Long-term Metadata Management & Quality Assurance in Digital Curation

A Dissertation
Submitted In Partial Fulfilment Of
The Requirements for the Degree Of

MASTER OF SCIENCE

In

Network Centred Computing,
E-Commerce

in the

Faculty Of Science

The University of Reading

by

Arif Bin Siraj Shaon

August 22, 2005

University Supervisor: Prof. V. N. Alexandrov
Placement Supervisor: Kerstin Kleese - van Dam &
Mr. Shoaib Sufi

Acknowledgements

Utmost gratitude to Mr. Shoaib Sufi, the Deputy Group Leader of the CCLRC eScience Data Management group and also the placement supervisor of the project, for his constructive suggestions and comments on different aspects of the project throughout the project period.

Many thanks to the head of the CCLRC eScience Data Management group Kerstin Kleese van Dam, who also was the external supervisor and instigator of the project jointly with Mr. Shoaib Sufi.

Thanks also to Professor Vassil Alexandrov who was the University supervisor for the project.

Furthermore, the author is also grateful to the Council for the Central Laboratory of the Research Councils (CCLRC) for providing financial support for this project.

Abstract

With the rapid advancements in the realm of data management especially in terms of data volume, data quality and data availability; the necessity for adequate, well managed and high quality Metadata is becoming increasingly essential for successful long-term high quality data preservation. Data preservation over substantially long periods of time is needed to enable burgeoning amounts of data, being produced today, to be accessible with its quality intact and independent of associated software or hardware, to e.g. future scientists or researchers in order to aid in their experiments and research. From this perspective, well-managed and high quality metadata holds the key to avoiding the high cost of replicating 'expensive to produce' data as well as ensuring the proper and efficient use of these data over the long term with dynamic evolvements in related technologies.

This dissertation details the main achievements of a MSc. project that endeavours to address the aforementioned issues by conducting an in-depth research on various aspects of Metadata management, such as current approaches & techniques for Metadata management & quality assurance, existing tools, standards etc. In addition, as devised on the basis of the assessed results of this extensive and scrupulous investigation, this thesis provides detailed plan of work for the coming 2.5 years, which subsumes specific recommendations for developing a working prototype of metadata management system in the context of digital curation.

Contents

List of Figures

List of Tables

Nomenclature

ASCII	American Standard Code for Information Interchange
CIMI	Consortium for the Computer Interchange of Museum Information
CCLRC	Council for the Central Laboratory of the Research Councils
CIP	Catalogue Interoperability Protocol
CEOS	Committee on Earth Observation Satellites
CVFS	Comprehensive Versioning File System
CSDGM	Content Standards for Digital Geo-spatial Metadata
CWM	Common Warehouse Metamodel
DTD	Document Type Definition
DDI	Data Documentation Initiative
DC	Dublin Core
DIF	Directory Interchange Format
EAD	Encoded Archival Description
ESA	European Space Agency
FGDC	Federal Geographic Data Committee
GCMD	Global Change Master Directory
GILS	Global Information Locator Service
HLSI	The Higher Level Skills for Industry Repository
HTML	Hyper Text Markup Language
ISO	International Organization for Standardization
IDN	International Directory Network
LOM	Learning Object Metadata
MARC	Machine Readable Catalogue Format
MOF	Meta Object Facility
MeSH	Medicine Medical Subject Headings
NEES	Network for Earthquake Engineering Simulation
OIM	Open Information Model
OMG	Object Management Group
RDF	Resource Description Framework
SeSDL	Staff Development Library
SDISS	Scientific Data and Information Super Server
TEI	Text Encoding Initiative
XML	eXtensible Markup Language
XSL	Extensible Stylesheet Language
XMI	XML Metadata Interchange
UML	Unified Modelling Language

Chapter 1

Introduction

1.1 Introducing Metadata

"Metadata is the foundation for an effective data-centric information system"
-Robert Craig, Director, Data Warehousing and Business Intelligence Division, at Hurwitz Group Inc.
(Framingham, Mass.)

Owing to the past decades of fast changing information technologies, a radical change is discerned in data management landscape. This enormous change has resulted in a dramatic rise in both the volume of collected and analyzed data as well as the speed with which these operations are performed. New dissemination media, such as the Internet, have also contributed to making data available to a broader public.

Furthermore, many varied structures and formats of documents in electronic type are employed to manage the organizational information resources in different areas such as geography, museums, technology, literature, music, etc. This significantly enhances the performance of the operations on these documents as well as their management. However, on the other hand, different data formats may crucially affect the system integration and information sharing among these organizations [SKR03].

Under the challenges set by these new technical possibilities and enhancements, the word "Metadata" is becoming increasingly prevalent in the humanities and elsewhere, especially in relation to the online discovery and exchange of electronic information, with its concept growing in importance in a spectacular way [GEG04]. In essence, Metadata is an emerging approach to organising digital information in order to enhance retrieval, preservation and interoperability. In addition, judiciously crafted Metadata facilitates significantly enhanced effectiveness of searching, hence increased accessibility to information [MMC98].

