Basin Plan and has a purpose to achieve the best possible environmental outcomes using the increased, but still finite, amount of water made available under the Basin Plan.

Figure 6-54. Map showing the area of interest for the MDB Project and the corresponding Landsat scene coverage (from Melrose *et al.*, 2013).

Vegetation monitoring is possible using EOS data as it has a unique spectral signature that distinguishes it from other types of land cover, using the visible and infrared ranges of the electromagnetic spectrum. Long time series analyses of these unique properties allow an assessment of the effectiveness of changes in environmental watering regimes.

Using the AGDC, Geoscience Australia delivered a range of vegetation indices and statistical products derived from spectral ratios of ten years of multi-band ARG25 data (Melrose *et al*., 2013). These derived products were used by Monash University to classify broad floodplain vegetation types and condition using an Artificial Neural Network model developed by Cunningham *et al*. (2009). The Vegetation Indices generated using the AGDC include (Equations 6-1 through 6-6):

1) Enhanced Vegetation Index (*EVI*)

$$EVI = G \times \frac{(NIR - RED)}{(NIR + (C1 \times RED) - (C2 \times BLUE) + L)} \tag{6-1}$$

where: $C1 = 6$, $C2 = 7.5$, $G = 2.5$, $L = 1$

2) Normalized Difference Vegetation Index (*NDVI*)

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \tag{6-2}$$

3) Normalized Difference Soil Index (*NDSI*)

$$NDSI = \frac{(RED - SWIR)}{(RED + SWIR)} \tag{6-3}$$

4) Normalized Difference Moisture Index (*NDMI*)

$$NDMI = \frac{(NIR - SWIR)}{(NIR + SWIR)} \tag{6-4}$$

5) Soil Adjusted Total Vegetation Index (*SATVI*)

$$SATVI = \left[\frac{(SWIR - RED)}{(SWIR + RED + L)} \right] \times (1 + L)) - \left(\frac{LWIR}{2} \right) \tag{6-5}$$

where: $L = 0.5$.

6) Specific Leaf Area Vegetation Index (*SLAVI*)

$$SLAVI = \frac{NIR}{(RED + SWIR)} \tag{6-6}$$

where: *BLUE* = Landsat-5/Landsat-7 Band 1 ($0.45\mu m$ – $0.52\mu m$) – visible blue; *RED* = Landsat-5/Landsat-7 Band 3 ($0.63\mu m$ – $0.69\mu m$) – visible red; *NIR* = Landsat-5/Landsat-7 Band 4 ($0.76\mu m$ – $0.90\mu m$) – near infrared; *SWIR* = Landsat-5/Landsat-7 Band 5 ($1.55\mu m$ – $1.75\mu m$) – short wave infrared; *LWIR* = Landsat-5/Landsat-7 Band 7 ($2.08\mu m$ – $2.35\mu m$) – long wave infrared.

Summary statistics were generated for each reflectance band (LS-5/LS-7 Bands1 – 7) and each index (*EVI*, *NDVI*, *NDSI*, *NDMI*, *SATVI*, and *SLAVI*) aggregated over each seasonal time period (Table 6-8). These statistical quantities included Sum, Valid Observations, Mean, Variance, Standard Deviation,

Skewness, Kurtosis, Max, Min, Median (non-interpolated value), Median Index (zero based index), 1st Quantile (non-interpolated value), 3rd Quantile (non-interpolated value) and Geometric Mean calculations.

Table 6-8. Date Ranges Used for Seasonal Aggregation of Data Products.

Season	Date Range*
Summer	01 December to 31 March
Autumn	01 March to 30 June
Winter	01 June to 30 September
Spring	01 September to 31 December

*Defined seasons overlap by one month

Figure 6-55 illustrates the processing workflow used to produce the time-series summary satellite images, vegetation indices and statistical data products. A total of 203,148 grid tiles of Landsat-5 (63,674 tiles) and Landsat-7 (139,474 tiles) data were processed; equivalent to 26,343 scenes (8,667 Landsat-5 and 17,676 Landsat-7). This represented 2.112 x 10^9 pixels/observations of ARG25 data to which the above analyses were applied. An example output of NDVI from this workflow is shown in Figure 6-56. A typical processing run on the VAYU HPC facility took less than 12 hours; compared to an estimated three years using traditional processing workflows.

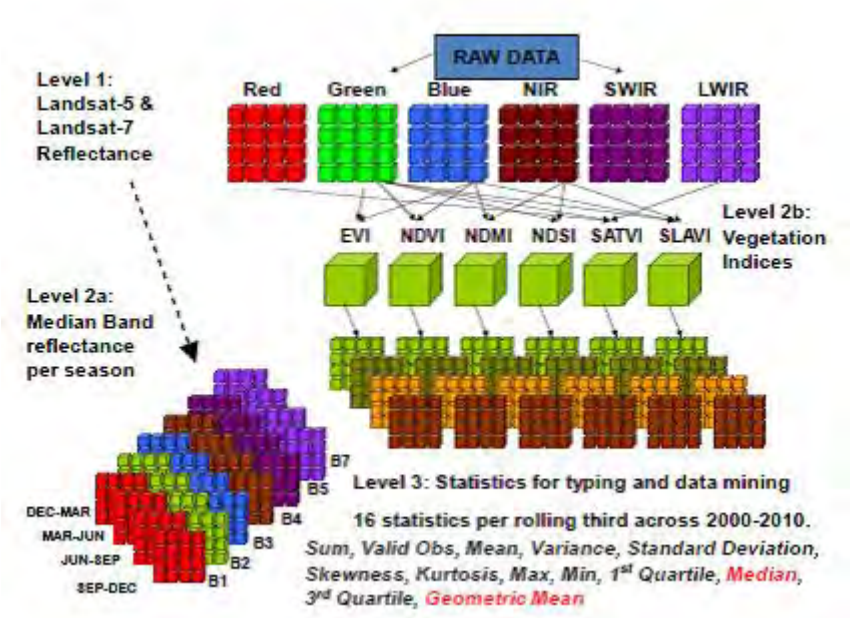


Figure 6-55. Graphical workflow of data processing and outputs (Melrose *et al*., 2013).

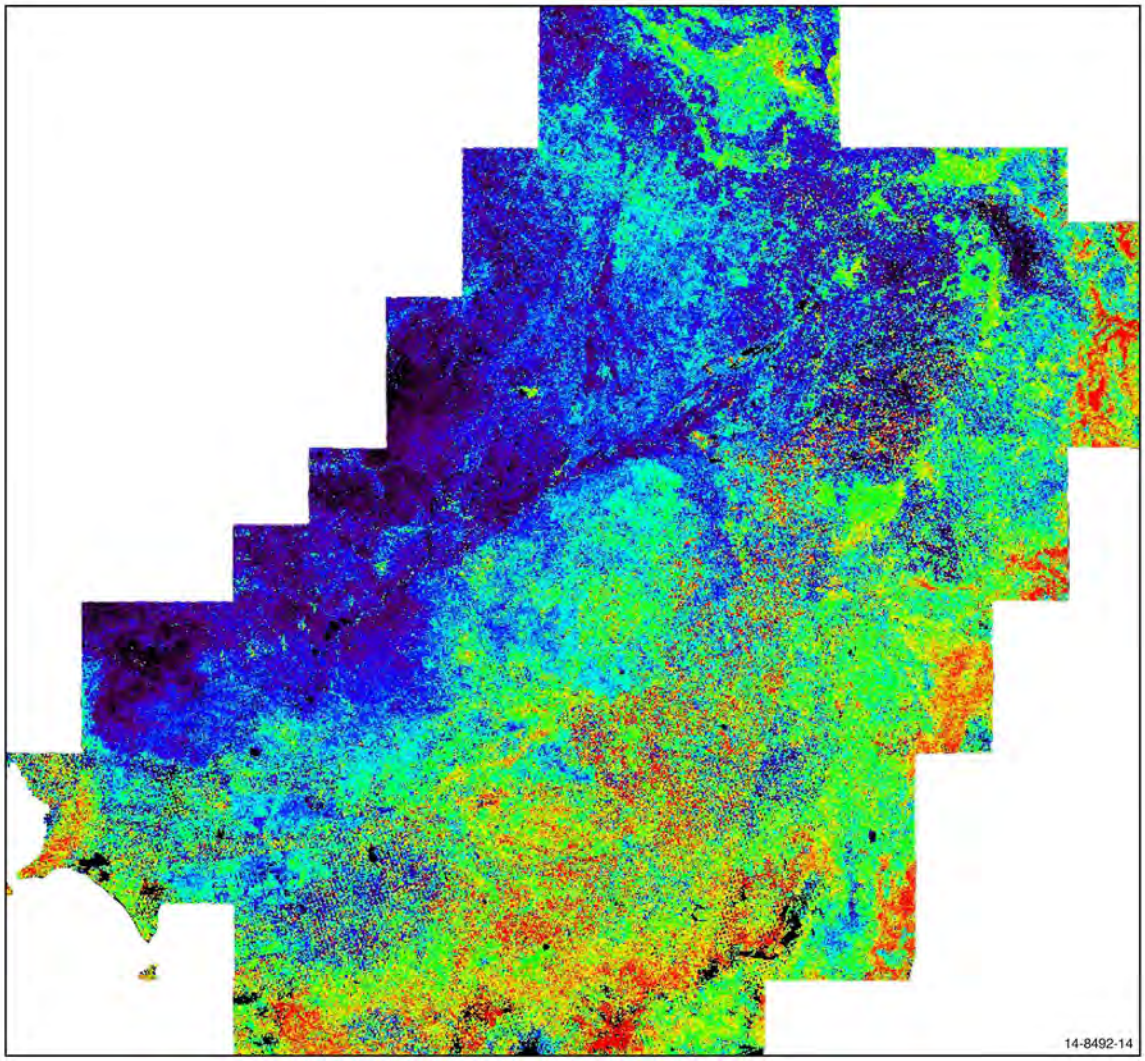


Figure 6-56. The first multi-scene, time-series mosaic created over the Murray-Darling Basin for the median Normalized Difference Vegetation Index (NDVI). NDVI is related directly to the photosynthetic capacity and hence energy absorption of plant canopies (Sellers, 1985; Myneni *et al.*, 1995), and has values ranging from -1 to +1. Low values of NDVI (close to -1, shown in dark blue) indicate lower photosynthetic activity associated with either unhealthy or sparse vegetation. Higher values of NDVI (close to +1, shown in yellows and reds) indicate higher photosynthetic activity associated with healthy or dense vegetation. The mosaic contains cloud-masked composite images created from 957 Landsat-5 and Landsat-7 scenes captured from 01 December 2008 to 30 March 2009.

4.1.2 *The Road to Big Data Interoperability*

The increasing integration of geospatial data into our everyday lives is driving an increasing expectation of spatial information on-demand, and with minimal delay. For the custodians of these data there is increasing pressure to deliver data and information as a service and to enable the rapid generation of these integrated information products. However, there remains a gap between expectation and the present reality. Combining multiple sources of geospatial information on-demand is a necessary key step in the geospatial-intelligence cycle, and a major challenge.

As demonstrated by the AGDC, this time gap can be bridged by converting traditional data archives into standardized architectures that support parallel processing in distributed and/or high performance computer

environments. Implementing global interoperability standards is critical to enable data fusion to be achieved without the bulk movement and refactoring of data across different networks. A common framework is required that will enable linking very large multi-resolution and multi-domain datasets, and enable the next generation of analytical processes.

4.1.2.1 Global Data Cube Hubs

As a tool, the Australian Geoscience Data Cube (AGDC) is not bound to any specific piece of ICT infrastructure. It is an open source spatial data infrastructure (available from github https://github.com/GeoscienceAustralia/agdc) with a growing and diverse user community. Consequently, there is potential for expanding the AGDC approach to create a global network of data cubes where large geoscientific and geospatial datasets can be stored and collocated in suitable HPC/HPD facilities. These data infrastructures can be used to enable big data analytical processes to be adopted. Essentially this will facilitate creating regional data hubs where data are stored, managed and accessed using HPD infrastructures (Figure 6-57).

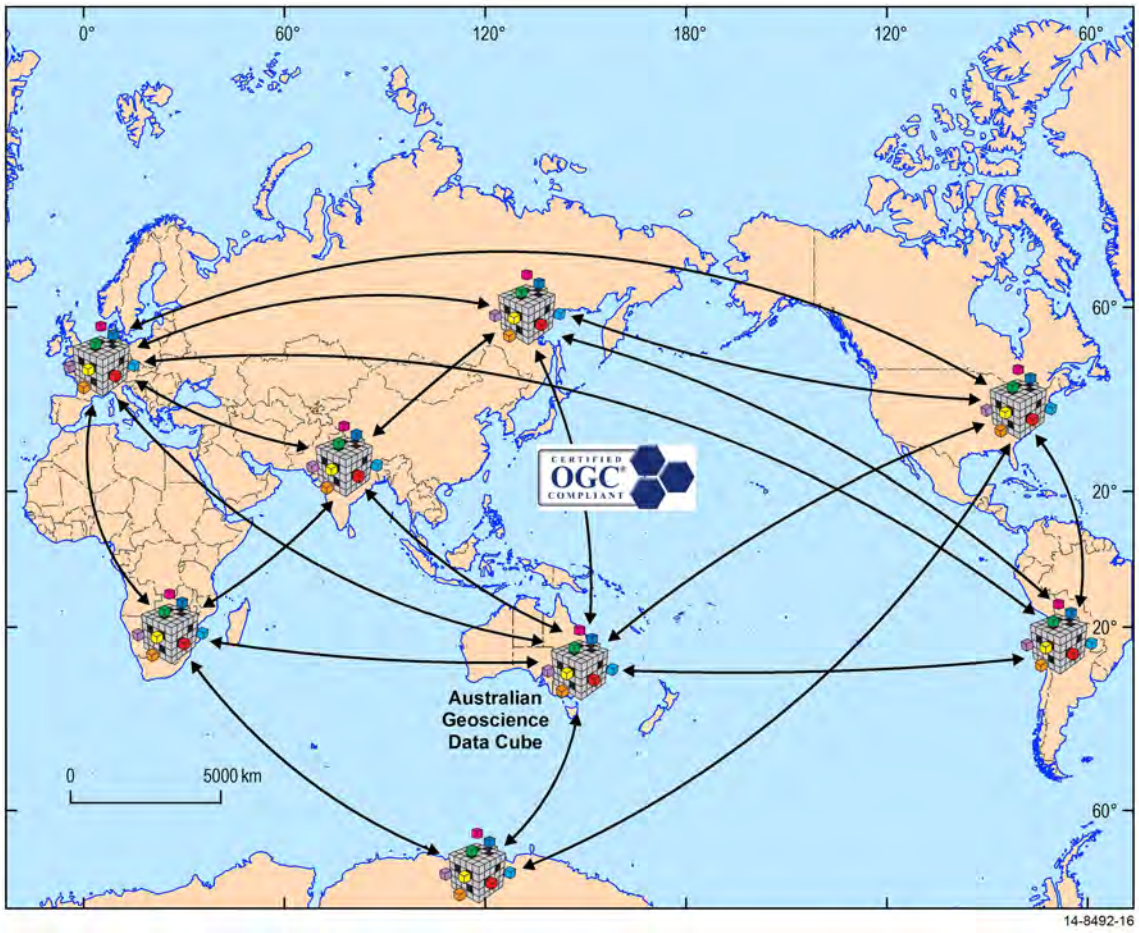


Figure 6-57. Conceptual representation of a global network to interact with distributed Data Cubes utilizing OGC Web Service Architectures.

The seamless exchange of data and information between these regional data-cube hubs requires two key pieces of technology. Firstly, the data must be organized using consistent but flexible data infrastructures, such as Discrete Global Grid Systems (DGGS), that enable analysis and fusion of multiple types of data

(e.g. raster, vector and point data types) at multiple spatial and temporal resolutions. Secondly, these regional hubs and their data infrastructures must be able to communicate using open standards that facilitate data interoperability and the brokering of computation without requiring bulk transfer of large geospatial datasets. The Open Geospatial Consortium (OGC®) has recognized that the geosciences can only achieve their full potential through fusion of diverse Earth observation and socio-economic data and information. In a multiple provider environment, fusion is only possible with an information system architecture based upon open standards (Percivall, 2013).

4.1.2.2 Discrete Global Grid Systems

Recognizing this need, the OGC has established a new Standards Working Group (SWG) to develop such an interoperability standard for DGGSs (www.opengeospatial.org/projects/groups/dggsswg). It is being led by Geoscience, Australia, Landcare Research, New Zealand, and The PYXIS Innovation Ltd. Their aim is to define the core qualities of DGGSs, to make them interoperable with conventional and other DGGS data sources, and to standardize operations on them through OGC Web Service architectures.

A DGGS represents the Earth using a tessellation of nested cells and is designed to ensure a repeatable representation of measurements that is better suited to today's requirements and technologies rather than primarily for traditional navigation and manual charting purposes. As with the various coordinate reference systems, there are a number of different implementations of DGGSs, each tailored to specific data types and/or principle use cases. DGGSs are a transformative technology because they remove the key obstacles to enabling real-time data integration with less human intervention. They also better enable geoprocessing performance to scale in step with rapid advances in hardware, software and business models.

Combining or integrating layers is easier in a DGGS because items of information automatically line up spatially. This is much like overlaying information across congruent rasters, and far easier than having to perform overlay using points, lines and areas. DGGS transforms ‘paper-age’ Earth coordinates to ‘computer-age’ coordinates, vastly reducing the computational burden and errors imposed by ‘paper-age’ coordinate transformations.

There are many possible DGGSs, each with their own advantages and disadvantages, with variations in geometry, alignment and granularity of cells. Some working examples of DGGS implementations include:

- The PYXIS WorldView™ client application, which uses the Icosahedral Snyder Equal Area Aperture 3 Hexagonal DGGS (Snyder Grid) (Figure 6-58). WorldView™ has been used to demonstrate successfully multi-source on-demand data integration and analysis within several OGC Open Web Services cross-community interoperability test-beds and the Group on Earth Observations (GEOSS) architectural pilot projects; and,
- The New Zealand government research institute, *Landcare Research,* is currently developing an open-source DGGS geographical analysis system for worldwide scientific collaboration called SCENZ-Grid (Figure 6-59). SCENZ-Grid uses a rectangular DGGS of 3x3 tessellations of the six faces of a Hierarchical Equal Area iso-Latitude Pixelated ellipsoidal cube (HEALPix). This is designed for use in HPC and cloud architectures and is being developed for inter-disciplinary environmental modeling. HEALPix originated at NASA's JPL for astrophysical analyses of massive full-sky data-sets.

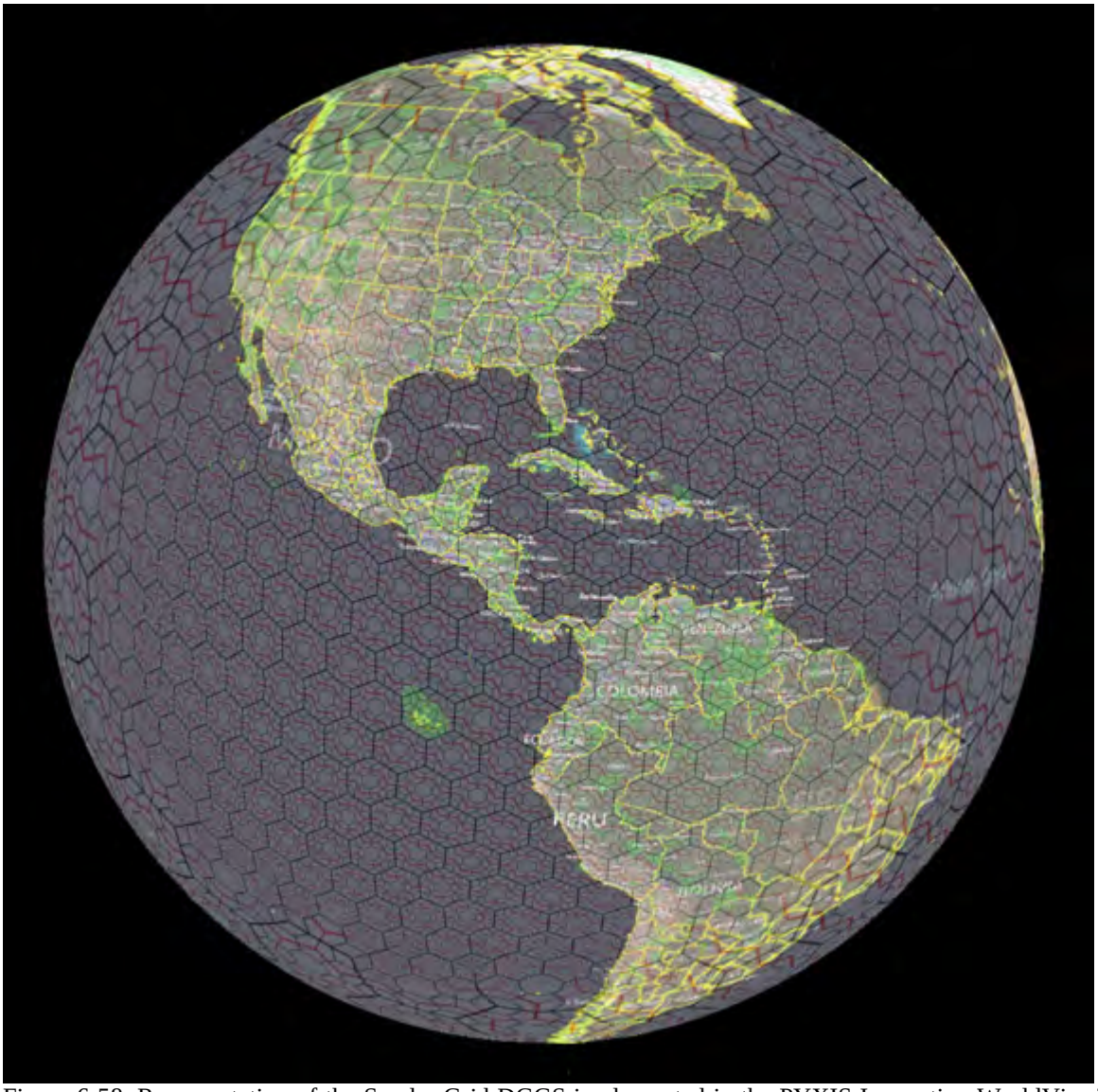


Figure 6-58. Representation of the Snyder Grid DGGS implemented in the PYXIS Innovation WorldView™ client application.

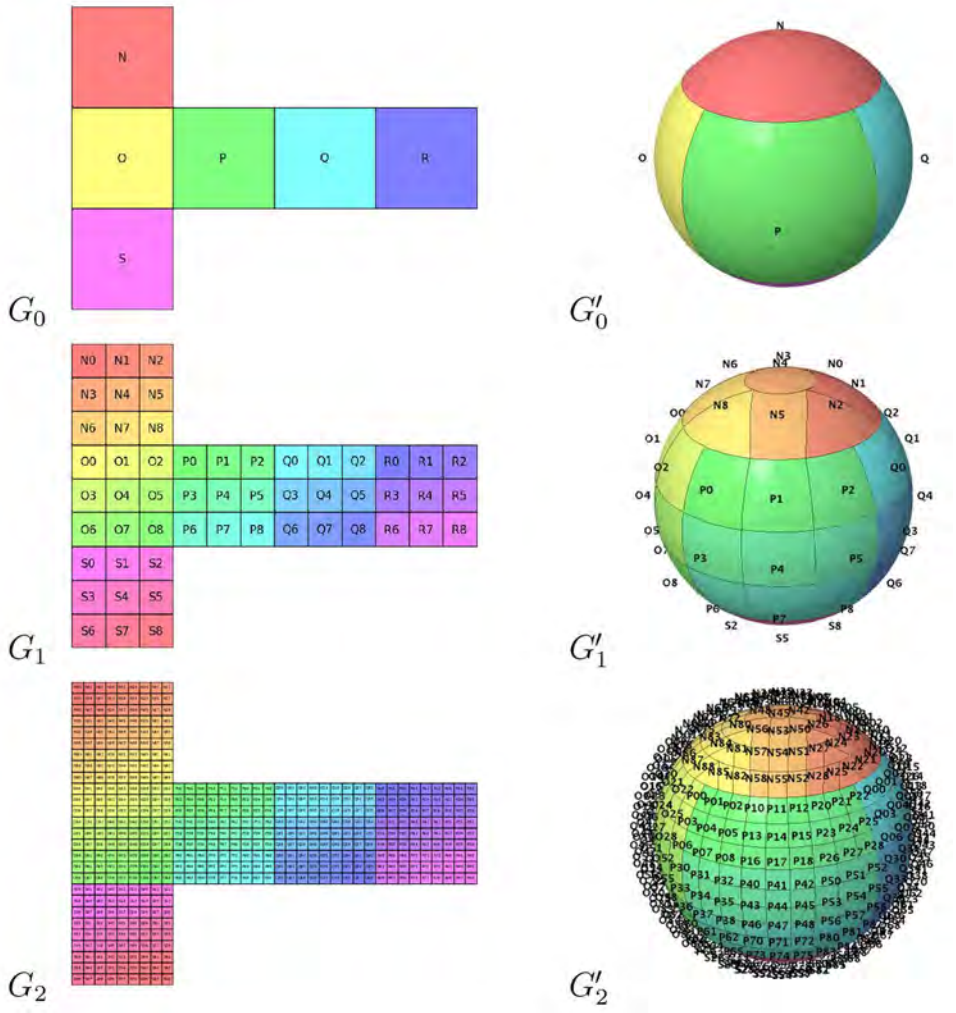


Figure 6-59. The first three planar and ellipsoidal grids of the (0,0)-HEALPix map projection of SCENZ-Grid with Nside = 3 and cells labeled with their IDs (Gibb *et al.*, 2013).