1.2 Scope and Objectives of the Project

"Metadata Management refers to the content, structure, and designs necessary to manage the vocabulary and other metadata that describes statistical data, designs and processes. ... includes the development of metadata models ..., building metadata registries to organise the metadata ..., developing statistical terminologies which define and organise terms ..."
- Bargmeyer and Gillman, METIS 2000

Metadata, a fundamental role of the digital content, has now become an important part of the global information construction in planning, processing, restoring and managing. For example, in large distributed storage systems, avoiding bottlenecks is critical to achieving high performance and scalability. One potential bottleneck is metadata access. Although the size of metadata is generally small compared to the overall storage capacity of such a system, 50% to 80% of all file system accesses are to metadata [JDG85], so the careful management of metadata is crucial. In addition, it is recognised that in an indefinitely large resource space, effective management of networked information will increasingly rely on effective management of metadata.

In addition to its role in ensuring the proper and efficient management and use of data, metadata quality is also important for effective resource discovery. Poor quality Metadata compromises discovery in various ways including poor recall, poor precision, and inconsistency of search results and important resources being missed and remain unused. Standard and efficient approach to ensure Metadata quality is to incorporate quality assurance process into Metadata management [BJH03].

Furthermore, from the perspective of a business organisation, it is absolutely vital to turn data into reliable, reusable information assets to improve operational efficiency and customer centricity in order to survive and thrive. Within a business domain, well-managed metadata helps the users understand the nature of the information in addition to its location and what value the information provides to users. In other words, well-managed metadata is a key to eliminating information silos, rapidly deploying information solutions, integrating disparate data sources, finding and sharing information assets, and to making information coherent [SPMNA].

However, the main interest of this project, in fact, lies in some other more challenging and imperative realm of metadata management than those above mentioned generic metadata domains. Within complex information domains, such as scientific domain, large amounts of data are being generated and published. This large volume of published data needs to be maintained (i.e. preserved) and highly available (i.e. published) by long term curation process in order to serve it to the future generations. This will, consequently, assist in avoiding the high cost of replicating data (e.g. data about a ground-breaking scientific or medical experiment) that will be expensive to reproduce in the foreseeable and the distant future. In other words, data being produced today will need to be accessible to the future scientists or researchers in order to provide aid in their experiments and research. In addition, this preserved information may well be used to facilitate cross discipline research as well as checking if a particular research or study has been repeated elsewhere; naturally a very useful activity for funding bodies, scientists etc.

Evidently, the efficient use of these data in future will be achievable, only if their quality or integrity remains intact over time. Yet the changes in technologies and increased flexibility in their use result in transforming and putting the integrity of very data they create at jeopardy. The main problem, herein, lies with the fact that the capacity to manage the long-term stewardship of digital information has been relatively slower to develop than that to create and consume it. The problem is exacerbated by the increasingly ephemeral time horizon beyond which preservation of digital resources becomes an exigency, a grim consequence of the frailty of digital storage media, with significant contribution from rapid obsolescence of storage and rendering environments (Lavoie, 2004). Therefore, with rapid evolution and enhancements in related technologies and data formats, this task of ensuring data quality for long periods of time, i.e. successful long-term (where long-term may imply long-term enough be concerned about the obsolescence of technology, or it may mean centuries or decades) data preservation, may seem incredibly daunting.

Under the challenges set by the daunting task of successful long-term data preservation, the word 'Metadata' is becoming increasingly prevalent, with a growing awareness of the role that it can play in accomplishing such a task. In fact, the digital preservation community has already envisaged the need of good quality and well-managed metadata for reducing the likelihood of the digital object becoming un-useable over substantially long periods of time. Metadata's assistance in reconstruction or accessibility of preserved data, however, bears the same predicament as that of the efficient use of digital information over time: long-term metadata quality and integrity assurance notwithstanding the rapid evolvements of metadata formats and related technology. The only solution to this problem is employment of a well-conceived, efficient as well as scalable curation (Appendix A) plan or strategy for metadata over substantially long periods of time. In effect, curation has the ability to inhibit metadata from becoming out of step with the original data or undergoing additions, deletions or transformations which change the meaning without being valid. In other words, in order to ascertain the overall quality and integrity of metadata over a sustained period of time, thereby assisting in successful long-term digital preservation, effective long-term metadata management or curation is indispensable.

Over the past few years, several organized and arguably successful endeavors (e.g. The NEDLIB[1] project) have been made in order to find an effective solution for successful long-term data preservation. However, the territory of long-term metadata management, thus far, is even somewhat unexplored, let alone conquered. In other words, no acceptable methods exist to date for effective management and preservation of metadata for long periods of time. This instalment aims to explore the primary ground for tackling this highly complex and significant but as yet unresolved issue of long-term Metadata curation.

Therefore, realizing the high significance and pervasiveness of metadata management and its quality assurance for the purpose of successful long-term data preservation, the main objectives of this MSc. project were to perform an in-depth research on different recognized metadata standards, thorough analysis of current approaches and methods to manage metadata and its quality assurance; and review of existing metadata management tools. All findings were to be summarized and assessed in terms of the relevance for data curation (See Appendix A). The results of this extensive and meticulous investigation were to contribute to

[1] Networked European Deposite Library - http://www.kb.nl/coop/nedlib/

devising a detailed plan of work for the coming 2.5 years for developing a working prototype of metadata management system in the context of digital curation.

1.3 Structure of this Dissertation

This dissertation has the following structure. Chapter 1, i.e. this chapter, introduces the main project objectives, project scope etc. as well as specifying the core requirements of this project. In addition, this chapter provides detailed description of the plan that was followed for this project. Chapter 2 presents thorough discussion of relevant concepts and principles that served as the platform or basis for performing different project tasks. The following chapters, from 3 to 7, summarise results of main project tasks in the order that they were performed (see section 1.5). Finally, Chapter 8 provides a conclusion of this project, briefly evaluating the main achievements against the project goals and milestones while outlining the major difficulties and problems encountered during the course of this project.

1.4 Project Specification

Every project has its own specific requirements. For example, a project aiming to develop particular software would specify different functional and non-functional requirements based on how both the users and developers of the software would perceive its use. However, as for this MSc. research project, the main objectives of the project more or less sum up the main requirements.

As mentioned above, the main objective of this project was to perform an in-depth research on published works, tools, standards etc. for long-term metadata management in order to assess them in terms of relevance for successful preservation of high-quality data. At the final stage of this project, these research results were to be employed in devising a detailed work plan for the development of a working prototype of metadata management system. This working prototype would serve the main purpose of the project – *long-term metadata management and quality assurance in the context of digital curation*.

Nevertheless, it may be helpful to outline the key problems that were to be investigated throughout the 5 months of the project period.

- **Recognised Metadata Standards:** In depth research needed to be performed to assess the effectiveness and suitability of currently available metadata standards for long-term successful data preservation. The result of this assessment may well serve as the basis for developing a specific metadata standard for long-term metadata management.

- **Current Approaches to Metadata Management:** Thorough investigation of currently employed approaches in terms of published works, research efforts etc. for Metadata management were required. These approaches needed to be assessed in terms of relevance for data curation.

- **Metadata Quality Assurance & Versioning Techniques:** Existing techniques (e.g. *Metadata validation* against a set standard) for ensuring Metadata quality needed to be examined on the basis of different criteria, such as efficiency, intelligence etc. In addition, techniques and standards for Metadata versioning (see 2.4.4) and proper management and updates of Metadata versions for metadata management needed be investigated in details.

- **Existing Metadata Management Tools:** As this research was to contribute towards the development of a working prototype of Metadata management system, a significant part of this literature survey was required to focus on the assessment of the existing Metadata management tools based on their efficiency, drawbacks as well as degree of success in terms of relevance for long-term preservation and management of Metadata.

- **A List of Potential Collaborators:** Research needed to be carried out to locate and assemble a list of experts, research groups etc. who are working to achieve the similar objectives to those of this project and most likely to act as potential collaborators for this project.

- **A Plan of Work for the coming 2.5 years:** The final requirement for this project was to devise a plan of work for the coming 2.5 years to come to a working prototype of a Metadata management system, based on the outcomes of this extensive literature survey.

1.5 Project Management

No project work is a casual activity; it must be carefully planned, and this project was no exception. The project was undertaken within a six month time frame, first of which was spent on the preliminary research work in order to gain a thorough and explicit understanding of the problem domain. A detailed project plan was also formulated on the basis of the results of that primary research. Both the project plan and the preliminary research results were submitted in the form of Project Placement Preliminary Report to the project supervisor for approval. This section aims to provide that work plan in details; which was followed for this MSc. research project on long-term metadata management after the submission of the preliminary report.

1.5.1 Project Tasks

As approved by the project supervisor, followings are the most significant tasks (listed in the order they were performed) that were carried out to achieve the main project objectives:

Task 1 - Assessment of Recognised Metadata Standards

As it has been mentioned in section 1.3, the project required in-depth research on different recognised metadata standards in order to assess their relevance and degree of efficiency for long-term data preservation. Besides, profound knowledge about different Metadata standards was deemed helpful for the succeeding tasks in assessing their use in different Metadata management and quality assurance techniques. Therefore, assessment of the recognised metadata standards was decided to be the first step for this project. This task required approximately 2 weeks to complete.

Task 2 - Assessment of Current Approaches for Metadata Management

The next task was to dedicate significant amount of research efforts towards gathering detailed information about the current approaches for metadata management. This involved gathering all relevant published works (e.g. research efforts etc.) on metadata management techniques. In addition, the quality assurance techniques employed by these approaches were thoroughly examined. Finally, major efforts were dedicated towards assessing these approaches or techniques in terms of relevance for data curation. This task completed within approximately 3 weeks.

Task 3 - Assessment of Current Quality Assurance and Version management techniques

Aside form the quality assurance techniques employed by current approaches for metadata management, it was necessary to acquire information regarding any other general metadata quality assurance and metadata validation techniques and how these techniques ensure high quality data over long time. Besides, this research also included finding currently employed techniques for metadata validation against a set Metadata standard. In addition, investigation was performed in details in order to determine how these quality assurance techniques could be incorporated within the actual metadata management process.