This variety in DGGS technologies and approaches presents both a challenge and an opportunity for the creation of a new global network of grid systems; rather than through the enforcement of a single DGGS, the development of open standards to enable interoperability through the use of different DGGSs appears the best approach. The DGGS SWG aims to increase awareness of the advantages of DGGSs in general to define the qualities of a DGGS; to make them interoperable with conventional and other DGGS data sources; and to standardize operations on them. The DGGS SWG will develop a version 1.0 OGC implementation standard that includes:

- A concise definition of DGGS as a spatial reference system.
- The essential properties of a conformant DGGS.
- The variability within these properties that classify types of DGGS.
- Elements of a spatial reference system identifier suitable for registering specific implementations of a DGGS.

It will also specify interoperable protocols within the standard or through extension documents to develop sample implementations.

4.1.3 *Acknowledgement*

Section 4.1 has been published with the permission of the Chief Executive Officer, Geoscience Australia. Development of the AGDC has relied on support and input from the following people and organizations:

- Dr. Stuart Minchin, Dr. Adam Lewis and Dr. Lesley Wyborn for their strategic vision, energy and foresight that has helped to drive scientific development behind the AGDC.
- The National Earth and Marine Observations (NEMO) Group at Geoscience Australia (GA) for delivering the science and technical innovation necessary.
- The International Forest Carbon Initiative provided the necessary technological baseline for Unlocking the Landsat Archive (ULA) project to be delivered.
- The members of the ULA Project were made possible through the financial contribution of the Australian Space Research Program: in particular, Lockheed Martin Australia; Geoscience Australia; the Victorian Partnership for Advanced Computing (VPAC); the Co-operative Research Centre for Spatial Information (CRC-SI); and, the National Computational Infrastructure (NCI).
- The Commonwealth Scientific and Industrial Research Organization (CSIRO);
- Landsat data were supplied by the U. S. Geological Survey (USGS) and is distributed by Geoscience Australia under the Creative Commons Attribution 3.0 Australia License; and,
- The long term acquisition plan supported by the Landsat program in combination with Geoscience Australia's longstanding investment in the on-ground infrastructure required to capture Landsat imagery.

Implementation of FC25 was made possible by new scientific and technical capabilities, the collaborative framework established by the Terrestrial Ecosystem Research Network (TERN) through the National Collaborative Research Infrastructure Strategy (NCRIS), and the leadership and capabilities of Geoscience Australia and the Joint Remote Sensing Research Program.

The effort, diligence and constructive advice from technical reviewers of Section 4.1 – in particular by Dr. Brendan Brooke, Dr. Medhavy Thankappan and Mr. Jeff Kingwell are greatly appreciated.

4.2 *The New Mexico Resource Geographic Information System (NM RGIS) - Development and Evolution of a Geospatially Enabled Data Management System*

4.2.1 *Introduction*

In 1990 New Mexico recognized the value of having a centralized resource within the state for accessing key geospatial data assets. With that recognition, the New Mexico Resource Geographic Information System (RGIS) program was established at the *Earth Data Analysis Center* (EDAC), then known as the *Technology Application Center* (TAC) at the University of New Mexico. Throughout the 1990s geospatial professionals could order pre-made disks and custom collections of data from RGIS, initially by phone, FAX, email, mail and by visiting EDAC's offices. The emergence of the web as an information access platform provided an opportunity to streamline access and discovery of RGIS' data holdings, and in 1996 EDAC released the RGIS clearinghouse website which included an online data order form, a catalog of available datasets, contact information for ordering data, and links to metadata search tools provided as part of RGIS'

participation in the US National Spatial Data Infrastructure (NSDI) program sponsored by the Federal Geographic Data Committee (FGDC).

In 2001 development of an updated web interface for RGIS was started, and shortly thereafter the new web site was rolled out, with a dynamically generated catalog of datasets available for immediate download. Additional capabilities were developed as the RGIS data collection evolved and grew over the succeeding 14 years. These capabilities included the server platform and interface which provided expanded search and discovery tools, expanded methods and options for accessing the data holdings, and the development of capabilities for the RGIS system to interoperate with other applications through support for open standards defined by organizations such as the Open Geospatial Consortium (OGC) (http://www.opengeospatial.org/), the World Wide Web Consortium (http://www.w3.org/), the Federal Geographic Data Committee (https://www.fgdc.gov/), and the International Organization for Standardization (ISO) (http://www.iso.org/iso/home.html).

The current RGIS system consists of a web services-oriented architecture that provides three distinct tiers: (1) data management; (2) web services; and (3) client. Each of these horizontal tiers interacts only with those adjacent to it in the stacked model, represented in Figure 6-60. This separation between tiers provides a degree of flexibility when expanding the capabilities of the data management, service, and client tiers. While this tiered model has been developed explicitly for the current system, earlier platforms for RGIS have embodied more general components in which capabilities have been provided by separately running systems, seamlessly integrated into the RGIS user interface. Section 4.3.2 outlines the progressive development of the RGIS system, the driving factors that influenced its evolution, and its current platform characteristics and uses.

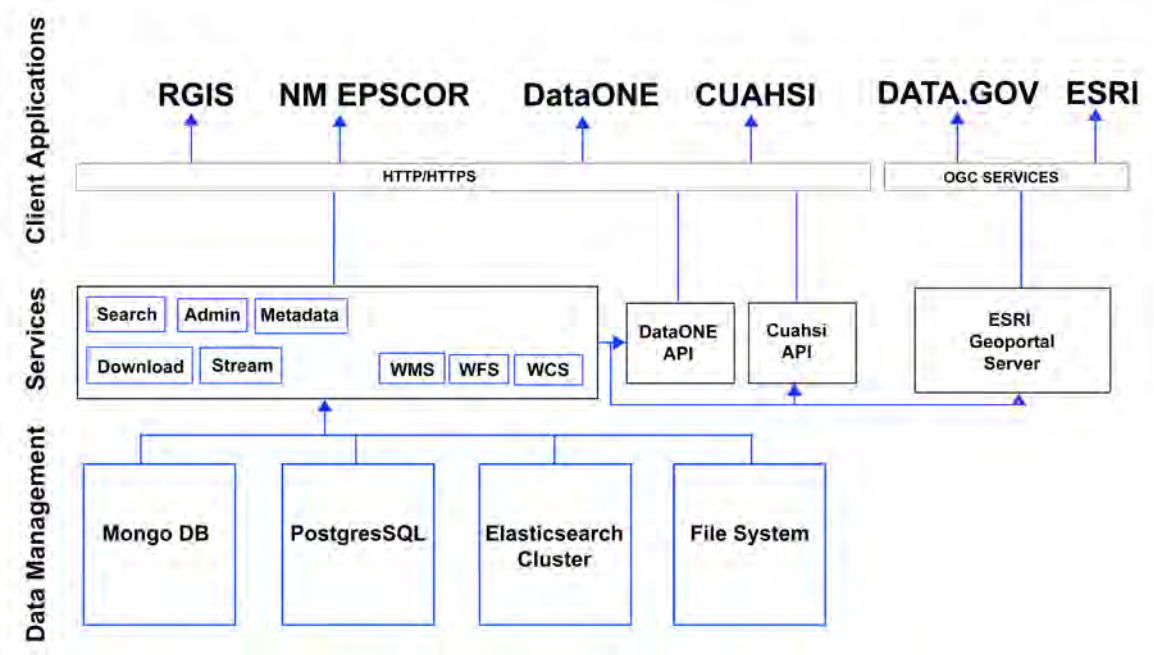


Figure 6-60. Service Oriented Architecture.

4.2.2 *Historical Development of RGIS*

The development of RGIS has its roots in EDAC's focus on making geospatial data as available as possible to the broadest population of users. Over the years this overarching programmatic goal has had a lasting influence on the decisions made about the architecture of the system and the prioritization of alternative capabilities that may be developed.

4.2.2.1 Roots in National Spatial Data Infrastructure Development

While establishment of the RGIS program in 1990 predates the US National Spatial Data Infrastructure (Clinton 1993) by three years, the focus of RGIS on data and documentation interoperability through integration of metadata, data, and visualization standards was well established in 1996 when RGIS implemented an early node in the FGDC's geospatial data clearinghouse system; a key component of the emerging US NSDI. Through funding from the FGDC Cooperative Agreement Program (CAP), RGIS established its clearinghouse node based upon FGDC *Content Standard for Digital Geospatial Metadata* (CSDGM, FGDC 1998) for over 2000 datasets available from the RGIS program.

In 2001 EDAC received an FGDC CAP grant to begin developing capacity for deploying OGC Web Map Services (WMS, de La Beaujardière, 2002) based on data holdings in the RGIS program. At about the same time, development of a searchable, interactive catalog of RGIS data holdings was started, with staged releases during 2001 and 2002. The new web catalog allowed users to directly access and download data holdings in the RGIS program without submitting an order form, waiting for the delivery of data on physical media, or paying a service fee associated with their request. Capability was based upon a database-driven web site that dynamically searched for data meeting basic search criteria, and provided links to cached data packages that included metadata for immediate download. Throughout this period the FGDC clearinghouse node established in the mid-1990s continued to provide the primary metadata search capability for data within the RGIS system. As improved access methods were developed for the RGIS data product (i.e. the direct downloads available through the web site) metadata upon which the clearinghouse node was based became outdated.

In 2003 RGIS transitioned from the FGDC clearinghouse node established in 1996 to contributing metadata for the early version of the NSDI's *Geospatial One-Stop* (GOS) portal. The metadata collection used to populate GOS was the same one used by the FGDC clearinghouse node with little modification or expansion to the direct access methods available through the RGIS web site. In many respects this metadata collection (still based upon the FGDC CSDGM) could be considered a legacy metadata resource that enabled rapid assimilation into GOS to support early capability development without providing all of the documentation features that GOS or other current metadata search platforms could support.

Until 2005 the focus in RGIS had been on the slow accumulation of individual datasets and their associated metadata. In 2005-2006 the process changed significantly with the acquisition of a new digital orthophotography collection (1m resolution) for the entire state of New Mexico. This collection was based upon a digital acquisition of four-band imagery, and when processed produced over 17,000 RGB and CIR digital ortho-photo quarter quadrangles (DOQQ) that increased the RGIS data collection by a factor of four in the course of two years. Metadata for these DOQQs were automatically generated by the contractor that produced the image files. This large metadata collection provided a degree of detail and consistency not

previously available for RGIS data. Given anticipated high demand for these data, the RGIS system's capabilities were expanded to include automated packaging of data and metadata, a request queueing system for handling bulk data requests, automatic generation of OGC WMS services for all of the new DOQQs, and enhanced metadata views that included live data previews based upon the generated WMS services for each dataset. In spite of these enhancements, the metadata-based search capabilities provided by GOS were still separate from the basic search capabilities supported by the RGIS web site.

Following over a year of development and ten years after the initial release of RGIS' online catalog, in 2011 a new RGIS website and underlying server platform (named the *Geographic Storage, Transformation and Retrieval Engine* – GSToRE (http://gstore.unm.edu)) was released as a product of combined support from the RGIS program and the National Science Foundation's EPSCoR Program. The key objectives of the development of the new platform and associated web interface included:

- An explicit RESTful (Fielding 2000) web services model through which client applications, including the RGIS web interface, access the available data and metadata within the system;
- Support for multiple OGC service standards for data access and visualization (Web Map, Web Feature and Web Coverage Services, WMS, WFS and WCS respectively) (de La Beaujardière 2006; Vretanos 2005; Whiteside 2006);
- Support for multiple metadata standards, starting with FGDC CSDGM and ISO 19115 (and related standards), but being extensible to support additional metadata representations as appropriate for evolving application and data requirements;
- Support for non-geospatial data and appropriate metadata within the system services that streamline exposure (through inclusion in standards-based metadata) of available data and visualization services in external systems such as Data.gov (http://catalog.data.gov/organization/edac-unm-edu) and DataONE (https://www.dataone.org/).

The expanding list of OGC, other web services and metadata standards included in the specification for the GSToRE platform align with continued support for RGIS and other research data management projects, participation in the US NSDI and the Global Spatial Data Infrastructure (GSDI), and a logical extension of the 20+ year evolution of the RGIS program as an interoperable component within the larger US NSDI network.

4.2.3 *Evolving Data Volume, Velocity and Variety as Drivers for RGIS Development*

While the NSDI capabilities and services described in the previous section provide a foundational requirement set for the capabilities of RGIS and GSToRE, other aspects of the flow of data into the system have driven the technical development of the system. Specifically, a combination of increasing data volume, velocity, and variety since 2001 combined with continuously changing models for client and user interaction with web platforms have led to periodic updates to the RGIS server platform and corresponding web interface (Figure 6-61).

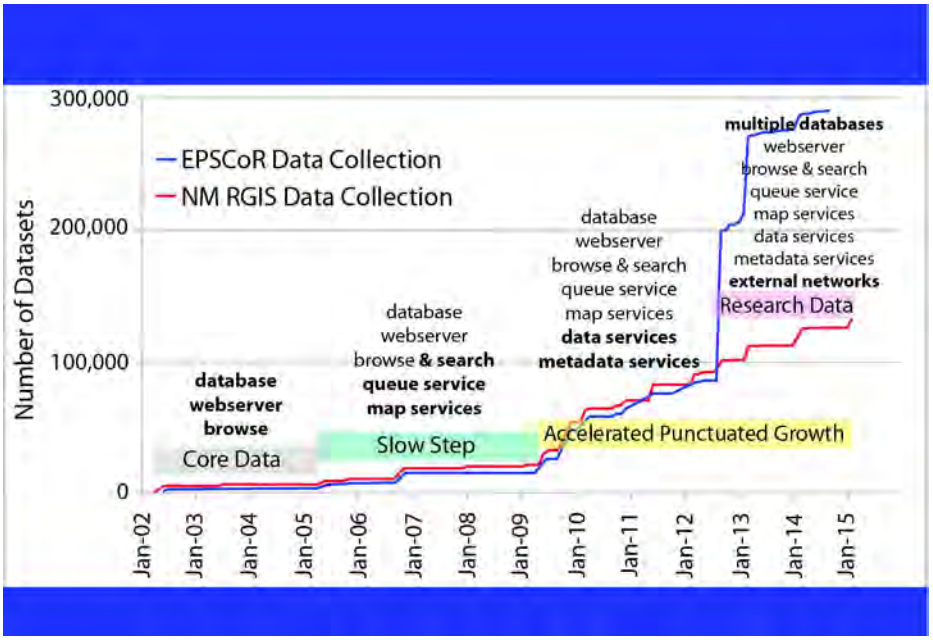


Figure 6-61. Advances in technical developments and increases in data volume for the RGIS server platform.

The number of datasets included in the core data management system upon which RGIS is based has gone through four qualitatively different phases between 2002 and 2015. In 2012 research data products associated with NM EPSCoR were added to GSToRE. Both RGIS and EPSCoR's data discovery and access interfaces are described above as part of the NSDI integration discussion.

The following phases are identified in the figure:

- Core Data are the initial datasets that were added to the system when it was first brought online in 2001, and are maintained as the foundation data for RGIS. Additional data were added slowly to the system in small groups or individually.
- Slow Step - the addition of two large collections of digital ortho-photos to RGIS yielded a four-fold increase in the number of datasets within the system punctuated by relatively slow growth similar to that observed during the preceding core data phase.
- Accelerated Punctuated Growth - during this phase diverse collections of data are periodically added to the system at an increasing rate, yielding a much faster rate of increase in both the types and number of data managed within the system. Development of the GSToRE platform began at this time.
- Research Data - the research data phase overlaps the previous phase and is specifically related to the rapid integration of datasets and associated materials associated with, and in support of, the NM EPSCoR program.

4.2.4 *Current RGIS Platform and Capabilities*

4.2.4.1 Tiered Architectural Model

The RGIS application and underlying GSToRE platform, developed at EDAC, comprises a services-oriented architecture with distinct data management, services, and client tiers. Each of these tiers is optimized for its unique functional role and only interacts with an adjacent tier. For example, the data

management tier can be accessed only through the services tier, and clients interact only with the services tier to access data managed in the data management tier. The data management tier includes multiple databases, including a geospatially enabled PostgreSQL/PostGIS Object Relational database, a JSON document-based MongoDB database, and Elasticsearch for metadata search. Additionally, the system includes file-based storage of data that are more appropriately stored outside of the database management systems. The services tier provides a set of RESTful Web services with which client applications (i.e., web applications, desktop GIS, analytic tools) interact to discover and access the data held in the system. Web services designed thus far include ingest and delivery of scientific data, search and retrieval capabilities, and data transformation services. These services include OGC Web Map, Web Coverage and Web Feature Services that are automatically generated through the REST web services provided by GSToRE.

- Data Management Tier

The data management tier includes multiple databases, including a geospatially enabled PostgreSQL/PostGIS Object Relational database, a JSON document-based MongoDB database, and Elastic search for metadata search. Each database provides specific data management services as illustrated in Figure 6-62.

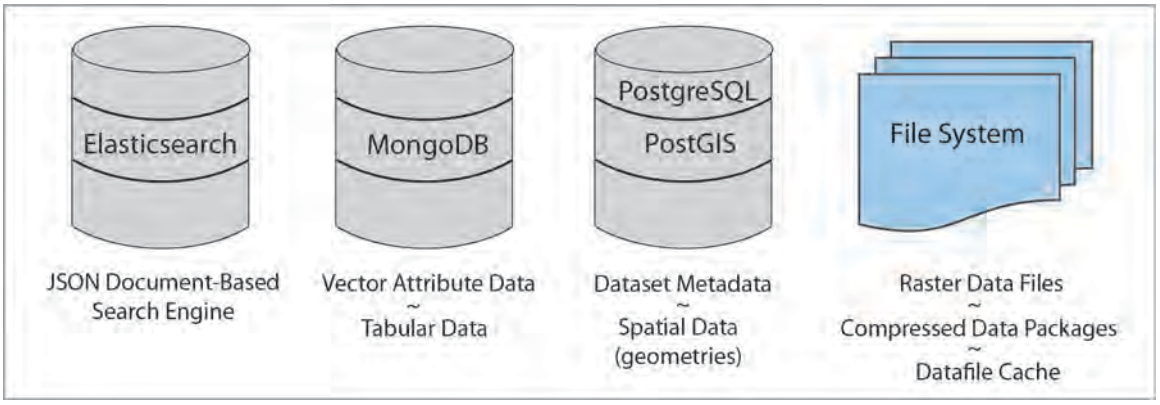


Figure 6-62. GSToRE database components and their roles in the system.

- Web Services – Application Programming Interface (API)

The services tier provides a set of RESTful Web services with which client applications (i.e., web applications, desktop GIS, analytic tools) interact to discover and access the data held in the system (Figure 6-63). Web services support a wide variety of scientific data and include ingest, search, data format transformation, map visualization (through OGC WMS), data extraction and delivery (through OGC WFS, WCS and general REST requests), and metadata delivery.

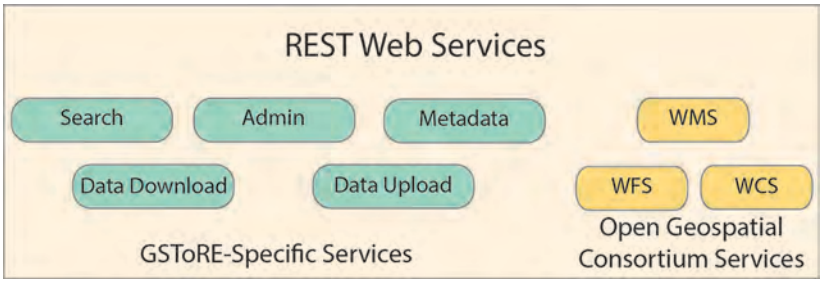


Figure 6-63. Classes of web services available through the GSToRE platform.

- Client Applications

Client applications include web applications, desktop applications, and external systems that interact with the GSToRE web services to harvest metadata and data. The RGIS web interface (Figure 6-64) is an example of a web application that uses the data discovery and access services of GSToRE to provide the content presented through the RGIS user interface to users. The implementation of support for the DataONE Tier 1 Member Node API is an illustration of the integration of GSToRE content into external systems through metadata harvest. The published GSToRE API (http://gstore.unm.edu) is the foundation upon which diverse applications can be built, both in direct support of EDAC projects and by external developers who would like to leverage the data and metadata content of GSToRE.

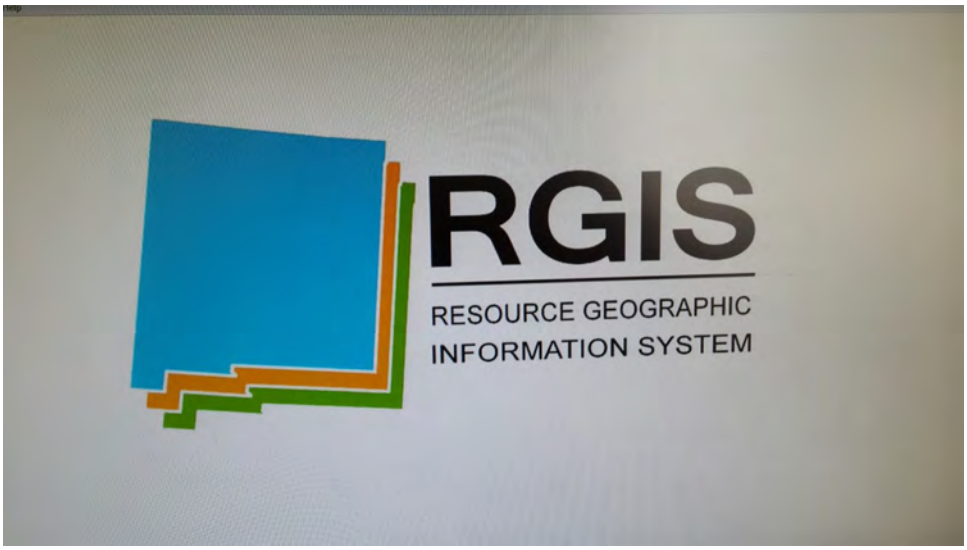


Figure 6-64. RGIS data search and browse

4.2.4.2 Ongoing Development and Integration into New Applications

When RGIS began its development, commercially available geospatial platforms did not exist yet. This led to the development of a custom geospatial data management discovery and access system to support the RGIS program requirements (GSToRE). This environment has changed recently with several commercial vendors now providing robust, web-based geospatial systems. These systems provide turn-key, user friendly interfaces and tools for managing and interacting with their content. As a result the RGIS portal is transitioning to the Hexagon Apollo Geospatial Portal (http://www.hexagongeospatial.com/products/provider-suite/erdas-apollo) system, which provides specialized support for the core geospatial requirements of the RGIS program and its users.

This continued focus on geospatial data and related services has allowed the RGIS program to leverage emerging commercial offerings in support of geospatial applications. The increasing need for robust support for more generalized research data management systems continues to drive the development and expansion of the capabilities of the GSToRE platform beyond the geospatial focus. While the RGIS program and portal are moving onto a commercial platform, the need for a customized and flexible web-enabled research data management system remains, particularly in support of ongoing and proposed data intensive research projects sponsored by the US National Science Foundation and others. This need is being met through continued development of the GSToRE platform.

Developers are automating the process of uploading researcher data into GSToRE and having their data available for other regional and national data warehouses to discover and access. This work includes implementation of the tier-4 member node DataONE for research metadata ingest and data replication for GSToRE assets, publishing metadata for harvest by Data.Gov, and development of workflows for integration into long-term data preservation systems such as UNM's LoboVault (http://repository.unm.edu).

The open source architecture used to develop GSToRE makes the platform very cost effective and flexible for other applications to build upon. An example is the NSF NM EPSCoR Track 2 Virtual Watershed Project. This project required a data warehouse capable of storing large researcher data sets, the ability to add data and metadata rapidly to the system for discovery and reuse, and support for other research teams to access data and documentation in support of their work. The GSToRE platform was a strong foundation upon which the Virtual Watershed system could be built. It had all the components needed from the storage of researcher data, to fast search of data and numerous web services that allow researchers to pull data from the system. By simply making a clone of the GSToRE platform, the Virtual Watershed developers were months ahead of schedule.

4.2.5 *Conclusions*

In the 25 years since initiating the RGIS program, the technical approach for providing access to geospatial resources that are the program's focus has evolved, not only as technology has changed but also as user expectations and needs have changed. Starting as a tightly integrated web-based data download system in 2001, and evolving over the subsequent 15 years into a multi-component web-enabled tiered architecture has allowed for rapid development of new capabilities and the flexibility to integrate new technologies as they become available and mature. The ongoing development of the RGIS portal and the GSToRE platform exemplify this flexibility and focus on continuous assessment and growth in how the systems that are used can continue to meet the expanding needs of the geospatial data and applications community that RGIS serves and the expanding research community that is using the capabilities of GSToRE to meet the increasingly stringent data management and sharing requirements of their projects and sponsors.

4.3 *DigitalGlobe Online Access to Remote Sensing Data*

4.3.1 *Introduction*

With increased collection capabilities of satellites that can collect millions of square kilometers of imagery and refresh the earth on a frequent basis, it is paramount for the users to have online access to the image catalogue, associated metadata, view the images as well as run analysis on the images. Remote sensing industry has taken advantages in cloud computing, open protocols and standards, as well as latest image processing, and analysis techniques to make imagery available online and on demand. The following sections discuss how imagery can be accessed online from various sources and the technology behind them.

4.3.1.1 Imagery Catalogue

Imagery catalogue represents raw imagery collected as well as processed imagery available online. Raw imagery catalogue will list of all the imagery collected by a data source, including all the imagery archive. In the case of Landsat, the image catalog dates back all the way to 1970’s, when the imagery was made

available commercially. Imagery is typically represented as scenes, with scene width and length depends on the sensor. Landsat has developed a fixed row/path grid scheme where users can search for imagery based on this scheme. Different providers use varying schemes to represent the image extents.

Online imagery catalogue is typically provided using a web interface that allows users to query based on various metadata parameters including sensor type, collection dates, sensor collection angles or off-nadir viewing angles, cloud cover, and others. Users can define their area of interest (AOI) using multiple options ranging from drawing a polygon, defining a centroid and a bounding box or circle, uploading AOI's as shapefiles or other industry standard formats. Further, users are also given the option of using a gazetteer to navigate to an area of interest or use web tools for zooming in and selecting their area of interest.

Typical online interfaces display queries that list images available for the AOI, that meet the user selected criteria. Some imagery providers have links to reduced resolution/browse images that allow users to determine if the imagery on their AOI is usable or not. The query results typically display images that also partially intersect the user's AOI.

Most of the online technologies use industry accepted software or protocols for access to the imagery catalogues. OGC standards for imagery and feature display, WFS and WMS, and typically employed for user interface as well as downloading the catalog. Few of the catalogues have also been exposed within Google earth engine in KMZ/KML formats.

4.3.1.2 Online Access to Imagery

Users are now able to view online imagery from variety of online platforms including google maps, bing maps, AGOL, Open standard web interfaces and others such as DigitalGlobe Cloud Services. This imagery is typically has been processed and is GIS ready and can be integrated into a customer GIS or webAPI that is based on maps. Google and Bing maps have introduced the concept of zoom levels where different levels can be derived from the same or multiple imagery sources. An example would be at zoom levels 13 and lower, Landsat imagery is displayed and from zoom levels 14 and higher, the high resolution imagery is resampled into multiple resolutions/zoom levels (zoom level 19-0.3 meters, 18–0.6 meters, 17–1.2 meters and so on). Users now also have an option the view imagery in 2D or 3D where imagery is typically overlaid or draped onto a Digital Terrain Model (DTM).

Some of the web viewing software has built in functionality to re-project imagery between projections to ensure display of multiple images at the same. Using some of the web tools, users can view imagery from tow dates with a slide bar that can display before and after images of an area of interest.

Users can download imagery online and on demand. The online image access will allow users to download both raw imagery as well as processed imagery. The raw imagery and associates spectral bands, can be downloaded in a GIS ready format or for ingest into platforms such as Google Earth. Several data compression and delivery acceleration techniques allow for faster download of imagery, even in internet band width challenged locations.

One of the commonly used technologies is OGC WCS protocol for downloading the selected imagery. There are several proprietary techniques that allow users to download imagery.

Users are now able to do variety of image processing and analysis tasks such as pan sharpening, extraction of vegetation indices such as NDVI, supervised classification, and other tasks. Further, online imagery

can also be now viewed as a False-color composite (FCC) where users can pick pre-defined color schemes or create their own band selections for display, instead of traditional natural color RGB composite.

4.4 *Coverages Uncovered: Agile Ad-Hoc Analytics on Spatio-Temporal Data Cubes*

4.4.1 *Introduction*

With the unprecedented increase of orbital sensor, *in-situ* measurement, and simulation data, as well as their derived products, there is an immense potential for getting new and timely insights - yet, their value has not so far been leveraged fully. Incidentally, such spatio-temporal sensor, image, simulation, and statistics data in practice typically constitute prime Big Data contributors. In view of such data too big to transport, the quest is on for high-level service interfaces for dissecting datasets and rejoining them with other datasets to allow users to ask any question, anytime, of any size, and enabling them to build their own product(s) on-the-fly.

The notion of coverages has proven instrumental in unifying regular and irregular grids, point clouds, and meshes so that such data can be accessed and processed through a simple, yet flexible and interoperable, service paradigm. The OGC Web Coverage Service (WCS) comprises a modular suite for accessing large coverage assets and establishes verifiable interoperability. WCS Core provides simple data sub-setting and encoding, whereas extensions add optional service facets up to *ad-hoc* spatio-temporal filtering and processing on massive datacubes. In computer programming contexts, a data cube (or datacube) is a three-dimensional (or higher) array of values, commonly used to describe a time series of image data. A data cube is also used in imaging spectroscopy, since a spectrally-resolved image is represented as a three-dimensional volume (cited from *https://en.wikipedia.org/wiki/Data_cube)*. The latter is accomplished by OGC's Big Earth Data query language, Web Coverage Processing Service (WCPS). By separating coverage data and service model, any service - such as WMS, WFS, SOS and WPS - can provide and consume coverages in addition to WCS, thereby enabling heterogeneous mashups with lossless information flow.

Big Data readiness of the OGC datacube standards becomes obvious when looking at the intercontinentally federated EarthServer initiative (Figure 6-65) (The EGU General Assembly, 2016). In its Phase 1, 2011 through 2014, EarthServer established common comprehension of all coverage types. Independent reviewers concluded after Phase 1 that the project has been shaping the Big Earth Data landscape through standardization activities within OGC, ISO and beyond. The underlying rasdaman Array DBMS (Baumann, 1994) will transform the way scientists in different fields of Earth Science will be able to access and use data in ways that were hitherto not possible. In Phase 2 from 2015 through 2018, EarthServer is focusing on massive datacubes, crossing the Petabyte frontier on satellite and climate data.

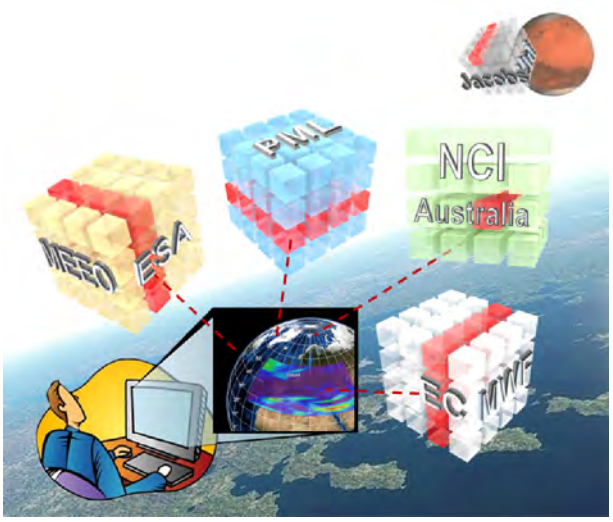


Figure 6-65. EarthServer Datacube Federation.

As a side note, EarthServer has also produced educational material with its 1-hour TV documentary "Big Earth Data – the digitized planet" (https://www.youtube.com/watch?v=ldsTcAvXwSc) which explains the volume, velocity, variety, and veracity of Earth data to society at large through real-life examples. The documentary has been broadcast by European TV stations four times in 2015 and is available in English, German, and French on YouTube (Baumgarten *et al.*, 2014).

This Section provides an overview of OGC datacube standards and their use in EarthServer, with particular emphasis on their implementation. In Section 4.4.2, OGC datacube standards are introduced. The underlying technology platform is presented in Section 4.4.3 and Section 4.4.4 addresses the services established. Section 4.4.5 concludes the contribution.

4.4.2 *Earth Datacube Standards*

The OGC Web Coverage (WCS) suite of standards arguably offers the most comprehensive functionality on spatio-temporal coverages, crafted along the coverage model. We will look at both data and service model. The normative documents of the coverage standards are available at OGC 2016, sample services are described by (rasdaman, 2016a), and a tutorial can be downloaded from (rasdaman, 2016b). Rasdaman is an array database management system for storing and retrieving massive, multi-dimensional sensors, images, and statistical data. It has no limits on the number of dimensions it can serve (see https://en.wikipedia.org/wiki/Rasdaman).

4.4.2.1 Coverage Data

The notion of *coverages* is defined in ISO 19123 (identical to OGC Abstract Topic 6) in an abstract way, and described in more detail in the OGC Coverage Implementation Scheme (CIS) which ISO intends to adopt as ISO 19123-2. In an abstract sense, *a coverage* is a function mapping points to values, describing the digital representation of some space/time varying phenomenon. In practice, this reduces to regular and irregular grids, point clouds, and general meshes.

While the abstract model allows many divergent and incompatible implementations, CIS 1.1, with its concrete data structure definitions and format mappings, is interoperable and conformance testable to the

pixel level. Relying on ISO 19123, CIS follows the functional approach and defines coverages as having a *domain set* that describes the so-called *direct positions* for which values are stored, and the *range set* containing these values (Figure 6-66). As the range values can be arbitrary (such as hyperspectral pixels, elevation, or land use), they need a description. This is given in the *range type*. In addition, for practical reasons, a metadata record has been added which can carry any kind of data an application may want to associate with the coverage (such as an ISO 19115 catalog entry).

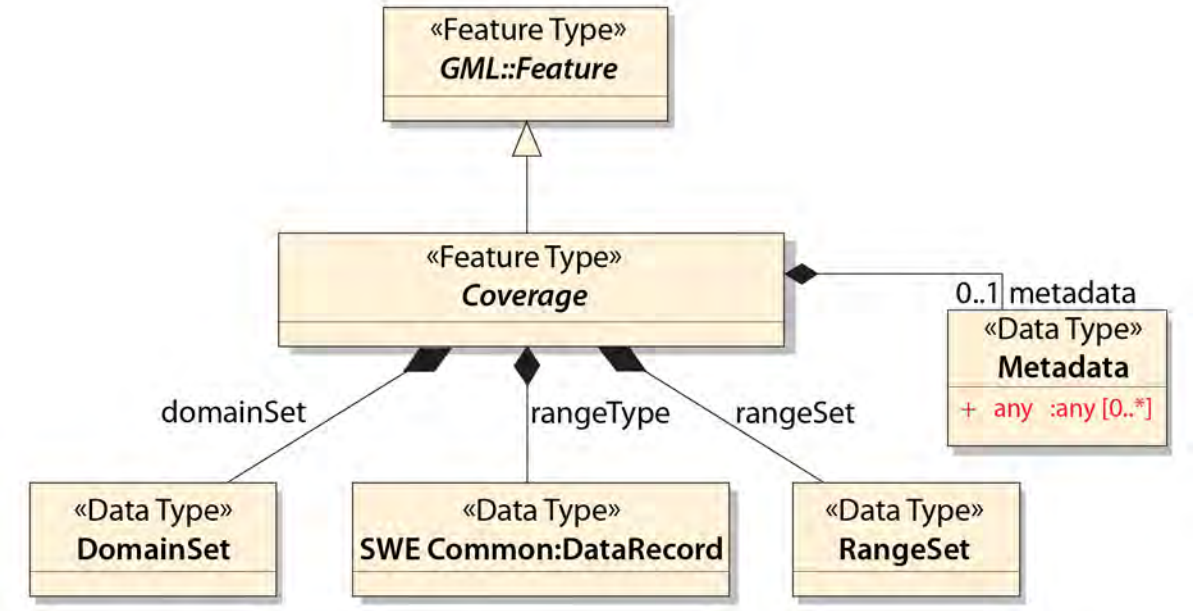


Figure 6-66. Basic conceptual coverage structure, as per CIS 1.1 (Baumann *et al.*, 2016).

In previous gridded raster data versions, CIS followed the original GML approach to define three types of grids: *Grid* (non-referenced rasters), *RectifiedGrid* (geo-referenced grids with straight lines and constant spacing), and *ReferenceableGrid* (all else). In particular, the last type is so general that handling is not straightforward. CIS 1.1, therefore, introduces a new type, *GeneralGrid*, that introduces various types of axes that can be compiled into any particular grid type. For example, a mix becomes possible, for example, as a satellite image time-series with two regular horizontal axes (Lat. and Lon.) and one irregular temporal axis. In the most general case, a General Grid may be defined algorithmically, e.g., through a SensorML model.

Coverage data can be modeled and encoded in various ways. The standard provides three informationally complete ASCII representations: GML, JSON, and RDF (based on JSON-LD). Further, any other – typically binary – format is also possible, such as GeoTIFF, NetCDF, and JPEG2000. As often both advantages (informational completeness and storage efficiency) are to be combined, *container formats* like multipart/MIME, zip, SAFE, GMLJP2, etc. are introduced that are capable holding a coverage consisting of several parts. The first part is GML, JSON, or RDF; while further parts can consist of any format, represented as a single file or several files, such as to represent tiling (Furtado and Baumann 1999).

4.4.2.2 Coverage Retrieval Service

Coverages as per CIS can be served through many services, such as a Web Feature Service (WFS); however, these can handle coverages only as "black boxes," whereas the Web Coverage Service (WCS) suite offers a variety of operations.

In the WCS Core (Baumann, 2012, Baumann, 2010b), only sub-setting and format encoding is established, keeping it tentatively simplistic. Sub-setting can mean *trimming*, where bounding coordinates are specified that reduce the coverage, but retain its dimension – such as a 2D cutout from a 2D map. *Slicing*, on the other hand, reduces the number of dimensions by extracting a slice at a particular position, like a particular image time-slice in an image time-series cube. Below is an example of a request retrieving a horizontal slice from a time-series datacube `c001`, returning it encoded in GeoTIFF:

http://www.acme.com/wcs ? SERVICE=WCS & VERSION=2.0
& REQUEST=GetCoverage & COVERAGEID=c001
& SUBSET=time("2009-11-06T23:20:52")
& FORMAT="image/tiff"

WCS Extensions add further functionality, such as:

- *Range sub-setting*, which defines extraction of "bands" or "variables" from complex range values;
- *Scaling*, where the resolution of gridded coverages can be reduced;
- *CRS*, that allows one to retrieve a coverage in a different Coordinate Reference System than the one in which it was stored (its so-called Native CRS); and,
- *Interpolation*, which allows clients to control interpolation applied in the server whenever interpolation occurs, such as in scaling or re-projection.

All these operations can be combined into a single request, allowing versatile data extraction. Several semantically equivalent protocol bindings are standardized, including GET/KVP, POST/XML, and SOAP. REST is under preparation, but usually leads to opposing opinions on whether slashes or ampersands are used. The most general extension enabling spatio-temporal datacube analytics is WCPS which is described in Section 4.4.2.3. A practical and important extension is discussed in Section 4.4.2.4.

4.4.2.3 Coverage Processing Service

OGC Web Coverage Processing Service (WCPS) is a geo raster query language: a client sends a WCPS query (i.e., a string) and gets back one or more coverages (or scalar values, in case of an aggregation). The language is high-level enough to allow manifold server-side optimizations and parallelization (see Section 4.4.2.4 for examples). For a comprehensive explanation refer to the standard (Baumann, 2009), concept paper (Baumann, 2010a), the online demonstrations (Baumann, 2016a), and the tutorial (Baumann, 2016b). As can be seen from the comparison and arithmetic operators, syntax is based on multi-dimensional map algebra, in this case: Array Algebra (Baumann, 1999). It offers operations that allow local, focal, zonal, and global operations to be expressed.

Notably, in EarthServer all "datacube" retrieval is done exclusively through WCS and WCPS, plus experiments on retrieval functionality which are being developed for proposing new standard components.

4.4.2.4 Maintaining Data Offerings

Large-scale ingestion of data turned out an issue early on in EarthServer. After some experimentation, a specification was developed which now is an adopted OGC standard named WCS-T (T for "Transaction") (Baumann, 2014). This standard allows Web-based insertion, deletion, and update of *coverages*, including updates of parts of a *coverage*. Coverage data to be uploaded into a server can be provided as part of the request or by reference to some location accessible to the server. In the latter case, the server fetches data by itself; this eases maintenance from a remote (and possibly lightweight) device considerably.

In practice, the rasdaman WCS-T implementation was extended into a complete ETL (*Extract, Transfer, Load*) tool offering a high-level interface for automating maintenance (Baumann *et al.*, 2015), such as standing ingestion workflows for time-series or for establishing datacubes from a large number of input files. A simple configuration file, called a recipe, captures all file locations, conventions (like file naming, accompanying metadata files, and potential conversion rules), and further information relevant for homogenizing. Syntax of such configuration files is based on JSON and well documented so that writing it is not an issue.

Meantime, WCS-T based recipes represent the standard method for ingesting and maintaining geo raster offerings through rasdaman, conveniently used by both developers and service operators. Actually, recipe templates often are exchanged and adopted across rasdaman installations.

4.4.3 *The EarthServer Datacube Technology*

A distinguishing criterion of the EarthServer approach is its flexible query capability, allowing "any query, anytime, on any size". The enabling technology behind this paradigm is rasdaman (raster data manager), a flexible and scalable array database system.

In the first place, rasdaman is domain agnostic and can serve n-D arrays irrespective of their semantics, such as satellite imagery, confocal microscopy imagery, cosmological simulations, or financial analysis data. The geo semantics is added in a specific layer which offers OGC interfaces for WMS, WCS and WCPS (see previous section), allowing for both regular and irregular grids and their re-projection, among others. Figure 6-67 shows a high level rasdaman architecture.

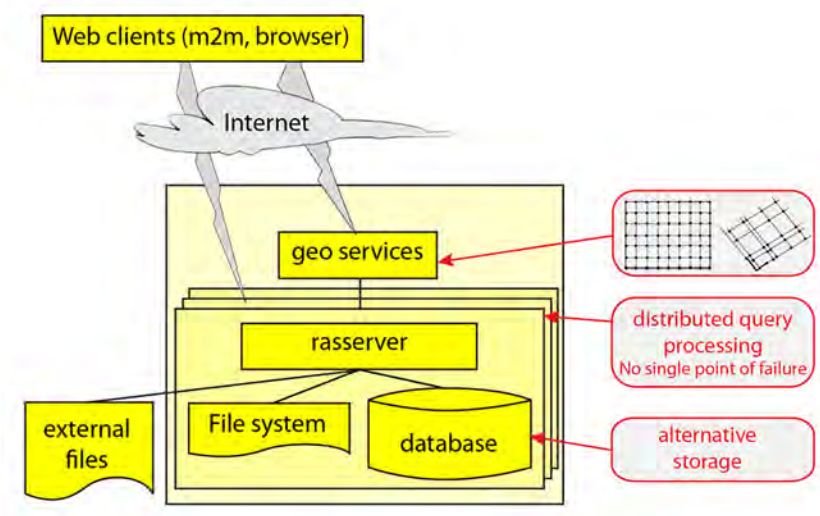


Figure 6-67. rasdaman architecture overview.

4.4.3.1 Array Storage

The rasdaman array database can maintain arrays in either a conventional database (such as PostgreSQL) or its own persistent store directly in any kind of file system. Additionally, rasdaman can tap into "external" files not under its control. Any given archive structure can be registered into rasdaman, including all file and directory naming conventions, auxiliary files, etc., which might hold relevant metadata such as image position, time taken, or sensor type.

A core concept of array storage in rasdaman is partitioning or *tiling*. Arrays are split into sub-arrays called *tiles* to achieve fast access. Tiling policy is a tuning parameter which allows adjusting partitions to any given query workload, measured or anticipated. As this mechanism turned out very powerful for users, a few strategies have been implemented in rasdaman to assist in finding the best tiling (Figure 6-68).

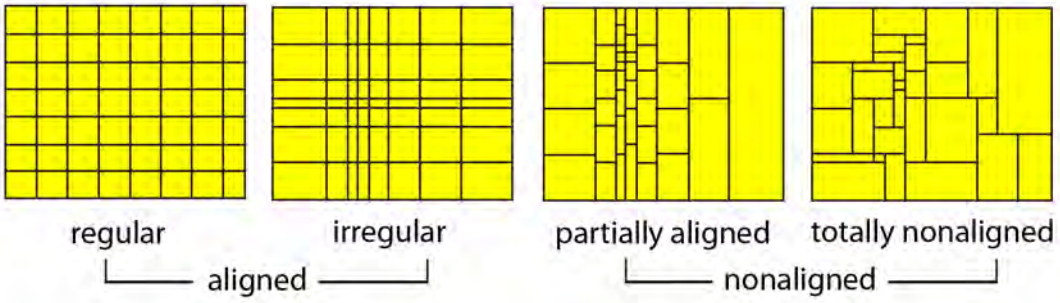


Figure 6-68. Tiling structures supported by rasdaman, per the Furtado classification (Furtado *et al.*, 1999).

In case no particular knowledge about user patterns is on hand, the system will automatically choose a regular partitioning that is adjusted to bus speed etc. of standard hardware architectures. Should there be a preference for particular directions (such as doing mainly time-series analysis) then Directional Tiling can be chosen where only sizing ratios between axes are indicated, such as 10:1:1 on t/x/y when ten times more time-series accesses are anticipated; the absolute sizing etc. will be determined by the system. In the most general case, the administrator only provides a list of – possibly overlapping – hotspots that need particularly fast response, and leaves all the details to the tiling engine (Figure 69). See (Baumann *et al.*, 2010) for a comprehensive overview on the strategies currently available in rasdaman.

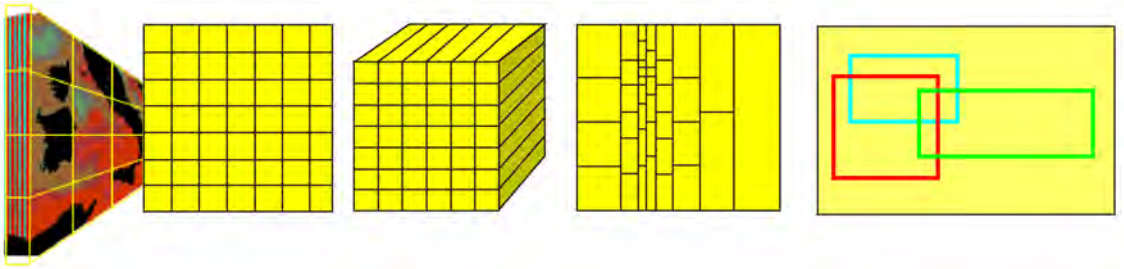


Figure 6-69. Tiling strategies supported by rasdaman.

Is such powerful partitioning necessary? The answer is yes, as the (vector) tiling example of Open Streetmap shows. Adaptive partitioning allows tuning much better towards user access patterns, which result in less data transfer and, ultimately, a better performance experience.

While ingest and automatic formatting into the predefined tiling strategy multi-dimensional optimizes access such import – i.e.: copying – of data is not always feasible on huge archives such as the 87 PB

climate archive of ECMWF. Therefore, linkage to external archives (called *in-situ* data by rasdaman) is indispensable. In this case, files are not merged into the database, but only registered once. Subsequent query processing dispatches automatically so that from a user's perspective there is no difference in query capabilities. Rasdaman milestone 9.3 permits an internal tiling of archive files, such as available with TIFF and NetCDF, that can be exploited for fine-grain reading (Figure 6-70). An automated distribution of tiles based on various criteria, optionally including redundancy, is under development.

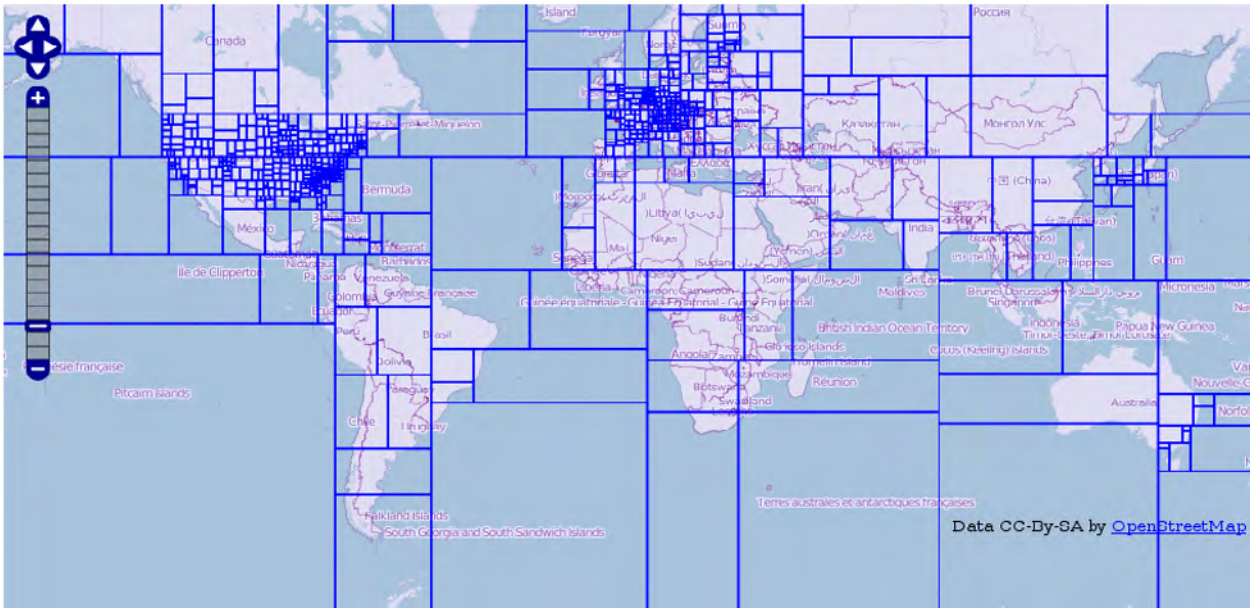


Figure 6-70. Open Streetmap tiling of its global vector networks (OSM 2016).

4.4.3.2 Array Processing

The rasdaman engine has been crafted from scratch, optimizing every single component for array processing. A series of highly effective optimizations is applied to queries, including:

- Query rewriting to find more efficient expressions of the same query; currently 150 rewriting rules are implemented.
- Query result caching is used to keep complete or partial query results in (shared) memory for reuse by subsequent queries; in particular, geographic or temporal overlap can be exploited advantageously.
- Array joins with optimized tile loading so as to minimize multiple loads when combining two arrays; (see Baumann and Merticariu, 2015) where this minimizing has been proven based on graph theory. This is not only effective in a local situation, but in particular when tiles have to be transported between compute nodes or even data centers in case of a distributed join.

After query analysis and optimization, the system fetches only the tiles required for answering the given query. Subsequent processing is highly parallelized. Locally, it assigns tiles to different CPUs and threads. In a cluster, query are split and parallelized across the nodes. The same mechanism is also used for distributing processing across data centers, where data transport becomes a particular issue. To maximize efficiency, rasdaman currently optimizes splitting along two criteria (Figure 6-71):

- Send queries to where the data sit ("shipping code to data").