Furthermore, exhaustive research was conducted in order to collect information about metadata versioning techniques, how these versions are controlled, managed etc. As part of the main project requirements, these quality assurance and version techniques were assessed on the basis of their potential contributions towards long-term preservation of high quality data. This task required about 3 weeks for its completion.

Task 4 - Production and Delivery of Interim Progress Report

Interim report was produced, detailing all findings and assessment results up till task 3 and delivered to the project supervisor for examination. Writing of the report required approximately 2 weeks. It should be noted, due to insufficient time (as indicated in the Gantt chart in Appendix H) before the deadline for this report, it was deliberately planned to perform task 4 in parallel to task 3.

Task 5 - Assessment of Existing Metadata Management Tools

The next step for this project was to examine all existing metadata management tools in terms of relevance for data curation (Appendix A). This examination included studying their metadata management and quality assurance techniques as well as assessing their degree of efficiency for long-term preservation for high quality data. In addition, other features of these tools such as, industry standards, user friendliness, robustness, customer range (i.e. how widely used and by whom) etc. were also examined in great details.

Task 6 - Assembling a List of Potential Collaborators

At this point of the project, research needed to be undertaken in order to locate all experts, research groups or organisations etc. that are working to achieve or have achieved similar objectives to that of this project and are most likely to act as potential collaborators for this project. After that, a list was to be produced to detail the contact information for these potential collaborators. This task required approximately 2 weeks to be completed.

Task 7 - Devising a Plan of Work for the Development of a Working Prototype

At this stage, most of the metadata management and quality assurance related concepts, techniques, tools etc. had been obtained. Therefore, based on the research results until this stage, detailed work plan for the coming 2.5 years was devised in order to develop a working prototype of Metadata management system. This plan of work subsumed relevant and potentially useful recommendations and comments (compiled in light of the experience and knowledge gained from this literature survey) in order to indicate the direction in which one should proceed to develop such system. This required approximately 2 weeks.

Task 8 - Production and Delivery of Dissertation Report & Preparation for Presentation

The final task for this project was to write this final report or dissertation thesis, presenting all findings and achievements for the entire project period. Considering the substantial amount of writing (e.g. approximately 25,000 words) to be done, this task was given approximately 3 weeks for its completion. Subsequently, preparation will need to be done for project colloquia or oral presentation. This is expected to take approximately 1 week.

1.5.2 Project Summary

The table below summarises all completed project tasks:

Project Step	Task	Duration (Weeks)
1	Assessment of Recognised Metadata Standards	2
2	Assessment of current Approaches for Metadata Management	3
3	Assessment of current Quality Assurance and Version management techniques	2
4	Production and Delivery of Interim Progress Report	2
5	Assessment of Existing Metadata Management Tools	3
6	Assembling a list of Potential Collaborators	2
7	Devising a Work plan for the Development of a Working Prototype	2
8	Production and Delivery of Dissertation Report & Preparation of Presentation	3+1

Table 1.1: Project Summary

1.5.3 Milestones for the Project

A Gantt chart, detailing completion time for all aforementioned project tasks, has been given in Appendix H.

Chapter 2

Review of Main Concepts & Issues

In order to provide an insight into the problem domain, this chapter presents a concise overview of different metadata management related issues, important definitions etc. However, the information provided in this chapter are only to serve as a foundation for the assessment and synthesis conducted on different aspects related to metadata management over long periods of time, therefore, should not be misconstrued as part of the main outcomes of the project.

2.1 Metadata Defined

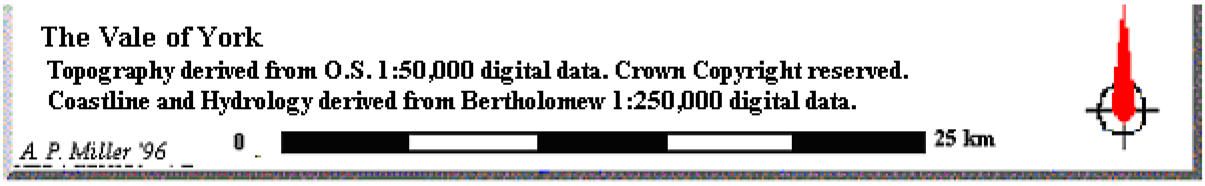
"Metadata is data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics"
- Lorcan Dempsey and Rachel Heery, UKOLN[2]

The word "metadata" was invented on the model of *meta-philosophy* (the philosophy of philosophy), *meta-language* (a language to talk about language) etc. in which the prefix *meta* expresses reflexive application of a concept to itself [JMS97]. Therefore, at the most basic, metadata can be considered as data about data. However, as conformity of the middle term ("about") of this definition is crucial to a common understanding of metadata, this classical and simple definition of metadata has become ubiquitous and is understood in different ways by many different professional communities [GEG04]. For example, from the bibliographic control outlook, the focus of "aboutness" is on the classification of the source data for identifying the location of information objects and facilitating the collocation of subject content. Conversely, from the computer science oriented data management perspective, "aboutness" may well accentuate on the enhancement of use in relation to the source data [PMD96]. Moreover, this metadata or "aboutness" is synonymous with its context in the sense of contextual information.