- Generate subqueries that do as much processing as ever possible locally, minimize the amount of data to be transported between nodes.

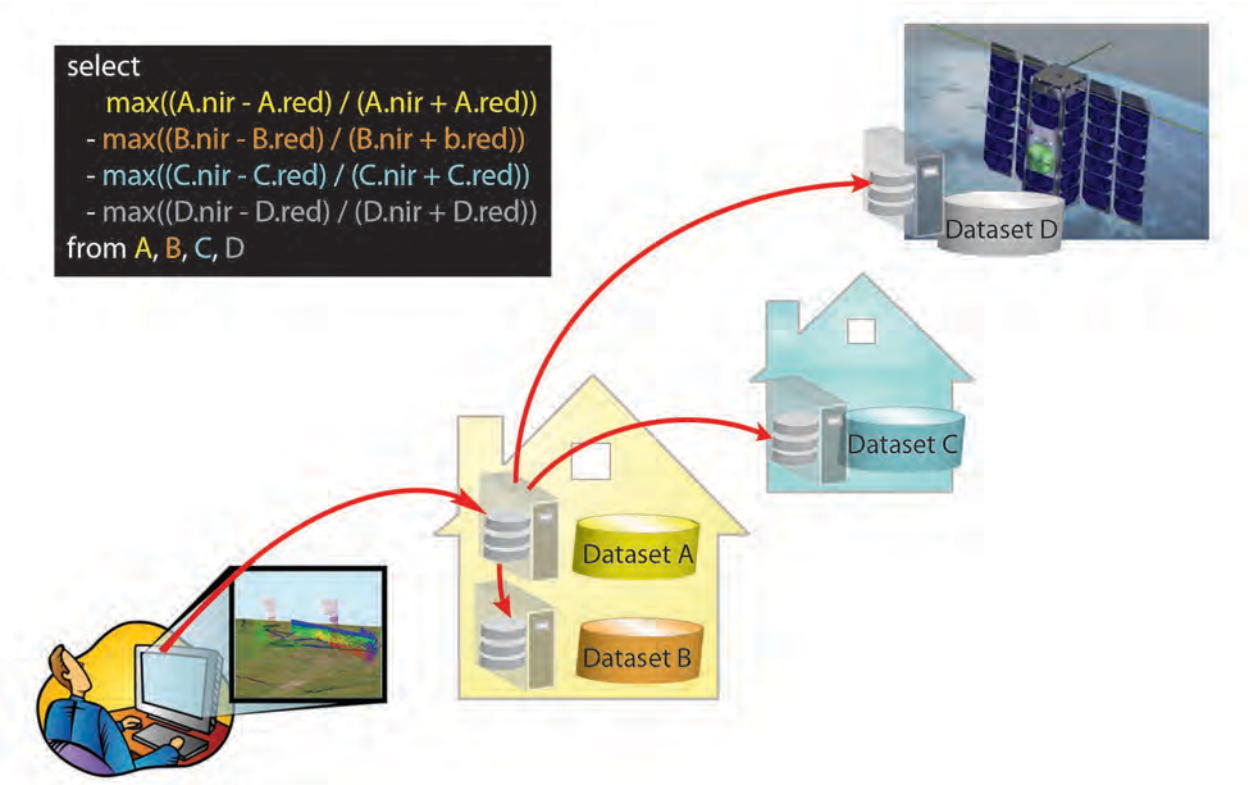


Figure 6-71. Rasdaman query splitting (the satellite onboard peer node, top right, will be launched in 2017 as part of the ESA OPS-SAT Mission).

This way, single queries have been successfully split across more than a thousand Amazon cloud nodes (Dumitru *et al.*, 2014). Figure 6-72 shows an experiment done on the rasdaman distributed query processing visualization workbench where nine Amazon nodes process a query on 1-TB processed in 212 milliseconds.

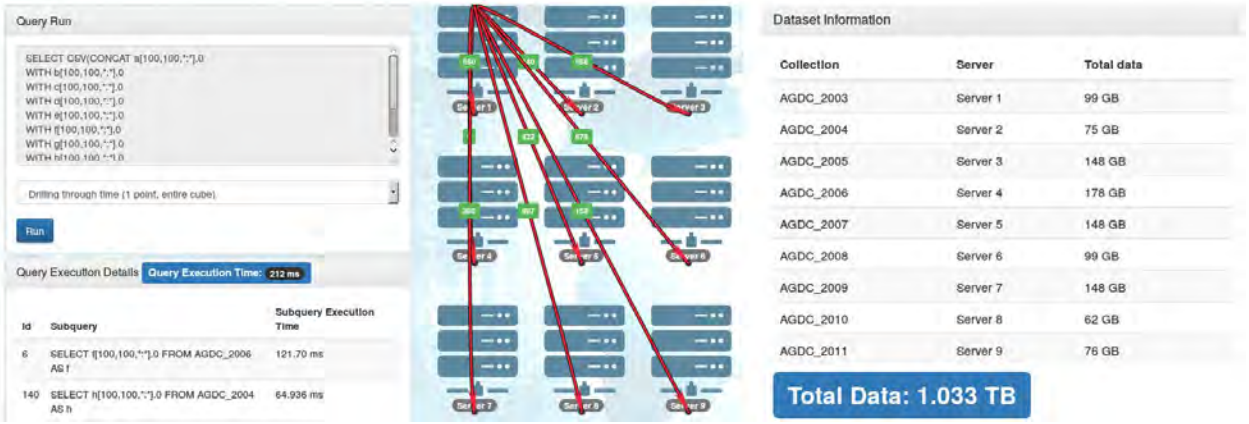


Figure 6-72. Visualization workbench for rasdaman distributed query processing.

Already earlier, in 2014, EarthServer had demonstrated early-stage integration by querying NASA and ESA simultaneously. NASA contributed drone imagery on Californian bush fires whereas ESA contributed worldwide DEM data; the intersection of both, California, allowed to combine this common location in

WCPS queries sent to a server at NASA Ames in California, in the *nasa.gov* domain, and ESA in Italy, in the *esa.int* domain (Figure 6-73). Both partial queries have been processed locally at their respective server, and the result has been merged in the client. Security Dimensions GmbH partner, as part of a collaboration with EU project CobWeb, supported this endeavor so that the different security policies of both servers addressed were taken into account properly. This handcrafted demonstration showed feasibility of distributed array query processing as such.

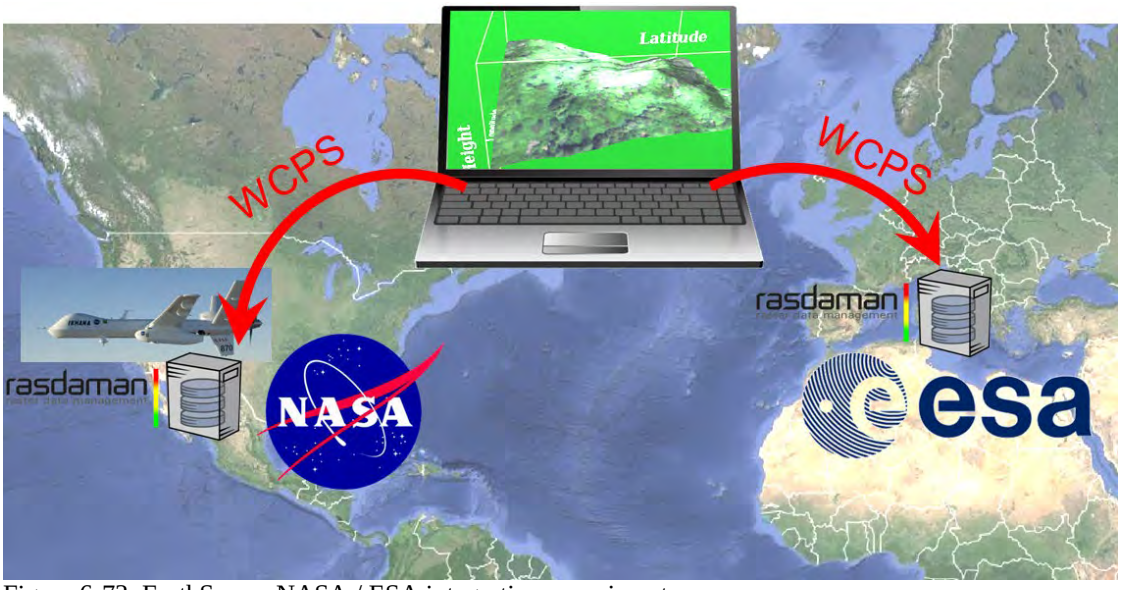


Figure 6-73. EarthServer NASA / ESA integration experiment.

Queries are analyzed by the receiving rasdaman node, and then distributed if the analyzer determines an advantage. At EGU with over 13,000 participants a dynamic splitting of queries has been demonstrated live. Data sets accessed in the sample query shown in Figure 6-65 consisted of climate data hosted by the European Centre for Medium-Range Weather Forecast (ECMWF) in Reading near London, UK, and Landsat 8 time-series hosted by National Computational Infrastructure (NCI) in Canberra, Australia. Participants could send queries to one of those partners, observe the query and data path taken, and inspect the result in NASA WorldWind (Figure 6-74).

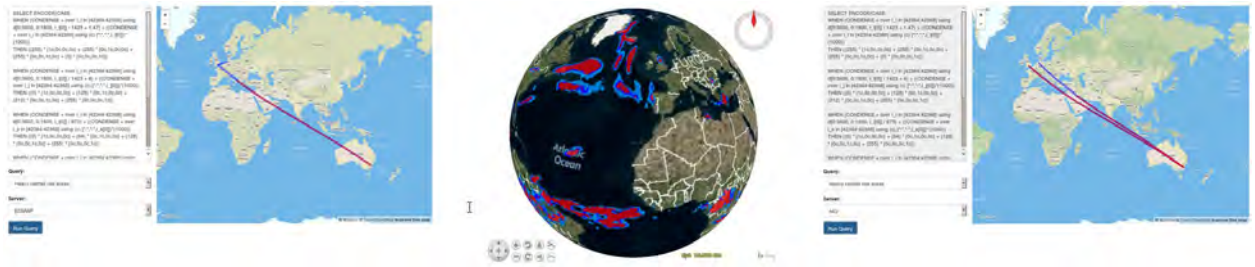


Figure 6-74. Dynamic query splitting with rasdaman: query sent to ECMWF, which spawns subquery to NCI (left); query result (center); and query sent to NCI, which spawns subquery to ECMWF (right).

4.4.3.3 Tool Integration

Even though the WCS, WCS, and WCPS protocols are open, adopted standards, they are not necessarily appropriate for end users – from WMS we are used to have Web clients like OpenLayers and Leaflet which

hide the request syntax, and the same holds for WCS requests and, although high-level and abstract, the WCPS language. In the end, all these interfaces are most useful as client/server communication protocols where end users are hidden from the syntax through visual point-and-click interfaces (like OpenLayers and NASA WorldWind) or, alternatively, through their own, well-known tools (like QGIS and python).

To this end, rasdaman already supports major GIS Web and programmatic clients, and more are under development. Among this list are MapServer, GDAL, EOxServer, OpenLayers, Leaflet, QGIS, and NASA WorldWind, C++, and Java. Python is in an advanced stage of development. Figure 6-75 shows retrieval QGIS accessing the PML service using the WCPS query *for c in (CCI_V2_monthly_chlor_a) return encode (c[ansi("2001-12-31T23:59:00")], "netcdf").*

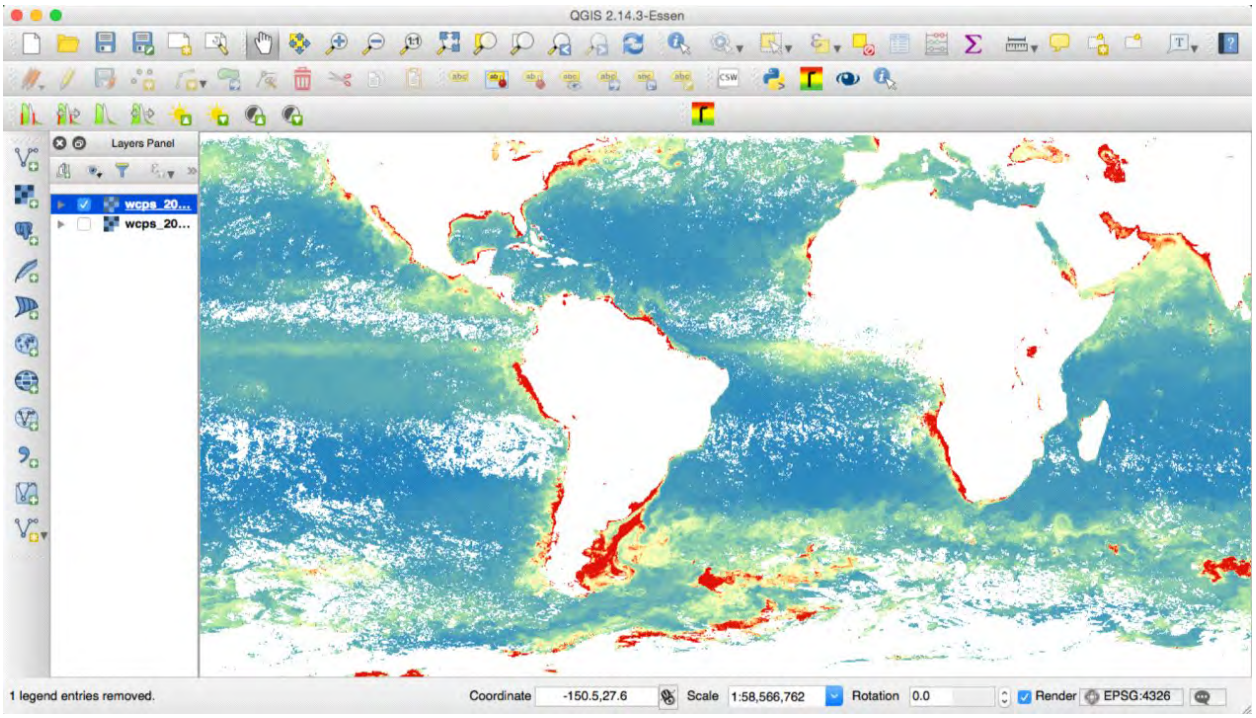


Figure 6-75. Sample QGIS access to a rasdaman service (service: Plymouth Marine Laboratory).

4.4.4 *EarthServer Services*

EarthServer encompasses a range of services on heterogeneous data and for various user groups:

- The European Space Agency (ESA), assisted by MEEO S.r.l., operates a service using satellite image time-series with a current volume of about 130TB.
- Plymouth Marine Laboratory (PML) is hosting around 75TB of Ocean Color Climate Change Initiative (OC-CCI) data.
- National Computational Infrastructure (NCI) Australia is hosting a service on Landsat 8 time-series datacubes.
- The European Centre for Medium-Range Weather Forecast (ECMWF) is growing a climate datacube which is foreseen to include its 87 PB MARS archive.

Jacobs University is hosting PlanetServer (http://planetserver.eu/), a service on Mars, Moon, and in future further solar system bodies.

Figure 6-76, Figure 6-77, and Figure 6-78 provide some impressions of the services; all are linked from www.earthserver.eu/services.

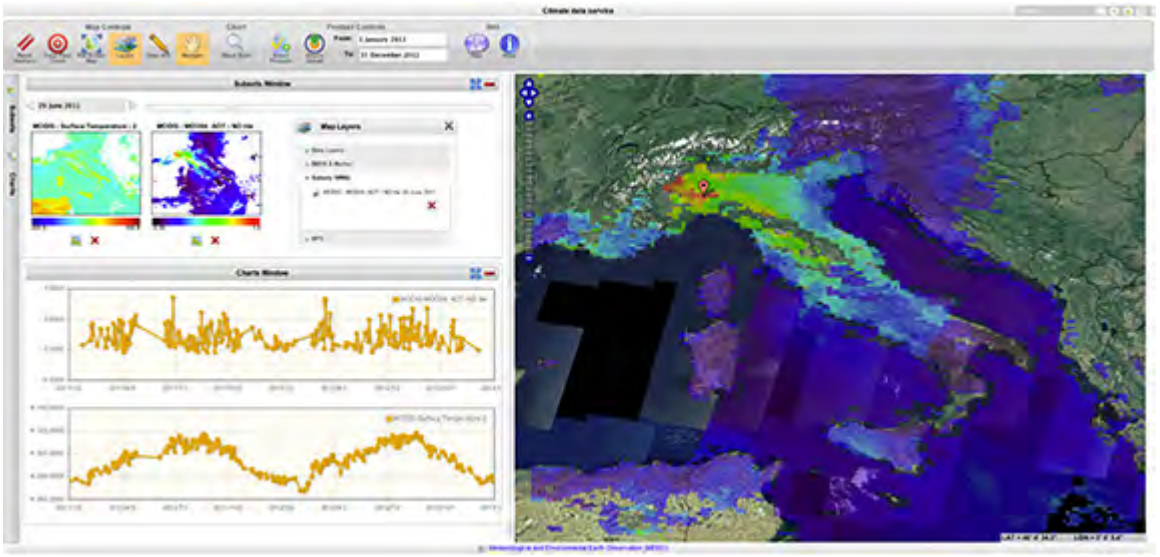


Figure 6-76. EarthServer Climate Analysis Service (source: MEEO ESA).

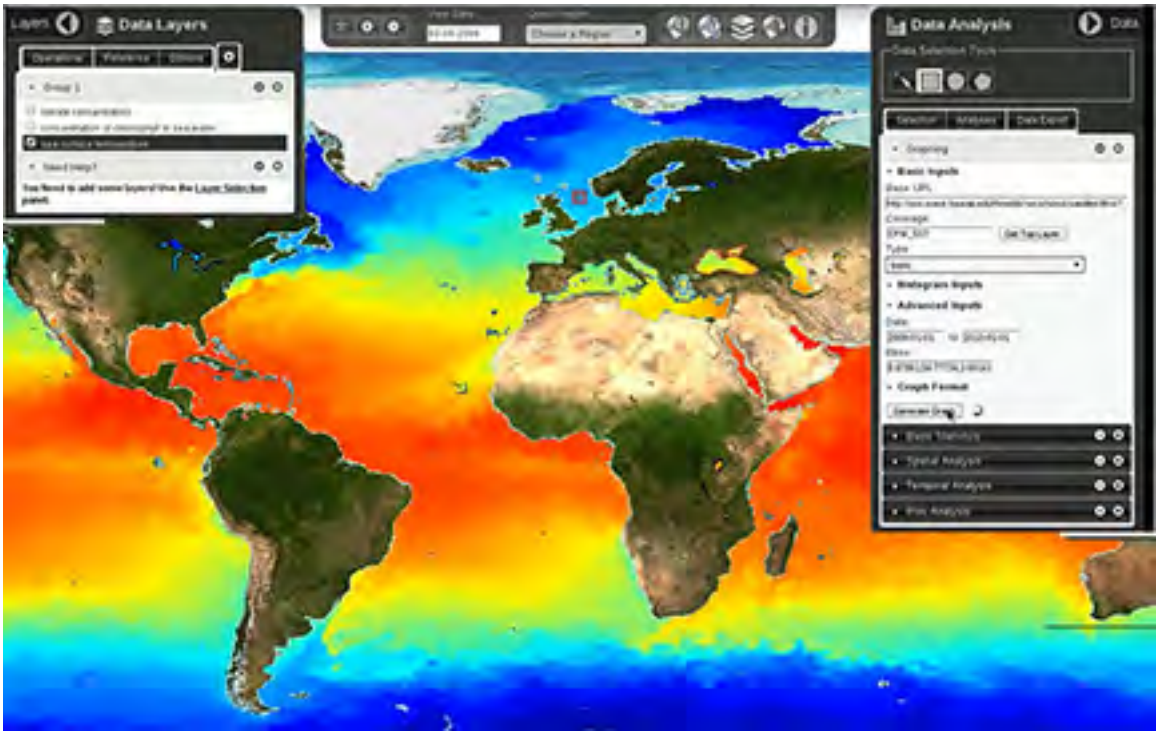


Figure 6-77. EarthServer Ocean Analysis Service (source: Plymouth Marine Laboratory).

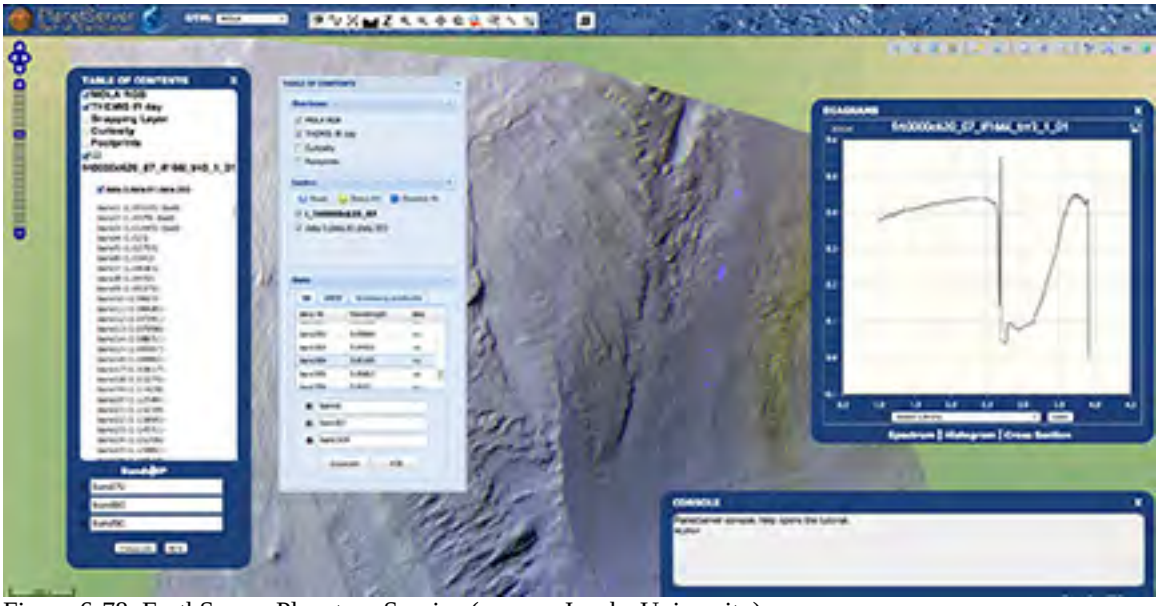


Figure 6-78. EarthServer Planetary Service (source: Jacobs University).

4.4.5 *Conclusion and Outlook*

With today's technology it is feasible to merge massive file sets into conveniently tractable spatio-temporal datacubes. Coverages provide a suitable abstraction for a "condensed" information representation. "O*ne cube says more than a million images*". OGC's modular coverage data and service model appears convenient for user-centric download, extraction, aggregation, and fusion of multi-dimensional data. OGC WCS offers scalable functionality ranging from simple download and extraction over scaling, re-projection, and further service facets up to powerful direct analytics.

In terms of technology, array databases in context of NoSQL database technology have proven instrumental in establishing flexible, scalable datacube services, demonstrated through operational services offering hundreds of Terabytes through the rasdaman array engine. Effectively, as intercontinental query splitting demonstrates, data center federations become possible where users can mix and match datacubes regardless of their location. The EarthServer initiative exploits this technology in the quest for on-demand federated datacube analytics.

Standardization is progressing rapidly. In OGC, CIS 1.1 is close to finalizing the adoption process. ISO TC211 is currently adopting OGC CIS, after which the abstract coverage standard 19123 will be modernized into 19123-1. Finally, it is planned to adopt WCS as an ISO standard. INSPIRE, the European legal framework for a common Spatial Data Infrastructure, is currently adopting WCS as a Coverage Download Service standard, including WCPS as an optional component offering additional functionality on coverages. The standards observing group of the US Federal Geographic Data Committee (FGDC) sees coverage processing in the style of WCS/WCPS as a future "mandatory standard".

Another ongoing activity in ISO is extending the SQL query language with multi-dimensional arrays in a domain-independent way. This candidate standard, called SQL/MDAs (Multi-Dimensional Arrays), is crafted along the rasdaman model as its blueprint. Altogether, EarthServer has a lead in Big Datacube standardization in OGC, ISO, INSPIRE, and beyond.

4.4.6 *Acknowledgement*

The author acknowledges gratefully the ingenuity and enthusiasm of the rasdaman team, especially Dimitar Misev, Alex Dumitru, and Vlad Merticariu. Further, he is grateful to the collaborators in EarthServer-1 and -2. Both projects have received financial support by the European Commission under the FP7 and H2020 programs.

4.5 *ESRI Online Access to Remote Sensing Data*

For the full value of Remote Sensing data to be realized, they need to be accessible online to multiple users within specific organization, or publically. Such online products could be the original source data, derived products, or information services. The massive volumes required by remote sensing data mean that simple file sharing techniques are impractical, so multiple technologies have been developed to make data accessible as Web Services.

4.5.1 *The exploding volumes of data*

Remotely sensed data coming from satellite, aerial, UAS and terrestrial platforms are processed to create an extensive range of products including orthoimages, elevation models, change maps and extracted features. Various notations are used to define the different levels of processing, but typically Level-0 refers to recorded measurements, Level-1 refers to data that have been resampled such that they can be located geometrically and represent some known measurement; Level-2 is typically orthorectified with potential color enhancements; and Level-3 is often a mosaic merged from multiple collections. Each of these processing levels is the source for further derived products needed for downstream analysis, or for human visualization and interpretation.

The large volumes of data and fast processing capability of modern servers mean that there is significant advantage to utilizing the *big data* concept to bring the *processing-to-the-data*, rather than *the data-to-the-processing*'. It is generally advantageous to process the data as much as possible prior to their transmission. Typically, the memory required to store each high level or derived product, is smaller than its source; but, at each higher level there is some information loss. With so many different products being created from the data source, the size of all the derived products can quickly exceed that of the source, further aggravating data storage and management issues. The source data are rarely destroyed, but may be archived to slower and cheaper storage systems.

4.5.2 *Transforming data into information*

The aim of online access to remote sensing data is to provide users with the appropriate product for visualization or further analysis over a computer network. This access needs to be fast and in many cases access control and security need to be implemented to ensure that only appropriate users have access. Use is best increased by creating applications that are simple to use, focused and targeted to specific user needs, and provide accurate and authoritative data. These factors drive the requirement to make imagery accessible via Web Services that enable access to imagery by browser and mobile applications, in addition to professional-grade desktop applications. Using Web Services, a single source of remotely sensed data can provide different products to multiple applications and increase their utility and value.

4.5.3 *Methods for Online Access*

Due to the large number of user requirements, there is no optimum method for online access to remotely sensed data. The following points provide ways of categorizing access into four methodologies, each of which has advantages and disadvantages:

- Data Download – Extraction and downloading of required sections of the data
- Tile cache – Access to pre-processed image tiles
- Dynamic Image Services – Access to user defined extents with server based processing
- Geoprocessing services – Access to processing tasks that return information products

4.5.3.1 Data Download

Traditionally, users have accessed image data by downloading required datasets based on a catalog of products or by using 'clip, zip and ship' services. In most such implementations, the products are stored in some form of tiered system and metadata about the products are stored in some web-accessible catalog. Users are able to *search* the catalog for datasets relevant to their specific application, *view* metadata and possibly *preview* reduced resolution renderings of the data. There are multiple methods by which such catalogs are published in a searchable form including using OGC CSW, providing webpages that get crawled and indexed by search engines, or by using feature services such as OGC WFS. Nearly all vendors for satellite imagery provide some form of web-based search tools to help users identify and order the required data. Once the appropriate datasets are selected, some form of access control, and possibly payment, is invoked. The delivery of data may be by multiple methods. Traditionally, the server packages data onto a specified temporary staging location and are then copied to a DVD or Hard Disk and shipped to the user. With the availability of faster internet connections, FTP or similar download protocols are often used. Users download their data to their local storage for further processing and analysis.

A range of DAP (Data Access Protocol) servers (such as OpenDAP) exist that also provide data download services. Such servers provide metadata about the holdings and provide an interface to enable dataset or subsets of the data to be downloaded. Such servers are used widely in Earth science disciplines when working with data in formats such as NDF and NetCDF.

With the advent of web accessible cloud storage, new opportunities for data download are evolving. Remotely sensed data can be stored in web accessible storage in cloud infrastructures so that data providers can provide customers with the relevant credentials to access the required storage directly. By storing data in suitably tiled data structures along with the associated metadata files on this cloud storage, individual tiles can be accessed on demand, often removing the need to download the complete datasets.

The key advantage of these data download methods is that once the data have been transferred, there is no reliance by the user on the data provider. The disadvantage is that the products are not web accessible directly. The volumes of data transfer can be very large, take a long time to execute, resulting in data duplication and data management issues for users.

4.5.3.2 Tile cache

In cases where the products for delivery are orthorectified mosaics of rendered imagery, the most efficient method for online web delivery is to use tile cache. By this process the source data are processed and rendered into very large numbers of small tiles based on a predefined tiling scheme in a predefined

projection. These tiles are stored in web-accessible storage typically fronted by a server that provides access control through a tile handler. Client applications are typically web applications that use a tile scheme that identifies the specific tiles required to cover a map extent at a suitable scale. The server returns the requested tiles that are then displayed. This method of serving imagery is very efficient and scalable as it makes extensive use of the caching capabilities of servers, networks and browsers, and puts almost no load on a server. The tiles are stored in data structures that enable them to be extracted with very little compute power. A single server can handle large numbers of requests since the server only acts as a tile handler to translate the request into an offset in the storage system and returns the appropriate data tile. Typically these tiles are defined as being JPEG or PNG tiles that can be displayed directly by web browsers without needing plugins. For any specified extent, web browsers can make multiple simultaneous requests to the servers that stream back the tiles that are then recomposed by the client into a seamless map. Web browsers also cached such JPEG and PNG tiles in the same way that they cache such images from web sites. As a result, if a user pans or zooms back to a previous location new tiles do not need to be requested from the servers. If the map display is in the same projection as the tiling scheme then neither the server not the application need to re-project; thus, simplifying the implementation.

Tile Cache access is the most widely used method of accessing geospatial imagery on the web. The base satellite imagery from all major map websites such as Google Maps, Bing Maps, and ArcGIS Online are based on this method of tile cache. Various tiling schemes can be used, but the most common is based on using the Web Mercator Auxiliary Sphere projection. Level-0 is a tile that covers the whole world in one 256x256 pixel tile. Each subsequent level has four times more tiles and one-half the resolution. Level 20 has a ground pixel size of about $15cm$ and 5.5^{11} theoretical tiles.

There are multiple protocols that can be used to access tile cache, including OCG WMTS, TMS, ArcGIS Map Server tiles, Virtual Earth and KML Super Overlays. Each of these protocols works by publishing information on the supported tiling scheme and then returning the appropriate tile based on the application defining a specific Level, Tile Column and Tile Row.

The key advantage of Tile cache is the massive scalability and its use of web caching, making it most suitable for static background imagery for simple visualization. The disadvantage to the method is that it requires the data to be pre-processed into RGB 8bit tiles that can be very time consuming and requires additional storage. Pre-processing and caching of data means that the data are typically static and cannot change. Temporal datasets can be handled by some protocols by adding an additional dimension to the tiling scheme, or by providing multiple such services.

Natural color imagery is typically compressed using JPEG compression to ensure small data size suitable for visualization, but not for analysis. One of the limitations of JPEG is that it does not support transparency. Consequently, portable network graphics (PNG's) are used where transparency is required. Similarly categorical data, or datasets that do not compress well with Joint Photographic Experts Group (JPEGs), use PNG.

A number of other variations on the same concept exist, especially when used by systems that are not limited for use on traditional web browsers. These protocols can then use a range of different tiling schemes, larger tiles sizes, and different compression methods. This is the basis of Discrete Global Grid Systems (DGGS).

A similar concept is provided by the JPEG2000, or JPIP (JPEG 2000 Interactive Protocol), as well as Enhanced Compression Wavelet (ECW) and MrSID. These also rely on the data being pre-processed and stored in a specific tiling scheme, but larger tiles and fewer levels are used. Each tile is stored as a collection of wavelets, and the specialized client applications can make requests for ranges of wavelets which are streamed back by the server and reconstructed by the client into pixel values for display at a range of scales. One of the advantages of formats such as JPEG2000 and JPIP is that they can handle a greater range of bands and bit depths. The primary disadvantage is that they require specialized plug-ins for web applications and do not exploit many of the web caching techniques.

4.5.3.3 Dynamic Image Services

Dynamic Image Services also return imagery or raster data, but requests are not based on a static tile structure. Clients make request by defining a bounding box in a specified projection along with the width and height (in pixels) and the data format required. The server accesses and resamples the source imagery and returns the required image or raster. In most web applications the extents of the requests are typically the extent of the current map display, but requests can be of various sizes.

To achieve such functionality, the server has to read the source data to apply appropriate transformations on the imagery. Depending on the server implementation, this can be as simple as re-projection of the source raster; but, more advanced servers can include other processing and rendering as well. There are multiple web protocols that provide such interfaces including OGC WMS, OGC CSW and ArcGIS Image Services. Various server technologies exist that provide dynamic image services.

- OGC WMS is used for simple web mapping applications. The clients can define a bounding box, width, height as well as a selected predefined style. WMS has also been extended to enable the passing multi-dimensional parameters including time and elevation. The imagery is returned as JPEG, PNG or Graphics Interchange Format (GIF).
- OGC WCS is similar to WMS, but is used for applications that require coverage data that can be used for client-side rendering or as input to further analysis. The data can be returned in various formats including Geography Markup Language (GML), GeoTIFF, HDF-EOS, or National Imagery Transmission Format (NITF).
- ArcGIS Image Services provide capabilities similar to both WMS and WCS, but are accessed through REST or SOAP APIs that provide a wide range of additional capabilities.

The disadvantage of dynamic images services are that the server must perform some processing in response to a request, which results in a greater compute burden on the servers. The servers must also have fast access to the imagery that forms the source of the service. The processing performed can be as simple as extracting and returning a small section of pre-rendered imagery, or much more complex processing that includes applying a range of geometric and radiometric functions. The advantages of such services are their greater flexibility and their ability to provide a potentially large range of derived products without needing them to be precomputed.

Advanced servers able to provide methods for working with multi-dimensional data enable client applications to filter the output by parameters such as time, elevation or other variables. Some servers provide a range of on-the-fly processing capabilities that apply geometric and radiometric functions to the dataset.

Geometric functions can include re-projection and orthorectification of the source data and consequently work directly with lower level products. Applications can access the data suitably georeferenced for display; or, if required get the original non sampled data values by specifically excluding any projections. When resampling the source data, the appropriate sampling method should be defined to gain either better radiometric correctness (e.g., nearest) *vs* better visual interpretability (e.g., bilinear or cubic convolution).

Radiometric functions can range from applying predefined styling on the pixels or enable the client applications to define user-defined stretches to the data for better interpretation, or to apply atmospheric corrections for better analysis. For satellite imagery, other radiometric processes can include color correction, pan-sharpening of higher resolution panchromatic imagery with lower resolution multispectral imagery, computing indices or applying classification. On datasets such as elevation or categorical data functions, other functions can include band arithmetic, or the computation of derived datasets such as hill-shade or slope maps. All such functions are 'local' in that the required processing is only applied to pixels within the requested spatial extent.

Depending on the processing applied, data transmitted from the server to client applications are considerably smaller than the original datasets. For visualization purposes, the images are typically returned to the client applications as lossy JPEG compressions, or as lossless Portable Network Graphics (PNGs). Applications can set the compression quality to maximize the compression for fast access over low bandwidth networks with some loss of quality; or, conversely, increase the quality value to get higher information content. Some protocols can return data in higher bit depths, or numbers of bands using different compressions more suitable for further analysis. These are used to enable client-based rendering and analysis. In addition to returning data values, the protocols can return metadata about the data being queried and returned, enabling client applications to document the sources appropriately.

Many datasets are not pixel aligned and with one another. Dynamic Mosaicking is a capability that enables the source for such services to be large collections of overlapping or disparate images, with the server being capable of fusing images from multiple sources based on the rules defined by the client application. This enables users to visualize and work with collections of images as a single service or with the individual datasets. Examples of such services are those that provide access to collections of multispectral and temporal satellite imagery such as Landsat, or from commercial satellite and aerial imagery providers. Metadata about the scenes and data values being returned are important and can be provided by the more advanced APIs.

The server needs to be highly optimized to be capable of extracting, processing and returning the required imagery and metadata within a couple of seconds, as is required in web environments. Such dynamic image services are typically optimized for smaller requests with the size of the output being equivalent to a computer screen. However, the size of the requests can be very large, enabling the export of complete datasets. Similarly, dynamic image services can be used as a data source for tile- cached-based services, by transforming tile requests into small extents.

With the advent of HTML5 and great web browser-based processing, applications are being created that have data values returned by the servers so as to enable client side processing that provides more interactive web experiences at the expense of more data transfer.

4.5.3.4 Geoprocessing Services

Geoprocessing services are a very generic method of providing online access to all forms of geospatial data including remotely sensing data, but make use of web-accessible tasks. The services expose the capabilities of a set of tasks. Each task can take as input a set of parameters and provides the output in the form of some data structure or reference data added to another service.

A typical example of such a geoprocessing service is the computation of a view-shed or downstream trace based on a terrain model. The inputs would typically be only a location and possibly parameters such as viewer height above ground. The output would be a set of features or rasters that define the visible terrain, or in the case of the downstream trace, a polyline that represents the trace. Geoprocessing services are applicable to be used for advanced image analysis applications such as feature identification and extraction. In many cases such geoprocessing services return just a simple answer from a complex process based on many different geospatial sources. The source of remote sensing data for such processes can be data stored locally on the servers, Tile Cache services that provide sufficient data values or Dynamic Image services. Most such geoprocessing services are run asynchronously and the server needs to be able to return the status until such time that the task is completed.

The key advantage of geoprocessing services is that it can encompass nearly all forms of geospatial data and can be used to create server-based applications that perform extensive processing. With the rapid increase in cloud computing and big data analytics, there are opportunities to implement very complex and computer-intensive tasks that expose themselves as such services. Examples of such protocols include OGC WPS and ArcGIS Server geoprocessing services.

4.6 *Accessing NetCDF Data*

Unidata developed NetCDF (*Network Common Data Form*) in 1988 to provide a format and data access software for data providers and application developers in the atmospheric sciences. The initial version was based loosely on ideas from a data model first implemented in NASA's Common Data Format, using three kinds of data objects to capture the meaning in data: Variables, Dimensions, and Attributes. The initial software supported C and Fortran programs.

Unidata's NetCDF software is a freely distributed collection of data access libraries implementing a data format that is self-describing, portable, scalable, appendable, extensible, sharable, and archiveable – all important characteristics for those who wish to create, access, and share array-oriented scientific data. NetCDF has been adopted widely by the atmospheric sciences community. For example, climate and ocean model outputs are commonly archived and accessed in NetCDF format. A large user community fosters support in applications used for analysis and visualization, APIs for common programming and scripting languages, international community conventions, such as Climate and Forecast (CF) metadata conventions, and standards for interoperability. In 2008 Unidata released NetCDF-4, implementing an enhanced data model and new data access interfaces built over the HDF5 storage layer and format. The resulting software provided compatibility with existing NetCDF programs and data, more powerful data modeling abstractions, and features for use in high performance computing such as parallel I/O.

Unidata and the NetCDF developer community continue to maintain and support NetCDF libraries for C, Fortran, Java, Python, C++, and other languages. NetCDF's strengths include an active user community,

continued funding from NSF for Unidata's contributions to earth science data infrastructure, incorporation of NetCDF in commercial and open-source applications, increasing contributions from open-source developers, and an active international standardization effort in development and maintenance of metadata conventions such as CF.

Beginning in 2009, the NetCDF format has been endorsed by several standards bodies:

- The NASA Earth Science Data Systems (ESDS) Standards Process Group endorsed NetCDF classic and 64-bit offset formats as appropriate standards for NASA Earth Science data, and later issued a report recommending NetCDF-4/HDF-5 file format for endorsement as an EOSDIS Approved Standard.
- The U.S. Integrated Ocean Observing System (IOOS) Data Management and Communications (DMAC) Subsystem endorsed NetCDF with Climate and Forecast (CF) conventions as a preferred data format.
- The U.S. Federal Geographic Data Committee (FGDC) endorsed NetCDF as a "Common Encoding Standard (CES)".
- The Open Geospatial Consortium (OGC) approved the OGC Network Common Data Form (NetCDF) Core Encoding Standard, NetCDF Binary Encoding Extension Standard - NetCDF Classic and 64-bit Offset Format, and the NetCDF Enhanced Data Model Extension to the OGC Network Common Data Form Core Encoding Standard, as official OGC standards.

4.6.1 *Data Models*

4.6.1.1 The "NetCDF classic" data model

The original data model implemented in all versions of NetCDF-C software supports a simple data model with *variables*, *dimensions*, and *attributes*, as shown in the following simplified UML diagram (Figure 6-79).

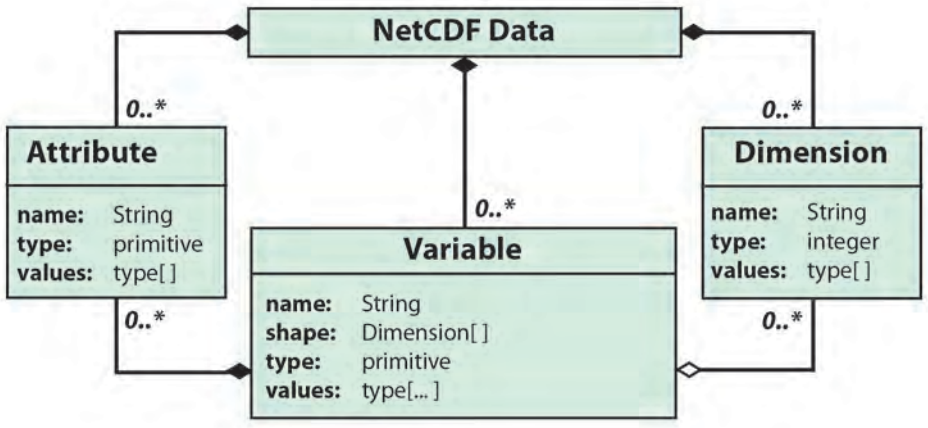


Figure 6-79. NetCDF classic data model. A NetCDF file has zero or more named variables, dimensions, and attributes. Variables also have attributes. Variables may share dimensions, indicating a common grid or coordinate system. One dimension may be of unlimited length. Primitive types for variables and attributes are: byte (8 bits), character, short integer (16 bits), integer (32 bits), floating point (32 bits), and double-precision floating point (64 bits).

Variables hold data values. A variable can hold a multidimensional array of values of the same type. Each variable has a name, type, shape, and attributes, in addition to values.

Dimensions are used to specify variable shapes, common grids, and simple coordinate systems. Dimensions shared by multiple variables indicate a common grid on which the variables have values. One distinguished *record dimension* may be of unlimited length, along which more data may be appended.

Attributes hold metadata containing information about the properties of a variable, such as its units of measure and special values to indicate missing data.

Variables and attributes have one of six primitive data types: character, byte, short, integer, float, or double.

4.6.1.2 The Enhanced Data Model for NetCDF-4

Although the NetCDF classic data model was simple to understand and explain, allowed an efficient implementation, and provided a good representation for gridded multidimensional data, it also had several important limitations:

- a limited set of primitive types
- a flat name space not ideal for organizing large sets of data objects
- a lack of nested structures and other user-defined types

The enhanced data model for NetCDF-4 extends the NetCDF classic data model to address these limitations in an upward-compatible way, adding more primitive types, named hierarchical *groups*, and *user-defined types* that may be nested, as illustrated in the following simplified UML diagram (Figure 6-80).

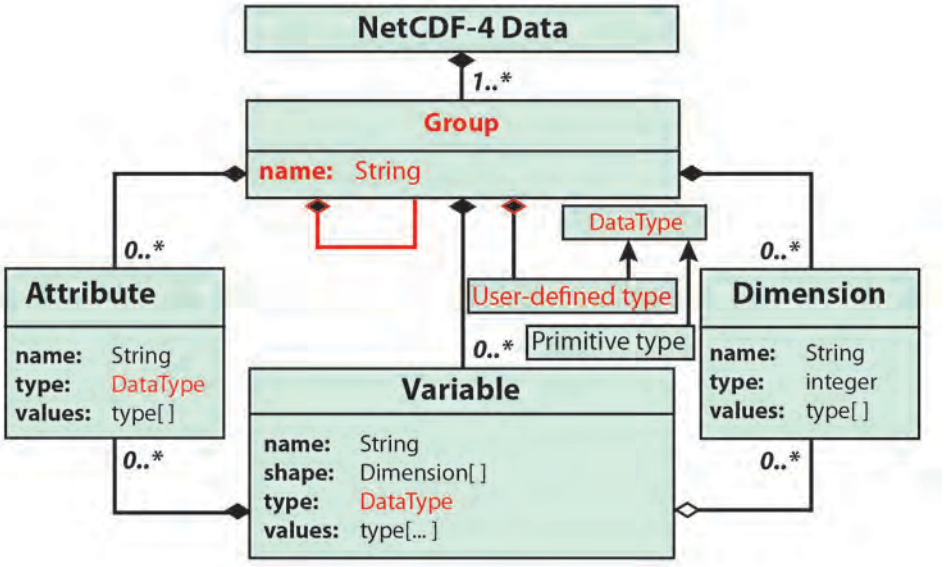


Figure 6-80. NetCDF enhanced data model with extensions from classic model in red.

A NetCDF file has a top-level unnamed group. Each group may contain one or more named subgroups, user-defined types, variables, dimensions, and attributes. Variables also have attributes. Variables may share dimensions, indicating a common grid or coordinate system. One or more dimensions may be of unlimited length. Primitive types for variables and attributes are: byte, unsigned byte, character, short

integer, unsigned short integer, integer, unsigned integer, 64-bit integer, unsigned 64-bit integer, and string. User-defined types include enumeration, opaque, compound, and variable length types, nested as needed.

By implementing the NetCDF enhanced data model over an HDF5 storage layer, additional performance benefits are also supported:

- Compression
- Chunking
- Multiple extendible dimensions
- Parallel I/O
- Efficient schema changes

The resulting data model is simpler than HDF5's with similar representational power. The NetCDF data model adds *shared dimensions*, which HDF5 lacks, to represent simple coordinate systems.

4.6.2 *Accessing NetCDF Data*

Developers may create custom applications to access NetCDF data through programming interfaces to libraries. The NetCDF C and Java libraries are independent implementations of such interfaces. Libraries for Fortran, Python, C++, and other programming languages are layered on top of the NetCDF C library.

The libraries make access available in several ways to multidimensional data from variables: a value at a time, a sub-array specified in terms of ranges of dimension indices, multidimensional sampled data slices specified with strides along selected dimensions or all values from a variable at once. Parallel I/O is supported for platforms with parallel file systems.

Data may be accessed from remote servers in the same way as from local files, implemented with OPeNDAP client protocols in the NetCDF C and Java libraries. This facilitates direct access to subsets on the server, so a small subset of a large file may be accessed efficiently, whether the data is local or remote.

Remote OPeNDAP servers such as Unidata's TDS (THREDDS Data Service), in addition to serving NetCDF data to OPeNDAP clients, also support access to data in many other formats, as if the data were CF-compliant NetCDF data. Similarly, such servers support access to subsets of data from multiple files as if the data were from a single file, through powerful and efficient data aggregation capabilities.

Open-source and commercial programs adapted to NetCDF access are available from numerous sources for browsing, analyzing, visualizing, and managing scientific data. Unidata maintains an up-to-date list of relevant software with brief descriptions and links to more information. Commercial software packages that "understand" NetCDF include ArcGIS, MATLAB, IDL, and Mathematica.

4.7 *WMO Standard Formats: BUFR, CREX and GRIB*

The World Meteorological Organization (WMO) is a specialized agency of the United Nations. It is the UN system's authoritative voice on the state and behavior of the Earth's atmosphere, its interaction with the oceans, the climate it produces and the resulting distribution of water resources. WMO had a membership of 191 Member States and Territories on 01 January 2013. Although established in 1950, WMO originated from the International Meteorological Organization (IMO), which was founded in 1873, and has retained the role of that body in facilitating international exchange of information relevant to weather, water and climate.

WMO Member States and Territories exchange their weather, water, and climate information using the WMO Information System (WIS). WIS is defined through WMO Technical Regulations that are regulated by an intergovernmental process (WMO 2012, WMO 2013a, b, WMO 2014 a, b). Although WIS itself is not limited in the formats of data it can exchange, users of the data need to be able to process the information rapidly to support safety-critical operations. WMO has specified data formats to support the operational exchange of information that must be used to represent information exchanged using the Global Telecommunications System (GTS) component of WIS (WMO 2012).

During its first thirty years, the World Weather Watch continued to exchange information in what are colloquially known as the *Traditional Alphanumeric Codes* (TAC), character-based data formats that had their roots in the exchange of data in the early years of the IMO. The meaning of characters in a TAC depends on the position of the character in the message relative to previous characters, and so a change to the TAC requires a change in the logic of software. This necessitated a change in the approach to data representations. WMO defined the Table-Driven Code Forms (TDCF) to overcome the issues associated with TAC, and in 1985 approved GRIB (Gridded Regularly distributed Information in Binary form) for operational use, with approval of BUFR (Binary Universal Form for the Representation of meteorological information) in 1988 and CREX (Character form for the Representation and EXchange of data) in 1998. GRIB is intended for the exchange of data on regular grids; BUFR and CREX are intended for exchanging irregularly distributed data.

The design principle behind the TDCF is that the message contains the definition of the data being exchanged as well as the data themselves. Because the TDCF were intended to be used to transfer large amounts of data over limited telecommunications lines, the definitions are recorded in tables external to the message (and formally published as WMO document WMO-No. 306 *Manual on Codes* (WMO 2014b). This approach to recording information about the data in tables is a key strength of the TDCF, as it enforces the use of controlled vocabularies so that operational users can build systems that can identify types of data reliably.

TDCF are used in the operations of weather services for the international exchange of many types of information including surface observations, observations from satellites, output from numerical models and analyses of frontal positions among many other types of data.

4.7.1 *BUFR (Binary Universal Form for the Representation of meteorological data)*

WMO data representations are given formal identifiers (FM numbers) that consist of a number for the code form (the WMO term for a data representation format) and an edition number for that form. BUFR is currently in its third edition with the formal identification FM94-XIV. This data representation and its supporting tables are defined by WMO (2014). An encoding of information in BUFR is called a BUFR message, which is divided into six sections.

- Section 0 identifies that the information is represented in BUFR (represented in CCIT IA5, allowing the format to be distinguished from other WMO formats), the edition of BUFR used (so that programs know how to interpret the message) and states the total length of the message (so that controlling software knows when to expect a different message to start, and provides a means of checking that the complete has been exchanged).

- Section 1 provides information on the center that created the message, the type of data represented (using a high level classification scheme; the detailed contents are listed in section 3), a representative date-time stamp for the information in the message, and indications of which versions of the tables are used in the message. It also allows centers to append local information to assist local management of the information, but this optional additional information cannot be interpreted by other centers unless the originating center provides instructions on how to do this.
- Section 2 (optional), is of arbitrary length and has no pre-defined contents or structure (beyond a requirement to state the length of the section).