Nevertheless, in light of its acknowledged role in the organisation of and access to networked information and significance in long-term digital preservation, metadata may be defined as structured, standardized information that is crafted specifically to describe another digital resource, in order to aid in the intelligent, efficient and enhanced discovery, retrieval,

[2] The UK Office for Library and Information Networking, University of Bath, UK.

use or preservation of that resource over time. For example, a paper map from the Ordnance Survey of Great Britain[3] as shown below associates metadata such as its scale, the date of survey and date of publication etc. However, a simpler example of metadata may be a service with three elements: *creator, function and availability* [MMC98].



The Vale of York

Topography derived from O.S. 1:50,000 digital data. Crown Copyright reserved.

Coastline and Hydrology derived from Bertholomew 1:250,000 digital data.

A P. Miller '96

0

25 km

Figure 2.1: An example of metadata [PMD96]

In less traditional information domain, the term *metadata* acquires an even broader scope. For example, an Internet resource provider might use metadata to refer to information being encoded into HTML[4] meta-tags for the purposes of making a Web site easier to find.

2.2 Categories of Metadata

Metadata has the proven ability to describe different aspects of a digital (and/or physical) information object, such as accessibility, preservation etc. in distinct, efficient and unambiguous manner. Owing to its versatile capacity, metadata is being increasingly prevalent in long-term digital curation environments. The versatility of metadata has also yielded different perspectives on it, which in turn has lead to a broad conception of metadata. This broad conception has further evolved, chiefly to aid in its proper understanding, into a range of distinct categories of metadata as outlined below:

2.2.1 Administrative Metadata

This type of metadata provides information documenting the life-cycle of a digital object and may include information (typically external to informational content of the digital object) regarding acquisition information, version control, archiving policy, audit trail, rights management, provenance, ownership and reproduction tracking etc. of the digital resource. In long-term digital curation, this metadata aids in efficient management and administration of the digital object throughout the curation period. Examples of this type of metadata may be data captured by the "Rights" element of the Dublin Core metadata standard and the "Access_Constraint" element of the Directory Interchange Format (DIF)[5] metadata standard. Administrative metadata may also include physical characteristics of digital objects. Examples include hardware and software documentation, digitisation information of a digital resource. This type of administrative metadata is often referred to as "Technical" metadata. Detailed, format-specific technical metadata is clearly necessary for implementing most

[3] Ordnance Survey - Britain's national mapping agency for best of British maps and mapping data - http://www.ordsvy.gov.uk/

[4] A coded format or page-description language used to create files or documents that can be formatted and displayed by World Wide Web browsers.

[5] Directory Interchange Format (DIF)
Writer's Guide, Version 9.4 - http://gcmd.gsfc.nasa.gov/User/difguide/difman.html

preservation strategies. The “software” and “hardware” elements as defined in PREMIS[6] data dictionary contain technical metadata about digital objects.

2.2.2 Descriptive Metadata

This category of metadata captures variety of descriptive information regarding a digital resource, such as identifying information, intellectual entities, annotation details, keywords etc. This descriptive information effectively assists in the process of retrieving the digital resource by enabling users to initially discover its existence, to locate it and then to determine if it is the resource that they require. It is also common practice to use this metadata to provide unique identification and links to organizations, files, or databases which have more extensive descriptive information about the resource (this is of particularly importance in the event that the digital file and its metadata are managed separately). Descriptive metadata can also help decision makers during preservation planning. For example, information represented by “Title”, “Identifier”, “Creator” etc. elements of Dublin Core metadata format may be classed as descriptive metadata.

2.2.3 Structural Metadata

In essence, structural metadata presents information in regards to the structure (i.e. internal organisation) of a digital object on various levels of complexity. It may also include information on relationships among different components or sections of a complex digital object for the purposes of navigation. This metadata is used primarily for storage of objects in a repository and for presentation of that object. Examples of this type are the table of contents, page numbers, and index of a journal or the types of reports (laboratory, imaging, consultant) etc. The “structure” element of DCC RI-label schema [7]has been defined to present structural metadata about digital objects.

Also of note, the borders of these three types of metadata however are not necessarily distinct from each other or exclusive of other types of metadata. Descriptive metadata, for example, is often referred to as a sub-class of administrative metadata. Preservation Metadata is also a specialised form of administrative metadata that can be used as a means of storing the technical information that supports the preservation of digital objects in changing technological environment and may also be considered Structural Metadata by definition[8]. Examples of this type of Metadata may be documentation of physical condition of resources, data refreshing and migration etc. Preservation Metadata is the most relevant metadata category to the long-term preservation of digital objects and is covered in more depth in “Preservation Metadata” chapter of this instalment.