  • Section 3 describes the contents of the message. It states the total number of subsets of data contained within the message (a data subset is a collection of data elements, such as the surface observations from a single station). It contains a flag to indicate whether the dataset contains observations or other types of information, and another flag to indicate whether or not the data are compressed. A characteristic of meteorological information is that the range of values in a given message may be much smaller than the total possible range, so that storing only the offset from the minimum value in the data collection can result in a much smaller message. The main part of Section 3 is the "data description", a list that contains at least one data descriptor(s) (table entry numbers) that describe the information contained within the message. In addition to the tables that define the contents, units and size of the data elements, other tables allow standard sequences to be recorded as a single grouping (such as all elements that are expected to be reported by a surface six-hourly reporting station) or to modify the meaning of other table entries (for example to allow more bits to represent a data item). Whereas the other sections use tables as code lists, the tables used in Section 3 are more complex and their features are summarized under section 5.
- Section 4 contains the data that are described in Section 3. Some of the information that has to be exchanged is not expressible as a number (for example the type of cloud). Such information is listed in "code tables" and the identifier for the entry in that code table that corresponds to the information to be reported forms the data item. Other entries may take the form of flags (where the table entry defines the meaning of a collection of logical states, such as whether certain types of quality control test were passed). Where data are marked in section 3 as being compressed, local reference values and offsets from these are stored. Otherwise the value stored is the positive integer value obtained by transforming the value to be reported using the scale and reference values and data width from the Table B entry.
- Section 5 contains the sequence *7777* represented using CCITT IA5.
- Tables used to support BUFR section 4 (Data Description)
- There are three types of tables supporting BUFR section 4: Table B (that describes the information being represented), Table C (that modifies how the Table B entries referred to in the data description should be interpreted), and Table D (that provides a means of standardizing the data contained in a message by using a single table entry reference to denote a pre-defined group of data descriptions).
- BUFR refers to table entries by a three-part identifier F X Y. F denotes the type of the descriptor: F=0 means that the descriptor describes an element, F=1 that the descriptor designates replication,

F=2 that the descriptor is an operator and F=3 that it defines a sequence. The meaning of X and Y depend on the value of F. For F=1 and F=2 X and Y are used to define the repetition (F=1) or type of operation (F=2). For F=0, X defines that "class" (themed sub-tables) and the identifier for the entry within that class (Y).

Table B provides the main functionality of BUFR. BUFR allows for a "Master Table" to be specified for different scientific disciplines, but only that for meteorology (Master Table 0) has been released. Each "Master Table" has its own Tables B and D. Each entry in Table B specifies the characteristics of how data are to be represented (Table 6-9).

Table 6-9. Meaning of BUFR Table B Entries.

Table Column Name	Meaning	Example (corresponding to entry 0 06 001)
Element name	Definition of the information being represented.	Longitude (high accuracy)
Unit	Units used to represent the information. In addition to physical units, this may specify "numerical", "code", "flag table", "CCITT IA5" (for text).	º (degrees)
Scale	The power of ten by which the reported value was multiplied to produce the integer value stored in the message.	5 (so that the integer value represents the longitude to the fifth decimal place)
Reference	The offset that was subtracted from the scaled reported value to obtain the value message.	-18000000 (so that the stored value lies in the range 0 to 35999999) stored in the
Data width (bits)	Number of bits used to represent the stored value.	26 (so the largest value that can be stored is 2^{26}-2; a stored value with all bits set to 1 means "missing data"). A point on the Greenwich meridian would have the stored value 18000000.

4.7.2 *CREX (Character form for the Representation and Exchange of data)*

CREX is closely related to BUFR and was developed to allow manual entry and interpretation of messages in situations where it was not possible to exchange information in binary form (for example, where telecommunications methods only permit the transmission of character information). Its structure is similar to BUFR, but compression is not used and, although it is still scaled and limited to a specified data width, the value in the message may be any integer. CREX is defined in WMO (2014) and has formal identifier FM95-XIV.

4.7.3 *GRIB (Gridded Regularly distributed Information in Binary form)*

GRIB is currently in its second edition and formally identified as FM92-XIV. This data representation and its supporting tables are defined by WMO (2014b). GRIB messages are divided into nine sections:

Section 0: Indicator section;
Section 1: Identification section;
Section 2: Local use section;
Section 3: Grid definition section;

Section 4: Product definition section;
Section 5: Data representation section;
Section 6: Bit-map section;
Section 7: Data section; and,
Section 8: End section.

The indentation scheme shown in the list of sections illustrates the groups of sections that may be repeated within a message.

- Section 0 starts with the characters *GRIB* represented in CCITT IA5 to allow the message type to be identified within a data stream. It indicates the "discipline" to which the data relate (valid values are defined in a code table), the edition number used (currently 2) and the total length of the message, including section 0.
- Section 1 (Identification) contains information about the center that originated the message, which international (Master) GRIB table version is used, which local GRIB table version is used to augment the Master table, the meaning of the reference time, the reference time the status of the data (for example "operational products") and the type of data (as defined by a code table entry).
- Section 2 (Local use) is available for centers to store information to assist their use of the data. The only required elements in this section are the length of the section and identification that it is section 2.
- Section 3 (Grid definition) is designed for efficient representation of gridded data. Section 3 provides information on the size of the grid (the number of data points within it), the definition of the grid (GRIB recognizes many different types of grid and projection and this section provides a method of recording parameters that fully define the grid even though different grids require different parameters). The group section 3 to section 7 may be repeated as many times as are needed to describe the data.
- Section 4 (Product definition) defines the type of information in the data (section 7) referred to by this section, both the type of processing (for example analysis) and the meaning (for example temperature). It also defines the vertical coordinates (or other layer specifications) that apply to the data in section 7. The group section 4 to section 7 may be repeated (within a "Section 3" level of grouping) as many times as needed to describe the data.
- Section 5 (Data representation) defines the structure of each datum in the data section. It records number of points in the data section; this is the number of points in the grid unless a bit map is used, in which case it is the (smaller) number of points for which value are stored. As for all other sections, the length of the section is expressed in octets and represented by a 64bit number; this places a (large) limit on the amount of data that can be included in one data subset. In practical terms, for operational data the operating practices of the GTS place more stringent limits on the size of GRIB messages.
- Section 6 (Bit-map) GRIB defines whether a bit map is used for the data in section 7; and if so, it contains the bit map. A bit map is used to indicate where values are omitted from the data section such as the packing or compression that is applied (for example so that values of sea surface temperature over land can be omitted).

- Section 7 (data) contains the data values at the points defined in sections 3, 5 and 6, using the data representation defined in section 5.
- Section 8 (End section) contains the sequence *7777* represented using CCITT IA5.

WMO data representations are used to support safety-critical operations throughout the world. The principal characteristics of the WMO data representations are that they are controlled strongly with international agreement on the content of the code tables used to support them, offer compression techniques that are matched to the characteristics of the information, and are used in operational systems be centers with widely differing technical capabilities.

4.8 *GMLJP2 Imagery Format*

The Open Geospatial Consortium (OGC) has defined a metadata standard for georeferencing JPEG2000 images with embedded XML using Geographic Markup Language (GML) format. JP2 and JPX files containing GMLJP2 markup can be located and displayed in the correct position on the Earth's surface by a suitable GIS similar to GeoTIFF images (Wikipedia).

4.8.1 *Annotating an image with graphics and transmitting it to another device.*

The composition of an image and some graphics if it is geographically annotated can be very complex and should respond to certain rules to ensure the transmission of all the information and if it is compressed improve the efficiency of the process. To meet the above requirements it has been foreseen the usage of GML for the annotation part and JPEG2000 for the imagery.

4.8.2 *Using GML standard for annotating an image*

The OGC Geography Markup Language (GML) standard is an XML grammar for the encoding of geographic information including geographic features, coverages, observations, topology, geometry, coordinate reference systems, and units of measure, time, and value objects. The Coverage is, loosely speaking, the digital representation of a spatio-temporally varying phenomenon; the formal definition of ISO 19123 (which is identical to OGC Abstract Topic 6) defines coverages formally as collection of direct positions in a coordinate space that may be defined in terms of up to three spatial dimensions as well as a temporal dimension. Examples of coverages include raster, triangulated irregular networks, point coverages and polygon coverages. Coverages are the prevailing data structures in a number of application areas, such as remote sensing, meteorology and mapping of bathymetry, elevation, soil and vegetation. The GML Application Schema – Coverages (GMLCOV) is a GML application schema uses GML 3.2.1 Coverages and SWE Common to generate a common application schema that can be used to describe coverage instances.

4.8.3 *ISO JPEG 2000 standard*

The ISO JPEG 2000 standard is a wavelet-based encoding for imagery that provides the ability to include XML data for description of the image within the JPEG 2000 data file. The JPEG 2000 standard does not, however, describe any means for including ancillary geographic information within the image, such as the geospatial coordinates of the image or annotations or references to features. This specification defines the

means by which the GMLCOV application schema and other means are to be used within JPEG 2000 for such purposes. For this reason a specific standard that is called GMLJP2 has been created.

4.8.4 *GMLJP2 standard*

GMLJP2 is a specific metadata standard for geographic imagery encoding. This specification includes the following elements:

- Specification of the uses of GMLCOV and GML within JPEG 2000 data files
- Packaging mechanisms for including GMLCOV and GML within JPEG 2000 data files, including brand field in File type box with value to signal JPX file and reader requirement box.
- Annotations, meaning associations between regions of interest and video, graphics, text, etc. and can be expressed in KML and its usage. The visualization of the coverage can also make use of KML. At the moment of writing other types of annotation styles are being considered like html5+svg but the progress made so far on encoding symbols in KML 2.3 seems quite promising.
- EO profile or other profiles for imagery metadata inclusion that can be Earth Observation Profile or others.

Figure 6-81 illustrates a basic example of the standard showing only a simple label for annotating a point in the image. In Figure 6-82 a more complex example is presented. Rather than just a label, it is possible to include a video or another image to provide a street view of an object (point) in the JPEG image.

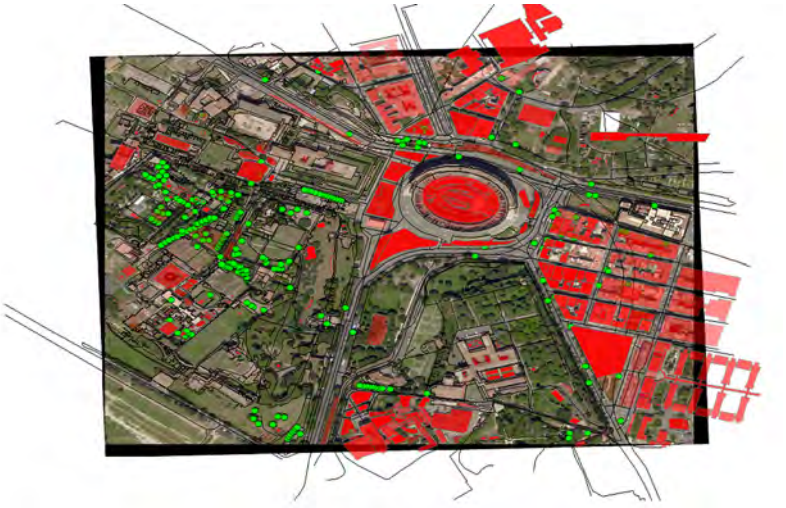


Figure 6-81. Example of a simple annotation.

Figure 6-82. Complex annotation with an image or video.

The implementation of the standard is based on open source software namely GDAL library and open-jpeg/jasper for the JPEG2000 codec. To minimize the impact of implementing the GDAL library, it has been decided to use a definition file that contains all the information to manage the encoding in the GML part of JPEG2000.

The support for creating the (conversion to) GMLJP2V2 format is built into our version of GDAL by extending the JPEG2000 driver. For example, the function:

```
static GDALDataset *JPEG2000CreateCopy(
                    const char *pszFilename,
                    GDALDataset *poSrcDS,
                    int bStrict,
                    char **papszOptions, /* <- options dictionary */
                    GDALProgressFunc pfnProgress,
                    void *pProgressData)
```

In JPEG2000dataset.cpp (frmts/ directory of the source) can be used for the purpose of converting/creating GMLJP2V2. Specifically, the key GMLJP2V2_DEF is inserted into the papszOptions dictionary argument of JPEG2000CreateCopy(...), with the value being one of the following:

- auto - in this case JPEG2000 driver generates GMLJP2 v2 file: i.e. the GMLJP2 box structure is present, with only minimal packaging and no GML data <GMLJP2CoverageCollection> element is computed automatically by GDAL from the information found in the input file.
- filename - the box structure and the xml sources of each of the GMLJP2 box is described in the input definition file of the name “filename”. The file path can either be relative or absolute. Paths involving white space (or special) characters can be quoted (e.g. "c:\some_dir\my_file
- definition.def").

If the GMLJP2V2_DEF key is not present in the papszOptions dictionary, JPEG2000CreateCopy(...) will act as it does for the default JPEG2000 driver. The syntax of the definition file is given below.

```
/*============================================================
    GMLJP2V2 conversion definition file.
    Specifies file box structure: root-instance and
    box sources/labels.
============================================================*/
{
    root-instance: {
        gml_id: "some_gml_id",      // specify GMLJP2CoverageCollection id here
        auto_coverage: 1,           // always 1, in future a 0 will indicate that
                                    // we wish to import gml as is
/*
here we specify the data file from which
the <gmljp2:eopMetadata> values are to be parsed.
*/
metadata: {
                {
// either absolute or path relative to
// the current definition file
             file: "dir/filename.XML",
             type: 0 // valid types {GEOEYE = 0,
// PLEIADES = 1, WORLDVIEW = 2 etc...}
               } // can have multiple number of these, comma separated
             },
 /*
        /*
            each gml file in gml_filelist will be included as <featureMember> of the root
GMLJP2CoverageCollection
        */
        gml_filelist: {
            {
                file: "converted/test_0.gml",                // can use relative or ab-
solute paths
                schema_location: "gmljp2://xml/schema_0.xsd"  // schema box label (link)
            },
            {
                file: "converted/test_1.gml",
                schema_location: "gmljp2://xml/schema_1.xsd"
            }
    }
    },
    box: {
      label: "schema_0.xsd", //
      file: "converted/test_0.xsd"
    },
    box: {
      label: "schema_1.xsd",
      file: "converted/test_1.xsd"
    }
  }
```

Each input file in the metadata array is converted to <gmljp2:eopMetadata> and inserted into <gmljp2:GMLJP2CoverageCollection>

under

<gmlcov:metadata xmlns:gmlcov="http://www.opengis.net/gmlcov/1.0">

<gmljp2:Metadata xmlns:gmljp2="http://www.opengis.net/gmljp2/2.0">

```
                    </gmljp2:Metadata>
</gmlcov:metadata>
```

4.8.5 *Future activities*

It is clear that the size of the imagery and the current development in the graphic libraries could envision an implementation of the above directly in the GPU.

4.9 *Compression Overview*

Compression is the process of transforming information so that it can be stored or conveyed in less space than the original information. With the rise of personal computing and the internet, compression has become ubiquitous as a means of minimizing data storage costs. For example, the music that you listen to daily, probably is stored in a compressed format, the most common being the MP3 format. Likewise, images, videos, and all manner of data are now compressed typically to minimize the consumption of resources like network bandwidth.

Given the amount and variety of geospatial data available today, it should come as no surprise that the remote sensing industry also takes advantage of compression for imagery and other data. This section focuses on imagery, providing an overview of how images are stored and represented on computers. The section also discusses why image compression is useful in geospatial workflows, highlight some common compression algorithms, and finally, compare the most common geospatial image formats.

4.9.1 *Why Use Compression?*

This is an exciting time for anyone who uses geospatial imagery. Increasingly powerful satellites capture more detailed imagery, including multispectral data with dozens of spectral bands. Unmanned aerial vehicles started making ripples in the remote sensing community, promising less expensive and more frequent image collection. LiDAR has skyrocketed. However, all of these innovations come with new challenges. As the number and variety of images increases, so does the problem of storing and accessing those data. One strategy for addressing these challenges is image compression.

4.9.2 *Benefits of Compression on a Local File System*

The most obvious benefit of image compression is the potential reduction in data storage costs. Depending on the compression algorithms used, it is not uncommon to observe an order-of-magnitude (or more) decrease in the storage space required. For example, consider the images in Figure 6-83. For many purposes, the images are equivalent, yet the compressed image on the right is more than 10 times smaller than the image on the left.

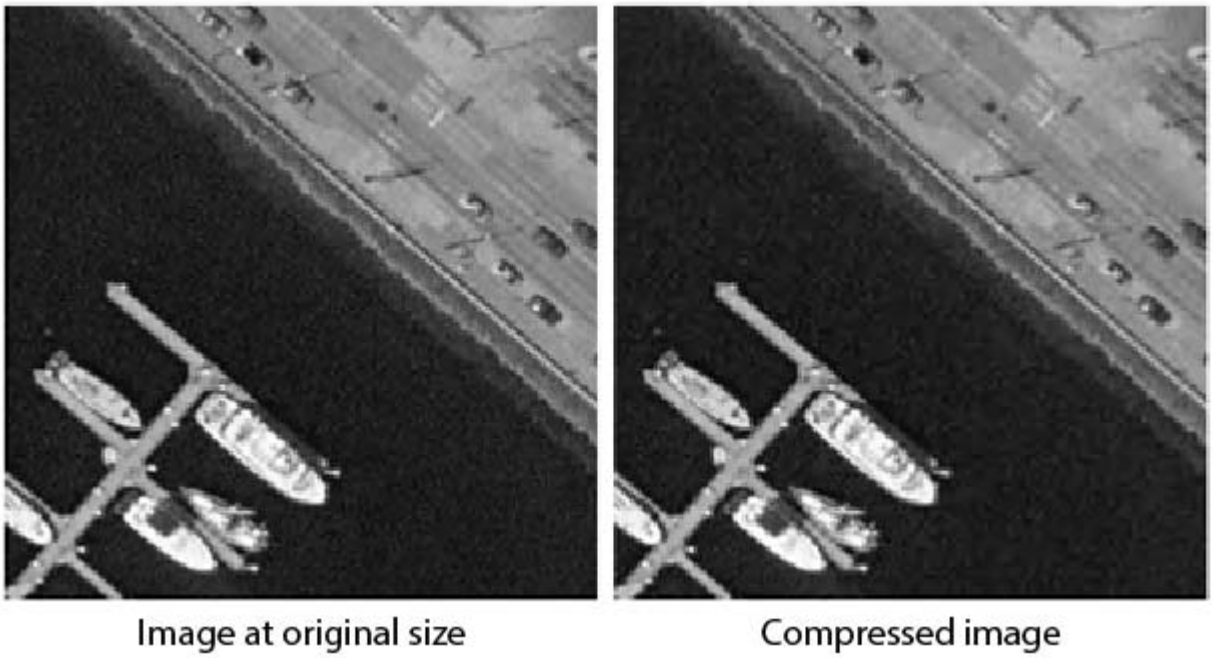


Figure 6-83. Original and compressed (10:1) gray scale images.

4.9.3 *Benefits of Compression on a Network*

When organizations distribute imagery to many users, there are two typical approaches for dealing with large datasets. If money and network capacity are plentiful, the challenges of large datasets can be addressed with increased hardware. The dataset is partitioned into small tiles and duplicated at multiple resolution levels. As a result, requests for imagery receive preprocessed image tiles which are then assembled by the client viewer into a single image. This solution is fast and scalable but can involve large costs because of the duplicate data and large network bandwidth required. Most consumer-level mapping applications work this way.

An alternative solution uses specialized image servers that leverage certain desirable qualities of modern compression techniques such as JPEG2000, MrSID, and ECW. Because these image formats incorporate multiple resolutions as a natural by-product of image compression, there is no need to store the same datasets at various resolutions. Furthermore, such formats support "selective decompression" which means arbitrary extents of arbitrary resolution can be decompressed. This avoids the computational cost of decompressing the entire dataset. Thus, not only are storage savings achieved, they are realized in such a way that access to the uncompressed pixel data is very fast.

Finally, to transmit images with a minimal network bandwidth, image compression is used in conjunction with a streaming protocol such as JPIP (JPEG2000 Interactive Protocol). In addition to its emerging use in geospatial applications, JPIP is used often in healthcare for medical imagery. Streaming compressed imagery not only makes it easier to maintain large-scale systems with many users, it also has applications for users in remote locations or on mobile devices that may have bandwidth constraints. A low resolution rendering can be passed and later refined by passing along only the missing details.

4.9.4 *Practical Applications*

4.9.4.1 Compression for Municipal Analysts

Given the increased data storage required for tile servers, some municipalities prefer to use compressed image servers. Also, because compressed image servers often store imagery as a single massive dataset, compressed image servers can reduce the need for a system administrator.

4.9.4.2 Compression for Deployed War-Fighters

War-fighters in the field often benefit from the ability to view satellite imagery of the surrounding terrain. However, because network access is limited often in the field, many war-fighters must rely on whatever imagery they are able to carry with them. The problem is compounded by the fact that portable devices offer limited storage space. As a result, it is common practice to store heavily compressed satellite imagery and other maps directly on mobile devices.

4.9.4.3 Drawbacks of Compression

Despite the many benefits of image compression, it is not suited for all geospatial applications. For small organizations with a modest amount of imagery, the cost of compression software suitable for geospatial work may outweigh the savings in storage space. Likewise, care must be taken to ensure that imagery is not compressed beyond the point of usefulness. For work that only requires visual identification of image features or in cases where imagery is used as a simple backdrop, high degrees of compression may be acceptable. On the other hand, high-precision work that demands absolute accuracy from each pixel should ensure that compression does not alter the imagery in any way. Another factor to consider is that most of the advanced compressed image formats in use by geospatial professionals are proprietary. A notable exception is the JPEG2000 format which will be discussed in more depth later. Public organizations that require the use of open standards may not be able to take advantage of common proprietary formats.

4.9.5 *Understanding Imagery*

4.9.5.1 Images as Pixels

Before discussing image compression, it is useful to understand how images are stored. Images are simply rectangular arrays of pixels, where each pixel represents a specific value (Figure 6-84). The simplest images are grayscale images, so called because they only represent shades of gray. Because these images only represent values between white and black, the image data can be stored in a single array of pixels, also known as an image band or sample. The most common range of values is 0 to 255, where 0 represents black and 255 represents white. Images that use this range of values are called 8-bit images because they have 2^8 or 256 possible values.

To represent colors, images typically store information in multiple arrays, the most common set of which is a red array, a green array, and a blue array. The shorthand for this type of image is RGB. In this case, each array of pixel values represents a different color, such that the combination of all three colors yields a very wide variety of possible colors. If each image band is allotted 8 bits, then each pixel can have one of 256 x 256 x 256 colors, for a total of 16,777,216 possible colors.

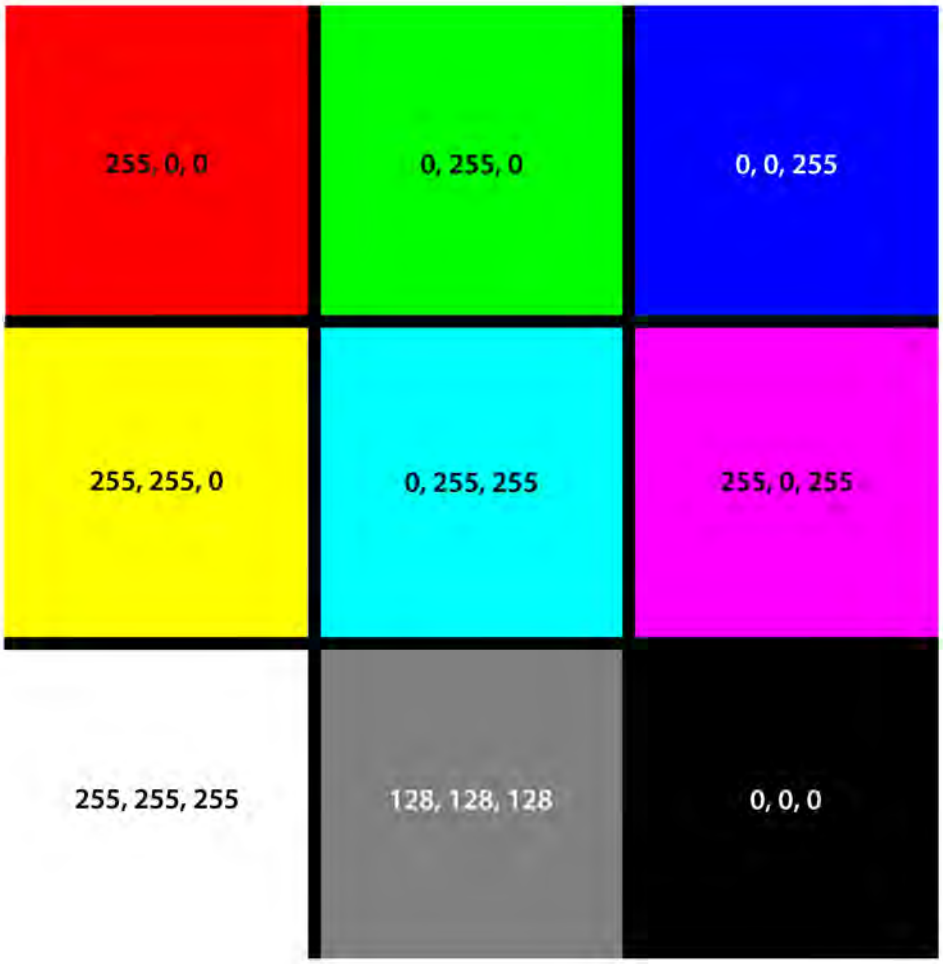


Figure 6-84. RGB pixel values for common colors.

The number of image bands and the range of pixel values for each image band can vary greatly depending on the type of image. Imagery with more than three bands is often referred to as multispectral, which means that it stores visible light in addition to light from other parts of the electromagnetic spectrum. For example, the remote sensing industry frequently makes use of four banded imagery which contains the typical red, green, and blue image bands, as well as an additional near infrared band. Indeed, some modern satellites capture multispectral imagery in dozens of image bands. Additionally, as the accuracy of cameras increases, 256 pixel values cease to suffice for each image band. As a result, it is becoming increasingly common to see images that are 12-bit, 16-bit, and even 32-bit.