[6] PREMIS (PREservation Metadata: Implementation Strategies) Working Group - http://www.oclc.org/research/projects/pmwg/

[7] http://dev.dcc.ac.uk/dcc-rilabel.xsd

[8] PADI-Metadata, National Library of Australia, Last accessed 06 August 2005 - http://www.nla.gov.au/padi/topics/30.html

2.3 Importance of Metadata

As it has been mentioned before, the importance of using metadata is diverse. This subsection elaborates on various benefits of using metadata [MMC98, JRB00], described in the previous section.

2.3.1 Understanding & Increased Accessibility

Metadata provides meaning to computer readable information. Without human language descriptions of their various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end users. The metadata provides the bridges between the producers of data and their users and convey information that is intrinsic for secondary analysts.

As far as access to preserved resources is concerned, rich and consistent metadata facilitates high precision discovery of those resources. Metadata can also make it possible to search across multiple collections or to create virtual collections from materials that are distributed across several repositories, but only if the descriptive metadata are the same or can be mapped across each site. In addition, Metadata is a key factor for ensuring the long-term access of digital resources. There is a continuous need for extending the existing metadata element set to be able to describe all available digital resources. Besides all these, Metadata can be used to describe the mapping between file instances and particular replica[9] locations, hence, plays a vital role in ensuring effectiveness of a Resource Location service[10].

Furthermore, well-structured metadata can facilitate an almost infinite number of ways to search for information, present results, and even manipulate information objects without compromising the integrity of those information objects. Also, metadata can be used to proper reuse of data objects.

2.3.2 Retention of Context & Assessing

Metadata plays a crucial role in documenting and maintaining different relationships between information objects as well as in indicating their authenticity, structural and procedural integrity and degree of completeness. In an archive, for example, metadata that documents the content, context, and structure of an archival record, consequently helps to distinguish that record from de-contextualized information.

In addition, Metadata provides end-user an opportunity to assess the quality and relevance of a collection of numbers. By describing methodologies and procedures, as well as features related to the context of a particular study, end users are allowed to decide whether or not a data collection is meeting their professional or scientific standards.

[9] Microsoft Definition: A copy of a public folder (i.e. data) that contains all of the folder's (i.e. data) contents, permissions, and design elements, such as forms behavior and views. Replication of data can reduce access latency, improve data locality, and increase robustness, scalability and performance for distributed applications.

[10] Used to maintain and provide access to mapping information from logical names for data items to target names, i.e. ensures accessibility of data that are near in terms of network latency.

2.3.3 Multi-Versioning & Preservation

Metadata can be used to provide link between the multiple versions and variants of data objects and capture the similarities and differences between each version. The metadata may also be used to distinguish what is qualitatively different between variant digitised versions and the hard copy of the original or parent object

As it has been mentioned before, Metadata also plays a very significant role in long-term data preservation. In general, metadata enables digital information objects to exist independently of the system that is currently being used to store and retrieve them, consequently enabling them to survive migrations through successive generations of computer hardware and software, or removal to entirely new delivery systems. However, it should also be noted that for the information objects to remain accessible and intelligible over time, it would also be essential to preserve and migrate this metadata.

Furthermore, Metadata holds the promise of being able to improve both the storage system-level performance, through more efficient staging (besides migration), and application-level performance, by allowing the user to make more informed choices about what data to retrieve.

2.4 Long-term Metadata Management: Main Requirements

The efficacy of Metadata management largely relies upon successful implementation of a number of requirements. Although metadata management requirements may be quite different according to the type of data described, the information outlined below attempts to provide a general overview of the main requirements; indeed the requirements of Metadata management or curation is an open research area.

2.4.1 Metadata Standard

Digital Preservation professionals have already perceived the necessity of a metadata standard[11] in forestalling obsolescence of metadata (hence obsolescence of the actual data or resource), due to dynamic technological changes. In the context of long-term data curation, it is essential that the structure, semantics and syntax of Metadata conform to a widely supported standard(s), so that it is effective for the widest possible constituency, maximises its longevity and facilitating automated processing.

As it would be impractical to even attempt to determine unequivocally what will be essential in order to curate metadata in the future, the metadata elements should reflect (along with other relevant information such as, metadata creator, creation date, version etc.) necessary assumptions about the future requirements in that regard. Furthermore, the metadata elements should be interchangeable with the elements of other approved recognised standards across other systems with minimal manipulation in order to ensure metadata interoperability[12]. This will consequently aid in minimization of overall metadata creation

[11] Fundamentally, a metadata standard or specification is a set of specified metadata elements or attributes (mandatory or optional) based on rules and guidance provided by the governing body or organisation(s).

[12] Metadata interoperability implies the possibility to unambiguously interchange between metadata schemes (in the textual and binary representation formats) such that components share similar meanings for all compliant parsers/decoders (Oltmans, 2001).

and maintenance cost. It may also be advantageous, as Rothenberg recommended, to define specific metadata e[13]lements that portray metadata quality.

2.4.2 Long-term Preservation

As mentioned before, long-term metadata curation is prerequisite of ensuring the successful long-term preservation of data. Therefore, metadata curation requires metadata to be preserved along with data in order to ensure its proper and effective descriptions over time.

To date, the dominant approach to long-term data preservation has been that of migration[14]. Unfortunately, it does pose the notable danger of data loss or in some cases the loss of original appearance and structure (i.e. 'look and feel') of data as well as being highly labour intensive. However, in the context of metadata preservation, 'look and feel' of metadata is not as imperative (e.g. using differing date/time formats) as that of the original data as long as it maintains its aptness for describing the original data accurately over time. Therefore, albeit the existence and availability of Emulation (which seeks to solve the problem of data 'look and feel' loss by mimicking the hardware/software of the original data analysis environment) Migration would appear to be a better solution for long-term Metadata preservation. If a superior or alternative preservation strategy is proposed this would be worth considering also as both Emulation and Migration have received criticism for being costly, highly technical, and labour intensive.

However, a classic unresolved data migration issue is that of tracking or migrating changes to the metadata itself. This issue is likely to arise when significant changes occur in the future to currently used metadata standards/formats. For example, an element contained within a contemporary metadata format might be replaced or even excluded in newer versions of that format, thus incurring the problem of migrating information under that element to corresponding element(s) (if any) of the new format. In order to successfully curate Metadata, a curation aware migration strategy needs to facilitate migration (ideally from old formats to new formats) and tracking/check changes (i.e. new formats to old formats) of metadata between metadata formats (e.g. maintaining an audit trail across versions) but also be flexible for addition of further requirements.

Furthermore, the total costs of preservation of digital data have not yet been determined. Most of the research to date [CPA96] has determined that preservation of digital data will be expensive, primarily because data preservation is a manual process and is very labor intensive. The cost of maintaining and preserving metadata also contributes to the total cost of long-term data preservation. For example, checks of the integrity of the data and metadata and checking for errors are additionally time-consuming tasks nevertheless are necessary.

2.4.3 Quality Assurance

As highlighted earlier in this paper, Quality assurance of metadata is an integral part of long-term metadata curation. It needs to be ensured that appropriate quality assurance

[13] Rothenberg, J. (1996): *Metadata to Support Data Quality and Longevity*, RAND, 1996 – http://www.computer.org/conferences/meta96/rothenberg_paper/ieee.data-quality.html

[14] The process of translating or transforming digital data from format or platform that is under the threat of becoming obsolete to a current format or platform.

procedures or mechanisms are in place to eliminate any quality flaws in a metadata record and thereby, ascertain its suitability for its intended purpose(s). As identified in [JSC03] some of the quality flaws that usually occur in metadata are as follows:

- Incorrect Content: The content of the metadata may be incorrect or out-of-date. This is mainly due to lack of validation and sanity checking at the time of metadata entry.

- Inconsistent Content: A lack of cataloguing rules and inconsistent approaches often cause inconsistency in the metadata content, especially, in cases where multiple people are involved in creating metadata.

- Non-interoperable Content: Lack of interoperability among different metadata formats across different systems generates non-interoperable metadata records. For example the date 01/12/2003 could be interpreted as 1 December or 12 January if projects based in the UK and USA make assumptions about the date format.

- Errors with Metadata Management Tools: Due to flaws in related functionality, metadata creation and management tools often output metadata in invalid formats.

- Errors with the Workflow Process: Metadata may become erroneous and corrupted through the workflow of different processing tools. As a simple example a MS Windows character such as © could be entered into a database and then output as an invalid character in a XML file.

In general, any Metadata quality assurance procedure should take three metadata quality levels into considerations: Semantic Structure (“format” or “element set”), Syntactic Structure (administrative wrapper or “schema”) and Data Values/Content. Procedures may also be required for periodic checking of the Metadata.

It should also be ensured that metadata creation and management tools have a rich set of functionality for metadata validation. In essence, the validation process ensures that metadata exhibits consistency across all records and conforms to some agreed standards [JCD01]. It should be noted that Metadata validation usually is done by checking metadata syntax such as, spelling etc. However, for metadata, perfect syntax does not guarantee a meaningful description of a data set. Therefore, validation for metadata semantics is also a strong requirement.

2.4.4 Versioning

Throughout the vibrant process of long-term metadata curation, metadata is prone to be volatile. This volatility may well be caused by updating of metadata which can involve the amendment or deletion of the metadata records, or the addition of new metadata. However, previous versions of metadata may need to be retrieved (e.g. in the case of annotation - who made the annotation and which version(s) of a value does it apply to) in order obtain vital information about the associated preserved information if required. It is therefore essential to be able to discriminate between metadata in different states which arise and co-exist over time by versioning[15] metadata information.

[15] Logically, a version is a complete snapshot of the state of an object at a particular point of time.

2.4.5 Metadata Storage Location

Different locations of metadata may be generalised as two locations, archiving systems and data warehouses (Appendix D). In general, archives are employed to retain information for its long-term value, or if one is an optimist, permanent value. Therefore, from the perspective of long-term data preservation, archiving system may be the more relevant of the aforementioned two possible metadata locations.

There are two main possibilities regarding the location of metadata in an archiving system; Metadata can be stored within the resource it describes or separate. Managing the metadata separately, for example in a database, normally makes the process of resource discovery more efficient. Yet, for the sake of integrity, all-important information should be tightly coupled with the resource. Keeping metadata close to the document itself is beneficial for the management of the system, as both will mutually persist in the archive [CLJ00]. However, the overall maintenance procedure is relatively more difficult for this possibility.