4.9.5.2 Measuring Image Size

To understand how much an image has been compressed, it is important to distinguish between the size of an image file and the size of an image. The size of an image file depends on the file format and the compression algorithm used to store the image. The image size is the number of bytes required to represent the image in its uncompressed, or raw, form. The image size is measured independently from the file format and stems from Equation 6-7:

$$Size = Length \cdot Width \cdot nSamples \cdot SampleSize/8 \quad (6\text{-}7)$$

where: *Size* is the number of bytes that make up the image, independent of the file format; *Length* is the number of rows of pixels in the image; *Width* is the number of columns of pixels in the image; *nSamples* is the number of bands or samples; and *Sample Size* is the number of bits for each image band.

4.9.6 *Understanding Compression*

4.9.6.1 Measuring Compression

Now that image size has been defined, one can begin to measure the degree to which an image is compressed. An intuitive means of doing this is with the compression ratio, which can be calculated according to Equation 6-8:

$$Compression\ Ratio = \frac{RawImageSize}{CompressedImageSize} \quad (6\text{-}8)$$

where: *Compression Ratio* is a simple ratio of the two sizes; *RawImageSize* is the image size as calculated in the previous section; *CompressedImageSize* is the size of the compressed image on the computer file system.

4.9.6.2 Lossy and Lossless Compression

Compression algorithms can be grouped into two broad categories: *lossless compression* and *lossy compression*. Lossless compression allows one to recover the original pixel values exactly as they appear in the raw image. At a basic level, lossless algorithms take advantage of the fact that many images contain redundant information. For example, it is intuitive that an image of a street will contain many neighboring pixels that are the same color, and can be represented with the same pixel value. By storing duplicate pixel values in groups, images can be compressed without losing any information.

Lossy compression allows only an approximate recovery of the original pixel values. Thus, some colors may appear differently in images compressed with lossy algorithms. However, these algorithms often afford compression rates far beyond what can be achieved by lossless algorithms, and often with such small changes to the original pixel values that the raw and compressed images are indistinguishable by human eyes.

4.9.6.3 Measuring Lossiness

When speaking of lossy compression, it is important to devise a measure of how much distortion is introduced. In rigorous discussion, the *mean-squared error*, or *MSE*, is used often for this metric. For a single-banded image this takes the form shown in Equation 6-9.

$$MSE(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \quad (6\text{-}9)$$

where: N is the number of pixels, x_i are the original values and y_i are the lossy values.

Sadly, while easy to calculate and understand, the *MSE* does not serve as a useful suitability indicator for the most common of all image processing operations: visualization. That is, the *MSE* value is not an accurate predictor of the distortion that could be discerned visually by the average person. For this scenario,

another term is introduced: visually lossless. An image is said to be *lossless* when an average person cannot discern visually that the image has been compressed.

4.9.7 *Compression Algorithms*

A compression algorithm is the set of instructions used to compress data and range from very simple to very complex. Efficiency of the algorithms varies widely and is dependent often on the type of data. For example, a compression algorithm used to store audio data would not work well necessarily for image data. This section discusses some of the most common compression algorithms.

4.9.7.1 Palletizing

Palletizing algorithms reduces the total number of colors used by an image, such that similar pixel values are grouped together into buckets. Because the pixel values are changed, palletizing is an example of a lossy compression algorithm, with the amount of lost data depending on how much the possible color values are reduced. For example, fifty shades of gray may be reduced to a single shade.

4.9.7.2 Run-Length Encoding

Run-length encoding is one of the simplest lossless compression algorithms and can be very effective for images that have large regions with the same color. To visualize run-length encoding, think of a waiter at a restaurant taking orders from a large group of people. Rather than remember the order of each individual, the waiter might remember that three orders in a row were for the special and one was for the soup. Similarly, an image may contain three green pixels, then one red pixel, then another green pixel. By grouping pixels with the same value when they occur in a sequence, the image can be compressed without any loss of information. Many image formats use run-length encoding in addition to other compression algorithms.

4.9.7.3 Lempel-Ziv-Welch Compression (extension of RLE, LZW-tiff, GIF, Zip, Deflate)

This compression algorithm is often abbreviated LZW and is named after its creators, Abraham Lempel, Jacob Ziv, and Terry Welch. LZW is another lossless compression algorithm and is used in several compression utilities, including the GIF image format.

4.9.7.4 Huffman Encoding and Arithmetic Encoding

Huffman encoding and arithmetic encoding are two similar lossless compression algorithms. The basic idea of these algorithms is that the most common values in an image should be represented with the fewest number of bits. Thus, an image that uses Huffman encoding would associate the most common pixel value with a short code, and then store the short code instead of the pixel value. Pixel values that do not occur often in the same image would alternatively have a long code. Arithmetic encoding works based on a similar probability-based idea; however, it stores values as a single number. Arithmetic encoding is used in many advanced compression image formats, especially in conjunction with wavelet-based encoding.

4.9.7.5 Wavelet-Based Compression

Wavelet-based compression, also called the wavelet transform, is a way of breaking down or transforming a high-resolution signal into two parts: a low-resolution approximation and a second part that shows the

details about what changed, also called detail coefficients. Very importantly, it does this in a way that allows one to recover the original from the approximation; that is, it is a lossless algorithm. Many modern image compression formats make use of this algorithm in conjunction with other algorithms.

Because this algorithm is born from digital signal processing, it is common to refer to a series of pixels as a signal. As an example, take the following high-resolution grayscale image (Figure 6-85).

Figure 6-85. Original image is 256 x 256 pixels.

Sometimes, one would rather have a smaller approximation of the original signal. The easiest way to create such an approximation is to throw out every other pixel. This creates an approximation that is a quarter the original size (Figure 6-86). If this image is scaled up to the original resolution, the image appears blocky because of the lost information (Figure 6-87).

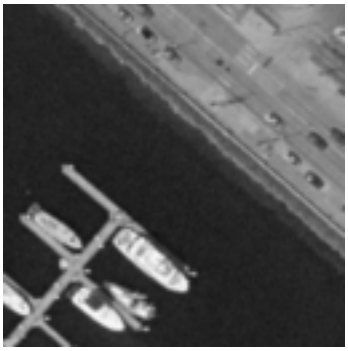


Figure 6-86. Down-sampled to 128 x 128 pixels.

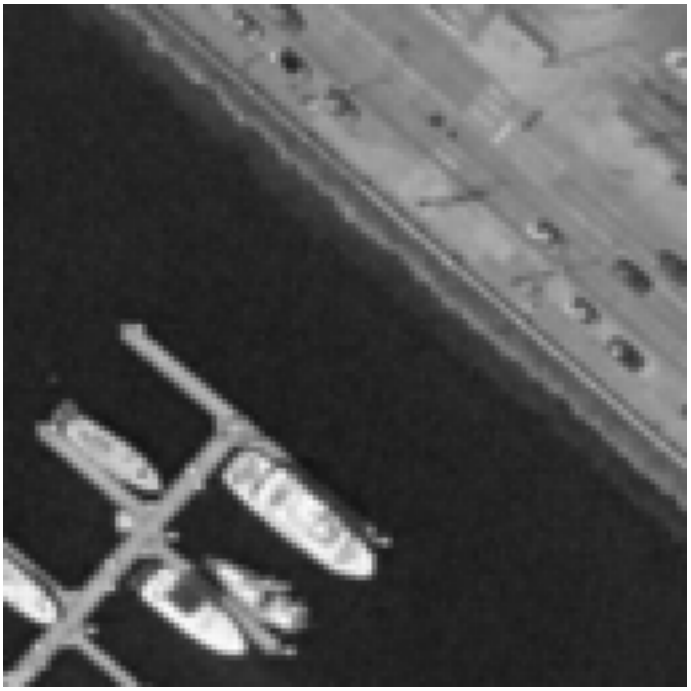


Figure 6-87. Resized down-sampled image showing loss of information.

Since pixels were discarded to derive the lower resolution version, there is no way to recover the high resolution version (Figure 6-85). However, if instead of discarding these pixels, the pixels are passed through the wavelet transform, these pixels can be recovered once more. This is typically illustrated by means of a decomposition diagram that shows the approximation in the upper left corner and the details coefficients over the rest of the image (Figure 6-88).

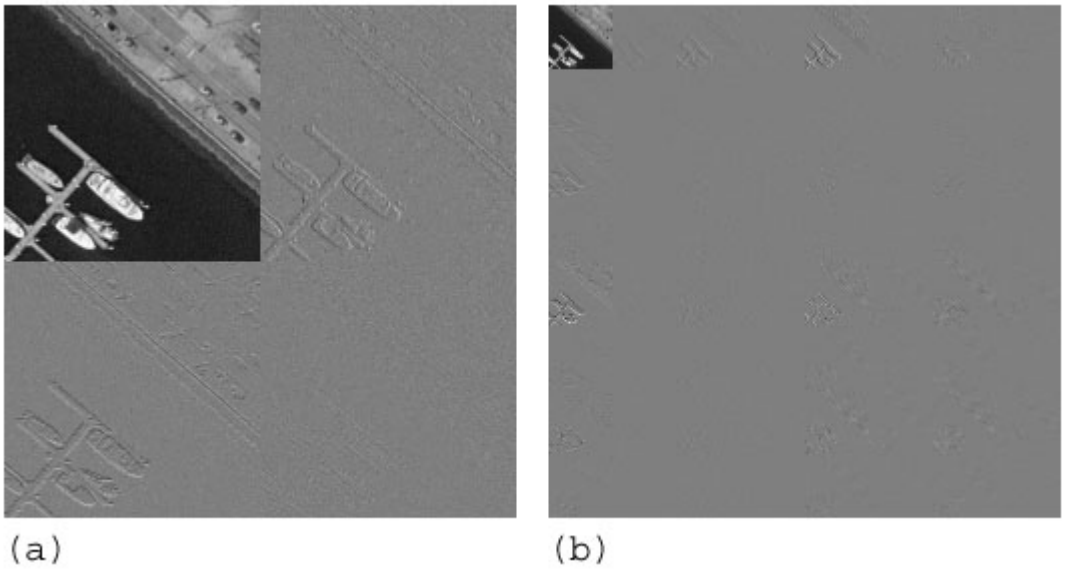


Figure 6-88. Wavelet transform of original image at one level (a); and at two levels (b).

Figure 6-88a is the wavelet transform of the original image. One can see the approximation in the upper left quarter. The other three quarters are the details coefficients that comprise the information missing from the approximation. You can roughly make out some of the structural details in the upper right and lower

left quadrants. On the whole though, the details coefficients have much less information than the approximation. At an intuitive level, this is because most of the information in the original is in the approximation. What isn't there shows up in the details coefficients. By recursively applying the transform to this result, one can get smaller and smaller approximations (Figure 6-88b).

The transform is lossless and the pixels in the original image can be completely reconstructed from the smallest approximation and all the wavelet coefficients. Optionally, this algorithm can be made lossy by simply discarding some of the small detail coefficients or by applying another compression algorithm to the detail coefficients.

As an added benefit, decomposing imagery using the wavelet transform yields two desirable properties. First, images are easily represented in multiple resolutions so that it is inexpensive from a processing point of view to display overviews of an image. Second, compression is to some extent localized; that is, it is possible to decompress pixels selectively from one portion of an image without reading all of the coefficients that comprise the vast bulk of the transform.

4.9.8 *Overview of Major Image Formats Used in the Geospatial Industry*

4.9.8.1 Traditional Formats

Although the number of compressed images in wide circulation has increased in recent years, the most common image formats are still uncompressed. Raw images, so called because they are stored in the same form as they were captured by a camera, are still widely used. Perhaps more common, is the TIFF format, which stands for Tagged Image File Format. TIFF is a widely supported image format whose specification is an ISO standard. The TIFF format is made suitable for geospatial applications with a set of metadata tags used to embed geo-referencing information and other metadata. This geospatial variant of the TIFF format is often referred to as GeoTIFF. Although not common, the TIFF specification permits some simple compression schemes including the run-length encoding and Huffman encoding algorithms mentioned previously.

4.9.8.2 Wavelet-Based Formats

Of the compressed image formats that have gained broad support in geospatial applications, the formats that rely on wavelet-based algorithms are probably the most common. There are three main wavelet-based algorithms:

- JPEG2000. An ISO-standard format, this is the only format of the group with a public specification. JPEG2000 was originally created to supplant the JPEG standard and offers improved compression ratios. This format lacks well-specified multispectral support.
- MrSID. A proprietary format originally developed at Los Alamos National Laboratory in 1992. This format is owned by LizardTech, a Seattle-based geospatial company. The latest iteration of this format, MrSID Generation 4 (MG4) includes support for multispectral imagery and LiDAR imagery.
- ECW. A proprietary format originally developed by Earth Resource Mapping and now owned by Hexagon Geospatial.

4.9.9 *Conclusion*

Image compression is a promising solution for the challenges of growing image sizes and increased image availability. Indeed, it can be a compelling solution for many organizations, especially organizations with large amounts of imagery and groups operating in bandwidth-constrained environments. However, image compression is not suited to all organizations and requires careful considerations of the desired applications of the imagery first.

5 SECTION AUTHORS

Section 1: John Faundeen (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA) and George Percivall (Open Geospatial Consortium, Wayland, MA, USA)

Section 2: John Faundeen (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA)

2.1: David Giaretta (Giaretta Associates Ltd, and PTAB Ltd, Yetminster, UK)

2.2: Mark Ferguson (National Archives and Records Administration, National Records Management Program, Records Management Services, Broomfield, CO, USA)

2.3: Steve Morris (North Carolina State University Libraries, Raleigh, NC, USA)

2.4: John Faundeen (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA) and B. Lazorchak (Library of Congress National Digital Information Infrastructure and Preservation Program, Washington, DC, USA)

2.5: Matt Martens (Stinger Ghaffarian Technologies, Sioux Falls, SD, USA), Chris Doescher and Tom Sohare (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA)

2.6: John Faundeen (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA)

2.7: John Faundeen (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA)

2.8: H.K. Ramapriyan (Science Systems and Applications, Inc., Lanham, MD, and NASA Goddard Space Flight Center, Greenbelt, MD, USA)

2.9: Vivek Navale (Center for Information Technology, National Institutes of Health, Bethesda, MD, USA)

2.10: John Faundeen (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA)

2.11: David Thau (Developer Advocate, Google Earth Engine, Mountain View, CA)

2.12: John Faundeen (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA)

Section 3: John Faundeen (U.S. Geological Survey, EROS Center, Sioux Falls, SD, USA)

3.1: Viv Hutchison (U.S. Geological Survey, Core Science Systems, Lakewood, CO, USA)

3.2: Siri Jodha S. Khalsa (University of Colorado, Boulder, CO, USA)

3.3: H.K. Ramapriyan (Science Systems and Applications, Inc., Lanham, MD, and NASA Goddard Space Flight Center, Greenbelt, MD, USA), N.L. James and J. Behnke (NASA Goddard Space Flight Center, Greenbelt, MD, USA)

3.4: Liping Di (Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA USA)

3.5: C.S. Lynnes (National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, MD, USA) and D.J. Newman (Aeronautics, Astronautics, and Engineering Systems,

M.I.T. School of Science, Cambridge, MA, USA)

3.6: Paul Uhlir (Scholar, National Academy of Sciences, and Consultant, Data Policy and Management, Annandale, VA 22003)

Section 4: George Percivall (Open Geospatial Consortium, Wayland, MA, USA)

4.1: Matthew Purss, Adam Lewis, Alex Ip, Leo Lymburner, Simon Oliver, Joshua Sixsmith, Fuqin Li, Rachel Melrose, Lan-Wei Wang, Norman Mueller (Geoscience Australia, Canberra, Australia) and Roger Edberg (National Computational Infrastructure, Australian National University, Acton, Australia)

4.2: Karl Benedict (College of University Libraries and Learning Sciences, University of New Mexico, Albuquerque, NM, USA), Shirley Baros and John Savickas (Earth Data Analysis Center, University of New Mexico, Albuquerque, NM, USA)

4.3: Kumar Navulur (DigitalGlobe, Westminster, CO, USA)

4.4: Peter Baumann (Jacobs University, Bremen, Germany)

4.5: Peter Becker (Esri, Redlands, CA, USA)

4.6: Russ Rew (UCAR Unidata Program, Boulder, CO, USA)

4.7: Stephen Foreman (World Meteorological Organization, Geneva, Switzerland)

4.8: Lucio Colaiacomo (European Union Satellite Center, Madrid, Spain)

4.9: Michael Rosen, J. Dominguez (LizardTech, Seattle, WA, USA) and Jeff Young (LizardTech, Denver, CO, USA)

6 REFERENCES

ABARES 2014. Australian ground cover reference sites database.

Altman, M., J. Bailey, K. Cariani, J. Corridan, J. Crabtree, B. Dessy, M. Gallinger, A. Goethals, A. Grotke, C. Hartman, B. Lazorchak, J. Mandelbaum, C.M. Morris, T. Owens, M. Phillips, J. Spencer, H. Tibbo, T. Walters and K. Wittenberg. 2013. 2014 National agenda for digital stewardship. National Digital Stewardship Alliance. http://www.digitalpreservation.gov/ndsa/documents/2014NationalAgenda.pdf.

ANZLIC. 2012. National Nested Grid (NNG) Specification Guideline.

APA. 2014a. Digital preservation glossary: http://www.alliancepermanentaccess.org/index.php/consultancy/dpglossary/.

APA. 2014b. Integrated view of digital preservation http://www.alliancepermanentaccess.org/index.php/community/common-vision/.

Arvidson, T., J. Gasch and S.N. Goward. 2001. Landsat 7's long-term acquisition plan - An innovative approach to building a global imagery archive. *Remote Sensing of Environment.* 78:13-26. doi: 10.1016/S0034-4257(01)00263-2.

Asrar, G. and H. Ramapriyan. 1995. Data and information system for Mission to Planet Earth. *Remote Sensing Reviews.* 13:1-25.

Australian Academy of Science. 2009. An Australian strategic plan for Earth Observations from Space. *Australian Academy of Science.*

Australian Government Information Management Office. 2013. Australian public service big data strategy. Department of Finance and Deregulation, Australian Government Information Management Office, Commonwealth of Australia.

Baumann, P. 1994. Management of Multidimensional Discrete Data. *Very Large Data Base Journal.* 3:401-444.

Baumann, P. 1999. A database array algebra for spatio-temporal data and beyond. Fourth International Workshop on Next Generation Information Technologies and Systems (NGITS '99). July 5-7. Zikhron Yaakov, Israel. Lecture Notes on Computer Science. 1649. Springer Verlag.

Baumann, P. 2009. Web Coverage Processing Service (WCPS) Language Interface Standard. Version 1.0.0 *Open Geospatial Consortium. Document* # OGC 08-068r2. Version 1.0.0.

Baumann, P. 2010a. The OGC Web Coverage Processing Service (WCPS) Standard. *Geoinformatica.*14(4):447-479. doi 10.1007/s10707-009-0087-2.

Baumann, P. 2010b. Beyond Rasters: Introducing the new OGC Web Coverage Service 2.0. *Proceedings ACM SIGSPATIAL GIS*. San Jose, California, USA. November 2-5.

Baumann, P., S.Feyzabadi and C. Jucovschi. 2010. Putting pixels in place: A storage layout language for scientific data. *Proceedings IEEE ICDM Workshop on Spatial and Spatiotemporal Data Mining*. Sydney, Australia.

Baumann, P., (ed.). 2012. OGC Web Coverage Service – Core. Version 2.0. OGC 09-110r4.

Baumann, P., P. Mazzetti, J. Ungar, R. Barbera, D. Barboni, A. Beccati, L. Bigagli, E. Boldrini, R.Bruno, A. Calanducci, P. Campalani, O.Clement, A. Dumitru, M. Grant, P. Herzig, G. Kakaletris, J. Laxton, P. Koltsida, K. Lipskoch, A.R. Mahdiraji, S. Mantovani, V. Merticariu, A. Messina, D. Misev, S. Natali, S. Nativi, J. Oosthoek, J. Passmore, M. Pappalardo, A.P. Rossi, F. Rundo, M. Sen, V. Sorbera, D. Sullivan, M. Torrisi, L. Trovato, , M.G. Veratelli and S. Wagner. 2015. Big data analytics for Earth Sciences: the EarthServer approach. *International Journal of Digital Earth.* 27pp. doi:10.1080/17538947.2014.1003106.

Baumann, P. and V. Merticariu. 2015. On the Efficient Evaluation of Array Joins. *Proceedings Workshop Big Data in the Geo Sciences* (co-located with IEEE Big Data). Santa Clara, California, USA. October 29.

Baumann, P., A. Dumitru and V. Merticariu. 2015. Grooming Big Data from Afar. *Proceedings 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Bellevue, Washington. 64:1-4.

Baumann, P., E.Hirschorn and J. Maso. 2016. OGC Coverage Implementation Schema (CIS) version 1.1. OGC 09-146r3. https://portal.opengeospatial.org/files/64632.

Baumgarten, T., H. Hoenig, and P. Baumann. 2014. Big Earth Data – the digitized planet. (TV documentary).

Baxter, A. 2014. SSD vs HDD. https://www.storagereview.com/ssd_vs_hdd.

Behnke, J., T.H. Watts, B. Kobler, D. Lowe, S. Fox and R. Meyer. 2005. EOSDIS Petabyte Archives: Tenth Anniversary. *Proceedings 22nd IEEE/13th* NASA Goddard Conference on Mass Storage Systems and Technologies (MSST'05). 13p.

Berk, A., G.P. Anderson, L.S. Bernstein, P.K Acharya, H. Dothe, M.W. Matthew, S.M. Adler-Golden, J.H. Jr. Chetwynd, S.C. Richtsmeier, B. Pukall, C.L. Allred, L.S. Jeong and M.L. Hoke. 1999. MODTRAN4 radiative transfer modeling for atmospheric correction. *Proceedings SPIE 3756. Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research* III, 348. doi: 10.1117/12.366388.

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. http://openaccess.mpg.de/Berlin-Declaration.

Berners-Lee, T. 2006. Linked data: http://www.w3.org/DesignIssues/LinkedData.html.

Bindschadler, R. 2003. Landsat coverage of the Earth at high latitudes. *Photogrammetric Engineering & Remote Sensing*. 69:1333-1339. doi: 10.14358/PERS.69.12.1333.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. Sustainable economics for a digital planet: Ensuring long-term access to digital information. http://blueribbontaskforce.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Boyd, D. and K. Crawford. 2012. Critical Questions for Big Data. *Information, Communication and Society*. 15: 662-679. doi: 10.1080/1369118X.2012.678878.

Bromley, D. Allen. 1991. Principles on full and open access to "Global Change" data, policy statements on data management for global change research. United States Office of Science and Technology Policy.

Byers, F. 2003. Information technology: Care and handling of DVDs: A guide for librarians and archivists. Council on Library and Information Resources, and National Institute of Standards and Technology.

CEOS and ESA. 2014. The Earth Observation Handbook: Special Edition for Rio+20 http://eohandbook.com/index14.html.

Clinton, W. J. 1993. Executive Order 12906 - Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure. Washington, DC: Federal Register.

CODATA. 2014. Nairobi Principles on Data Sharing for Science and Development in Developing Countries. CODATA. https://rd-alliance.org/sites/default/files/attachment/NairobiDataSharingPrinciples.pdf.

CODATA. 2015. The value of open data. Group on Earth Observations. https://www.earthobservations.org/documents.php?smid=200.

Coleman, T.W. 2013. Big Data Deemed Mission critical: http://www.internationalpolicydigest.org/2013/06/19/big-data-deemed-mission-critical/.

Columbia University Center for International Earth Science Information Network. 2002. Policy for preservation of digital resources: http://www.ciesin.org/policies.html.

Computer History Museum. 2015. Memory and storage: http://www.computerhistory.org/revolution/memory-storage/8/intro.

Consultative Committee for Space Data Systems. 2012. Reference model for an Open Archival Information System (OAIS) Magenta Book: http://public.ccsds.org/publications/archive/650x0m2.pdf.

Cunningham, S.C., R. MacNally, J. Read, P.J. Baker, M.White, J.R. Thomson and P. Griffioen. 2009. A robust technique for mapping vegetation condition across a major river system. *Ecosystems.* 12:207-219. doi: 10.1007/s10021-008-9218-0.

Data Sharing Working Group. 2015. GEOSS data sharing principles Post-2015. Group on Earth Observations (2014).https://www.earthobservations.org/documents/dswg/Annex%20III%20-%20GEOSS%20Data%20Sharing%20Principles%20Post-2015.pdf.

de La Beaujardière, J. (ed.). 2002. Web Map Service Implementation Specification. Open GIS Consortium implementation specification. Version 1.1.1.

de la Beaujardiere, J. (ed.). 2006. OpenGIS Web Map Server Implementation Specification, Version 1.3.0. OGC® 06-042: Open Geospatial Consortium.

de la Torre, M. (ed.). 2002. Assessing the values of cultural heritage. The Getty Conservation Institute: http://www.getty.edu/conservation/publications_resources/pdf_publications/pdf/assessing.pdf.

Devarakonda, R., B. Shrestha, G. Palanisamy, L. Hook, T. Killeffer, M. Krassovski, T. Boden, R. Cook, L. Zolly, V. Hutchison, M. Frame, A. Cialella and K. Lazer. 2014. OME: Tool for generating and managing metadata to handle Big Data. *Proceedings 2014 IEEE International Conference on Big Data*. 8-10 pp.

Digital Curation Centre. 2010. How to appraise and select research data for curation: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data.

Digital Globe Foundation. 2014. http://www.digitalglobefoundation.org/.

Digital Preservation Coalition. 2015. Digital preservation handbook: http://www.dpconline.org/advice/preservationhandbook.

Dumitru, A., V. Merticariu and P. Baumann. 2014. Exploring cloud opportunities from an array database perspective. *Proceedings ACM SIGMOD* workshop on data analytics in the cloud (DanaC'2014), June 22 - 27, 2014, Snowbird, USA.

Erway R. 2012. You've got to walk before you can run: First steps for managing born-digital content received on physical media. Online Computer Library Center, Inc.: http://www.oclc.org/content/dam/research/publications/library/2012/2012-06.pdf.

Federal Geographic Data Committee Metadata Ad Hoc Working Group. 1998. Content standard for digital geospatial metadata. https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf.

Federal Geographic Data Committee. 1998. Content standard for digital geospatial metadata (CSDGM). Version 2. http://www.fgdc.gov/metadata/geospatial-metadata-standards#csdgm.

Federal Geographic Data Committee. 2014a. Guidance on the selection and appraisal of geospatial content of enduring value.

Federal Geographic Data Committee. 2014b. National Geospatial Data Asset Management Plan http://www.fgdc.gov/policyandplanning/a-16/ngda-management-plan.

Federal Geographic Data Committee. 2015. https://www.fgdc.gov/.

Fielding, Roy. 2000. Architectural styles and the design of network-based software architectures. Ph.D. Dissertation, Information and Computer Science. University of California, Irvine. 162pp. pdf.

Franks, P. 2013. Records and Information Management. Neal-Schuman: Chicago. 424 pp.

Freed. N., J. Klensin and T. Hansen. 2013. Media type specification and registration procedures. Memo. Internet Engineering Task Force Request for Comments 6838: https://tools.ietf.org/html/rfc6838.

Fritsch, D. and D. Stallmann. 2000. Rigorous photogrammetric modelling processing of high-resolution satellite imagery. *International Archives of Photogrammetry and Remote Sensing*. 33:313-321.

Furtado, P. and P. Baumann. 1999. Storage of multi-dimensional arrays based on arbitrary tiling. ICDE'99. March 23-26, 1999. Sydney, Australia.

Geospatial Multistate Archive and Preservation Partnership. 2010. GeoMAPP Interim Report: 2007-2009. North Carolina Center for Geographic Information and Analysis. North Carolina Department of Cultural Resources.

Geospatial Multistate Archive and Preservation Partnership. 2011a. GeoMAPP Final Report 2007-2011. North Carolina Center for Geographic Information and Analysis, North Carolina Department of Cultural Resources: http://www.digitalpreservation.gov/multimedia/documents/GeoMAPP_FinalReport_final_20111231.pdf.

Geospatial Multistate Archive and Preservation Partnership. 2011b. GeoMAPP Geoarchiving Business Planning Toolkit: http://www.geomapp.com/publications_categories.htm#busplan.

Geospatial Multistate Archive and Preservation Partnership. 2011c.

Gibb, R., A. Raichev and M. Speth. 2013. The rHEALPix discrete global grid system. https://datastore.landcareresearch.co.nz/dataset/rhealpix-discrete-global-grid-system.

Group on Earth Observations. 2005. The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan. http://www.earthobservations.org/docs/10-Year%20Implementation%20Plan.pdf.

Gruen, A. and L. Zhang. 2002. Sensor modelling for aerial mobile mapping with three-line-scanner (TLS) imagery. *International Archives of Photogrammetry and Remote Sensing.* 34:139-146.

Gutman, G., R. Byrnes, J. Masek, S. Covington, C. Justice, S. Franks and R. Headley. 2008. Towards monitoring land-cover and land-use changes at a global scale: The global land survey 2005. *Photogrammetric Engineering and Remote Sensing.* 74:6-10.

Hanai, K. 2013. The Archival Stability of Metal Particulate Tapes. Fuji Film presentation, Library of Congress.

Hanai, K. and K. Kakuishi. 2002. The storage stability of metal particulate media: Chemical analysis and kinetics of lubricant and binder hydrolysis. *Proceedings 10th Goddard Conference on Mass Storage Systems*. Greenbelt, MD. 311-315 pp.

Hart, P. and C. Saunders.1997. Power and trust: critical factors in the adoption and use of electronic data interchange. *Organization Science.* 8:23-42. doi: 10.1287/orsc.8.1.23.

Hansen, J., M. Sato and R. Ruedy. 2012. Perception of climate change. *Proceeding National Academy of Science.* 109: 14726-14727, E2415-E2423, doi:10.1073/pnas.1205276109.

Hattori, S., T. Ono, C.S. Fraser and H. Hasegawa. 2000. Orientation of high-resolution satellite images based on affine projection. *International Archives of Photogrammetry and Remote Sensing.* 33:359-366.

High Level Expert Group on Scientific Data. 2010. Riding the wave: How Europe can gain from the rising tide of scientific data. European Commission, European Union.

Hilker, T., M.A. Wulder, N.C. Coops, N. Seitz, J.C. White, F. Gao, J.G. Masek and G. Stenhouse. 2009. Generation of dense time-series synthetic Landsat data through data blending with MODIS using a spatial and temporal adaptive reflectance fusion model. *Remote Sensing of Environment* 113:1988-1999. doi: 10.1016/j.rse.2009.05.011.

International Federation of Data Organizations for Social Science. 2015. Data preservation: http://ifdo.org/wordpress/?page_id=18.

Inui, M. 2013. International standards and off-line archiving through the use of recordable optical discs. *Proceedings Imaging Science and Technology Archiving Conference.* pp.137-142.

Irish R.R., J.L. Barker, S.N. Goward and T. Arvidson. 2006. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. *Photogrammetric Engineering and Remote Sensing.* 72:1179-1188. doi: 10.14358/PERS.72.10.1179.

ISO Technical Committee 20 S 13. 2006. Producer-Archive Interface Methodology Abstract Standard. *International Organization for Standardization*.

ISO Technical Committee 20 S 13. 2012a. Audit and Certification of Trustworthy Digital Repositories. http://public.ccsds.org/publications/archive/652x0m1.pdf.

ISO Technical Committee 20 S 13. 2012b. Reference Model for an Open Archival Information System (OAIS). *International Organization for Standardization*. http://public.ccsds.org/publications/archive/650x0m2.pdf.

ISO Technical Committee 20 S 13. 2014. *International Organization for Standardization*.

ISO. 2001. ISO Standard 15489-1: Information and Documentation - Records Management: Part 1: General Website. ISO: Geneva. https://www.iso.org/obp/ui/#iso:std:iso:15489:-1:ed-1:v1:en.

ISO. 2003. 19115:2003 Standard for Geographic Information- Metadata: http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020.

Kahle, B. and A. Medlar. 1991. An information system for corporate users: Wide area information servers. *Online.* 15:56-60.

Kerekes, Z. 2015. Charting the rise of the solid state disk market: http://www.storagesearch.com/chartingtherise-eofssds.html.

Komorowski, M. 2009. A history of storage cost: http://www.mkomo.com/cost-per-gigabyte.

Komorowski, M. 2014. A history of storage cost - Update: http://www.mkomo.com/cost-per-gigabyte-update.

Laney, D. 2001. 3D data management: Controlling data volume, velocity and variety. pdf. Technical Report, META Group Inc.

Lazorchak, B. 2010. Library of Congress Takes Leadership. *ArcNews Summer* 32:1.

Lewis, A., L. Lymburner, M. Purss, B. Brooke, B. Evans, A. Ip, A. Dekker, J. Irons, S. Minchin and N. Mueller. 2016. Rapid, high-resolution detection of environmental change over continental scales from satellite data - the Earth Observation Data Cube. *International Journal of Digital Earth*. 9(1):106-111.

Li, F., D.L. Jupp, S. Reddy, L. Lymburner, N.Mueller, P.Tan and A. Islam. 2010. An evaluation of the use of atmospheric and BRDF correction to standardize Landsat Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.* 3:257-270. doi: 10.1109/JSTARS.2010.2042281.

Li, F., D.L. Jupp, M. Thankappan, L. Lymburner, N. Mueller, A. Lewis and A. Held. 2012. A Physics-based atmospheric and BRDF correction for landsat data over mountainous terrain. *Remote Sensing of Environment.* 124: 756-770. doi: 10.1016/j.rse.2012.06.018.

Li, F., D.L. Jupp, M. Thankappan, M. Paget, A. Lewis and A. Held. 2013. The variability of satellite derived surface BRDF shape over Australia from 2001 to 2011. *Proc. 2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. 255-258. doi: 10.1109/IGARSS.2013.6721140.

Library of Congress. 2014. Sustainability of digital formats planning for Library of Congress collections: http://www.digitalpreservation.gov/formats/index.shtml.

Lynnes, C., M. Hegde, C. Smit, J. Pan, K.C. Bryant, C. Chidambaram and P. Zhao. 2013. Volume, variety and veracity of Big Data analytics in NASA's Giovanni tool. *Proceedings American Geophysical Union.* Fall Meeting Abstracts.

Mattmann, C.A. 2013. Computing: A vision for data science. *Nature* 493:473-475. doi: 10.1038/493473a.

McCallum, J.C. 2014. Disk Drive Prices (1955-2019). https://jcmit.net/diskprice.htm.

McCallum, S. 2006. A look at new information retrieval protocols, SRU, OpenSearch/A9, CQL, and XQuery. *Proc. World Library and Information Congress*: 72nd IFLA General Conference and Council.

McGarva, G., S. Morris and G. Janee. 2009. Technology watch report: Preserving geospatial data. *Digital Preservation Coalition:* http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee.pdf.

Melrose, R., J. Kingwell, L. Lymburner and R. Coghlan. 2013. Murray-Darling Basin vegetation monitoring project: Using time-series Landsat satellite data for the assessment of vegetation. pdf. Geoscience Australia. doi: 10.11636/Record.2013.037.

Michener, W.K. 2015. Ecological data sharing. *Ecological Informatics*. doi: 10.1016/j.ecoinf.2015.06.010.

Moore, M. and D. Lowe. 2002. Providing rapid access to EOS data via Data Pools. *Proceedings SPIE Earth Observing Systems VII*. Pgs. 7-10.

Moore, R.L., J. D'Aoust, R.H. McDonald and D. Minor. 2007. Disk and tape storage cost models: https://libraries.ucsd.edu/chronopolis/_files/publications/dt_cost.pdf.

Morris, S. 2013. Issues in the appraisal and selection of geospatial data. National Digital Stewardship Alliance: http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Appraisal_Selection_report_final_102413.pdf.

Muir, J., M. Schmidt, D. Tindall, R. Trevithick, P. Scarth and J. Stewart. 2011. Field measurement of fractional ground cover: A technical handbook supporting the Australian collaborative land use and management program. Queensland Department of Environment and Resource Management, Australian Bureau of Agricultural and Resource Economics and Sciences: Canberra.

Murray-Darling Basin Authority. 2009. The Basin Plan: a concept statement. Murray-Darling Basin Authority.

Murray-Darling Basin Authority. 2011. The Living Murray Annual Environmental Watering Plan 2011-12.pdf. Murray-Darling Basin Authority: http://www.mdba.gov.au/publications/mdba-reports/living-murray-annual-environmental-watering-plan-2011%E2%80%9312.

Murray-Rust, P., C. Neylon, R. Pollock and J. Wilbanks. 2010. Panton principles for open data in science. http://pantonprinciples.org/.

Myneni, R.B., F.G. Hall, P.J. Sellers and A.L. Marshak. 1995. The interpretation of spectral vegetation indexes. *IEEE Transactions on Geoscience and Remote Sensing*. 33:481-486. doi: 10.1109/36.377948.

NASA GSFC. 2014. GCMD Keyword Community Guide: http://gcmd.nasa.gov/learn/keywords.html.

NASA GSFC. 2015. Common Metadata Repository (CMR) Life-Cycle Document (Rev. 1). Earth Science Data and Information System Project. Code 423, NASA Goddard Space Flight Center, Greenbelt, MD.

NASA. 1998. The Landsat-7 Science Data User's Handbook . NASA Goddard Space Flight Center, Greenbelt, MD.

National Archives and Records Administration. 2001. National Oceanic and Atmospheric Administration, Request for Records Disposition Authority N1-370-00-06.

National Archives and Records Administration. NARA 1571 Archival storage standards. 2002. http://www.archives.gov/foia/directives/nara1571.pdf.

National Archives and Records Administration. 2004. Strategic directions: Flexible scheduling.

National Archives and Records Administration. 2006. Appraisal policy of the National Archives and Records Administration: http://www.archives.gov/records-mgmt/publications/appraisal-policy.pdf.

National Archives and Records Administration. 2007. Frequently Asked Questions (FAQs) about selecting sustainable formats for electronic records.

National Archives and Records Administration. 2007a. National Archives and Records Administration Strategic Directions: Appraisal Policy.

National Archives and Records Administration. 2007b. Strategic directions appraisal policy: http://www.archives.gov/records-mgmt/initiatives/appraisal.html.

National Archives and Records Administration. 2008. National Oceanic and Atmospheric Administration, Request for Records Disposition Authority N1-370-03-10.

National Archives and Records Administration. 2009. Code of Federal Regulations, Title 36, Chapter XII, Subchapter B, Records Management.

National Archives and Records Administration. 2012. Knowledge Area (KA) 3: Records scheduling, participant guide.

National Archives and Records Administration. 2014. General records schedules, Transmittal 23. National Archives and Records Administration.

National Digital Stewardship Alliance, Geospatial Content Team. 2013. Issues in the Appraisal and Selection of Geospatial Data: An NDSA Report. http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_AppraisalSelection_report_final102413. pdf.

National Information Standards Organization. 1988. Z39.50.1988. Information retrieval service definition and protocol specification for library applications. https://www.loc.gov/z3950/agency/markup/01.html.

National Oceanic and Atmospheric Administration. 2008. NOAA procedure for scientific records appraisal and archive approval: Guide for data managers. U.S. Department of Commerce: https://www.ngdc.noaa.gov/wiki/images/0/0b/NOAA_Procedure_document_final.pdf.

National Science and Technology Council. 2013. National strategy for civil Earth observations: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/nstc_2013_earthobsstrategy.pdf.

National States Geographic Information Council (various dates). About the GIS Inventory.: http://gisinventory.net/about-the-gis-inventory.

Navale V. 2005. Predicting life expectancy of magnetic and optical media. RLG *DigiNews* 9: http://worldcat.org/arcviewer/1/OCC/2007/07/10/0000068919/viewer/file1.html#article3.

Neill. M. and Hewson, W. 2013. Smarter Uncle Sam: The big data forecast. http://www.meritalk.com/smarteruncleSam.

NIST and Library of Congress. 2005. Optical media longevity study: http://www.itl.nist.gov/iad/894.05/loc/definitions.html.

North Carolina Center for Geographic Information and Analysis. 2010. Geospatial multistate archive and preservation partnership interim report 2007-2009. *North Carolina State Archives.*

Nottingham M. 2010. Web Linking. Website. IETF Trust. https://tools.ietf.org/html/rfc5988.

Nottingham, M. and R. Sayre. 2005. The Atom syndication format Website. *The Internet Society*: http://www.ietf.org/rfc/rfc4287.txt.

Online Computer Library Center (OCLC). 2013. OCLC Research activity report. Dublin, Ohio. http://www.oclc.org/research/publications/library/2013/2012activityreport.pdf.

Office of Management and Budget. 2000. OMB Circular A-130. Executive Office of the President of the United States.

Office of Management and Budget. 2010. Geospatial line of business. OMB Circular A-16 Supplemental Guidance. Office of the President of the United States.

Office of Management and Budget. 2000. Circular No. A-130 (Revised). Executive Office of the President. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A130/a130trans4.pdf.

Office of the Australian Information Commissioner. 2011. Principles on open public sector information: https://www.oaic.gov.au/information-policy/information-policy-resources/principles-on-open-public-sector-information.

Open Access Directory. 2015. http://oad.simmons.edu/oadwiki/Declarations_in_support_of_OA.

Open Knowledge. *Open Data Handbook* (2015). http://opendatahandbook.org/guide/en/what-is-open-data/.

Open Streetmap Foundation. 2014. Copyright and License: http://www.openstreetmap.org/copyright.

Overpeck, J.T., G.A. Meehl, S. Bony and D.R. Easterling. 2011. Climate data challenges in the 21st Century. *Science. 331*:700-702. doi: 10.1126/science.1197869.

Oxera Consulting Ltd. 2013. What is the economic impact of Geo Services? Website. Oxera Consulting Ltd.: http://www.oxera.com/Latest-Thinking/Publications/Reports/2013/What-is-the-economic-impact-of-Geo-services.aspx.

Pearce-Moses, R. 2005. A glossary of archival and records terminology. *Society of American Archivists.* http://www2.archivists.org/glossary/terms/d/disposal.

Percivall, G. 2013. Geodata fusion study by the Open Geospatial Consortium. *Proc. SPIE, Geospatial InfoFusion III.* 87470A 8747 doi: 10.1117/12.2016226.

Peters, M. 2014. The technical and operational values of Barium Ferrite tape media. Website. Fuji Film: http://www.storagenewsletter.com/rubriques/tapes/technical-and-operational-values-of-barium-ferrite-tape-media/.

Pingdom, R. 2011. Would you pay $7,260 for a 3TB drive? Charting HDD and SSD prices over time: http://royal.pingdom.com/2011/12/19/would-you-pay-7260-for-a-3-tb-drive-charting-hdd-and-ssd-prices-over-time/.

Poli, D., L. Ahang and A. Gruen. 2004. Orientation of satellite and airborne imagery from multi-line push-broom sensors with a rigorous sensor model. *International Archives of Photogrammetry and Remote Sensing.* 35: 130-135. doi: 10.1.1.140.7214.

Porter, J. and C. Duke. 2013. Ethics of data sharing and reuse in Biology. *BioScience*. 63:483-489. doi:10.1525/bio.2013.63.6.10.

Purss, M.B., A. Lewis, R. Edberg, A. Ip, J. Sixsmith , G.Frankish, T.Chan, B.Evans and L. Hurst. 2013. Exploiting data intensive applications on high performance computers to unlock Australia's Landsat Archive. *Geophysical Research Abstracts.* 15:EGU2013-8049.

Purss, M.B.J., A. Lewis, S.Oliver, A. Ip, S.Sixsmith, B. Evans, R. Edberg, G. Frankish, L. Hurst and T. Chan. 2015. Unlocking the Australian Landsat archive from dark data to high performance data infrastructures. *Geophysical Resarch Journal.* 6:135-140. doi: 10.1016/j.grj. 2015. 02.010.

Raster Data Manager (rasdaman). 2016a. Big Earth Data Standards. http://www.rasdaman.org/.

Raster Data Manager (rasdaman). 2016b. Educational Material. http://www.rasdaman.org/.

Ravanbakhsh, M., L.W. Wang, C.S. Fraser and A. Lewis. 2012. Generation of the Australian Geographic Reference Image Through Long-Strip ALOS Prism Orientation. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.* 39:225-229.

RECODE Project. 2015. *Policy Guidelines for Open Access and Data Dissemination and Preservation*. European Commission. https://trilateralresearch.co.uk/wp-content/uploads/2018/09/RECODE-D5.1-POLICY-RECOMMENDATIONS-_FINAL.pdf.

Reed,.C., R. Singh, R. Lake, J. Lieberman and M. Maron. 2006. An Introduction to GeoRSS: A standards based approach for Geo-enabling RSS feeds. *OpenGeospatial Consortium.* OGC 06-050r3.

Richter, R., T. Kellenberger and H. Kaufmann. 2009. Comparison of topographic correction methods. *Remote Sensing.* 1:184-196. doi: 10.3390/rs1030184.

Ronsdorf, C., P. Mason, J. Holmes, U. Gerber, A. Streilein, M. Bos, A. Shaon, K. Naumann, M. Kirstein, G. Samuelsson, M. Rantala, S. Kvarteig, L.Ralsberg, J. Svennewall and W. Stobel. 2014. GI+100: Long-term preservation of digital geographic information: 16 Fundamental Principles Agreed by National Mapping Agencies and State Archives: http://www.eurosdr.net/research/project/eurosdr-archiving-working-group.

Saffady, W. 2004. Records and information management: Fundamentals of professional practice. *ARMA International*: Lenexa, KS.

Sawyer, D., L. Reich, D. Giaretta, P. Mazal, C. Huc, M. Nonon-Latapie and N. Peccia. 2002. The Open Information System (OAIS) reference model and its usage. http://ccsds.cosmos.ru/publications/documents/SO2002/SPACEOPS02_P_T5_39.PDF.

Scarth, P., A. Roder and M. Schmidt. 2010. Tracking grazing pressure and climate interaction - the role of Landsat fractional cover in time series analysis. *Proc. 15th Australasian Remote Sensing and Photogrammetry*.

Sellers, P.J. 1985. Canopy reflectance, photosynthesis, and transpiration. *International Journal of Remote Sensing*. 6:1335-1372. doi: 10.1080/01431168508948283.

Shahani, C.J., M.H. Youket and N. Weberg. 2004. Compact disc service life: An investigation of the estimated service life of prerecorded compact discs (CD-ROM). Library of Congress, Preservation Research and Testing Series: http://www.loc.gov/preservation/resources/rt/CDservicelife_rev.pdf.

Showstack, R. 2014. Award Program recognizes efforts to protect geoscience data. *EOS Transactions. American Geophysical Union.* 95:1-12. doi: 10.1002/2014EO010002.

Sixsmith J., S. Oliver and L. Lymburner. 2013. A hybrid approach to automated landsat pixel quality. IEEE Geoscience and remote sensing symposium (IGARSS): 4146-4149. http://dx.doi.org/10.1109/IGARSS.2013.6723746.

SKOS. 2006. Simple knowledge organization system. http://www.w3.org/2004/02/skos/.

Skupsky, D. 1991. Records retention procedures. Information Requirements Clearinghouse, Denver.

Slattery, O. 2004. Stability comparison of recordable optical discs: A study of error rates in harsh conditions. *J. Res. National Institute of Standards and Technology*. 109:517-524.

Smith, I. 2014. Cost of hard drive storage space.

Soenen, S.A., D.R. Peddle and C.A. Cobum. 2005. SCS+C: A modified sun-canopy-sensor topographic correction in forested terrain. *IEEE Transactions on Geoscience and Remote Sensing.* 43:2148-2159. doi: 10.1109/TGRS.2005.852480.

Stroeve, J., M.H. Marika, W. Meier, T. Scambos and M. Serreze. 2007. Arctic sea ice decline: Faster than forecast. *Geophysical Research Letters.* 34: L09501. doi: 10.1029/2007GL029703.

The Commonwealth of Australia. 2009. Budget Measures: Budget Paper No. 2.

The Commonwealth of Australia. 2013. Australia's satellite utilization policy. Department of Industry, Innovation, Science, Research and Tertiary Education. https://www.industry.gov.au/sites/default/files/May%202018/document/pdf/australias_satellite_utilisation_policy.pdf?acsf_files_redirect.

The Hague Declaration on Knowledge Discovery in the Digital Age. *LIBER*. 2014. http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/.

The Royal Society. 2012. *Science as an open enterprise.* RSUK. https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf.

Toutin, T. 2003. Block bundle adjustment of Landsat 7 ETM+ images over mountainous areas. *Photogrammetric Engineering & Remote Sensing.* 69:1341-1349. doi: 10.14358/PERS.69.12.1341.

U.S. Department of Defense. 2007. Electronic records management software applications design criteria standard. pdf. U.S. Department of Defense.

U.S. Geological Survey. 2013. Landsat 8 Fact Sheet. Earth Resources Observation and Science Center, U.S. Geological Survey.

van Bogart, J.W.C. 1994. Media stability studies. National Media Laboratory.

Vretanos, P.A. (ed.). 2005. Web feature service implementation specification. Version 1.1.0. Vol. OGC 04-094: Open Geospatial Consortium.

Watanabe, A. 2013. Optical library system with extended error coding for long-term preservation. *Institute of Electrical and Electronics Engineers.Massive Storage Conference.* https://www.ieee.org/. Long Beach, CA.

Weiss, R. 2002. Environmental stability studies and life expectancies of magnetic media for use with IBM 3590 and quantum digital linear tape systems. National Archives and Records Administration: https://www.archives.gov/research/electronic-records/magnetic-media-study.pdf.

Whiteside, A. and John D. Evans (eds.). 2006. Web Coverage Service (WCS) Implementation Specification Version 1.1.0. Vol. 06-083r8: *Open Geospatial Consortium.*

Wikipedia.2014. History of hard disk drives. http://en.wikipedia.org/w/index.php?title=History_of_hard_disk_drives&oldid=644810953.

Wikipedia. 2015a. Active Archive: http://en.wikipedia.org/wiki/Active Archive.

Wikipedia. 2015b. Solid-state Drive: http://en.wikipedia.org/wiki/Solid-state_drive.

World Meteorological Organization. 2012. Volume I - General Meteorological Standards and Recommended Practices. WMO. Geneva.

World Meteorological Organization. 2013. Guide to the WMO Information System. WMO. Geneva.

World Meteorological Organization. 2013. Manual on the WMO Information System. Annex VII to the WMO Technical Regulations. WMO. Geneva.

World Meteorological Organization. 2014a. Manual on Codes. International Codes. Volume I.1. (Annex II to WMO Technical Regulations). Part A - Alphanumeric Codes. WMO. Geneva.

World Meteorological Organization. 2014b. Manual on Codes. International Codes. Volume I.2 (Annex II to WMO Technical Regulations). PART B - Binary Codes. PART C - Common Features to Binary and Alphanumeric Codes . WMO. Geneva.

Yang, J. and T. Johnson. 2000. Metadata for Earth Observing Systems (EOS) data gateway. *Geoscience and Remote Sensing Symposium*. pp. 1205-1207.

Zetta, Inc. 2015. The history of computer storage.

Zhu, Z. and C.E. Woodcock. 2012. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment.* 118:83-94. doi: 10.1016/j.