Furthermore, metadata about different versions of data, such as label associations, audit records or branching information need to be stored separately from the data itself in order to facilitate efficient operations in many cases, such as selecting a version or set of versions by label.

2.4.6 Other Issues

Aside from the requirements outlined above, long-term metadata curation need take the following additional issues into account.

- <u>Metadata Policy:</u> A set of broad, high-level principles that form the guiding framework within which the Metadata curation can operate, must be defined (Lowe, 2002). The Metadata Policy would normally be a subsidiary policy of the organizational data policy statement, and as such should reference the same, e.g. legal issues regarding the use of data (or metadata) etc.

- <u>Access Constraints & Control:</u> There should be one authoritative source and registration process for each type of metadata to control unauthorised access to it. This effectively helps prevent any illegal or malicious modification to the metadata, hence ensures the overall consistency in the metadata records. In addition, this should help support audit control or trail facility for metadata. This audit trail information should be tied closely with the metadata archive, metadata standard and metadata policy, which needs to provide information about what standard, what policy, and where in the lifecycle the metadata was created or edited.

Chapter 3

Assessment of Recognised Metadata Standards

Metadata is considered to be good in quality when it conforms to a set of standard(s). In the context of long-term data curation (Appendix A), it is essential that the structure, semantics and syntax of Metadata conform to widely supported standard(s), so that it is effective for the widest possible constituency, maximises its longevity and processing can be automated as far as possible. In essence, a metadata standard or specification is a set of specified metadata elements or attributes based on rules and guidance provided by the governing body or organisation(s). In simple terms, metadata standards provide guidelines (agreed and accepted by communities concerned) to create and maintain metadata.

Due to a wide range of communities having an interest in metadata, there are a bewildering number of metadata standards and sub-sets or even super-sets of standards in existence or under development. In addition, the fact that Metadata Standards have long been a national and international priority for professionals in government, information management as well as archiving and library communities; has also resulted in a large number of efforts for standardizing metadata. Many of these have much commonality, but vary in the degree of complexity and the level of detail required to complete a Metadata entry [CHN01]. The project required examination of a significant number of recognised metadata standards in details and assessing them in terms of their suitability and usefulness for long-term metadata management and quality assurance. This chapter outlines the assessment results of a select few of those metadata standards.

3.1 Categories of Metadata Standards

In order to provide better understanding of this large number of metadata standards, several attempts have been made to divide these standards into distinct categories, based on different criteria, such as characteristics, origin etc. One such attempt is the one made by Lorcan Dempsey and Rachel Heery of UKOLN [LDR97], who describe three different categories of metadata standards on the basis of their completeness and structure as shown in table 3.1.

Category	Structure	Example
I	Simple Formats, Proprietary, Full text indexing	Lycos, AltaVisa, Yahoo etc.
II	Structured formats, Emerging standards, Field structure	Dublin Core, RFC, SOIF, LDIF, etc.
III	Rich formats, International standards, Elaborate tagging	CIMI, GILS, EAD, SOIF, TEI, MARC, LOM etc.

Table 3.1: Three Categories of Metadata Standard

In the table above, Category I generally includes unstructured indexes - the data currently created by web crawlers. These can be reasonably effective for finding a known item but less effective for discovery. Category II includes data, which contains a full enough description to allow a user to assess the usefulness or interest of a resource without having to retrieve it or connect to it. Finally, category III includes fuller descriptive formats, which may be used for location and discovery but also have a role in documenting objects. It is suggested that the trend is for category II to become more important as a general-purpose access route [LDR97].

Aside from these three categories, metadata standards also differ in whether they specify content, format, or use:

Standards	Description	Example
Content standards	Specifies information content, but not how to organize this information in a computer system or for a data transfer, or how to communicate or present the information.	ISO, FGDC, GILS, CLRC, DDI, Dublin Core etc.
Format standards	Specifies information for indexing, cataloguing etc.	MOF, CWM, MARC etc.
Use standards	Represents both data and metadata	XML, XMI, DTD, UML etc.

Table 3.2: Another categorization attempt for Metadata Standards

It should be noted that these categories of metadata standards are to provide only a clearer image of the metadata standards, not to outline the state of the art.

3.2 Dublin Core Metadata Standard

The Dublin Core (DC) metadata standard, probably the most well-known of all metadata standards, is a list of fifteen metadata elements that specifically intended to support resource discovery by enhancing existing network catalogues of electronic documents. It aims to provide a basis for semantic interoperability between other, probably more complicated, formats and resource discovery tools. In the home-page's own words: it specifies "a simple resource description record that has the potential to provide a foundation for electronic bibliographic description that may improve structured access to information on the Internet and promote interoperability among disparate description models".

The semantics of DC elements (approved as ANSI/NISO Standard X39.85 in 2001 and as ISO standard 15386 in early 2003) have been established through consensus by an international, cross-disciplinary group of professionals from librarianship, computer science, etc.; which can be used to describe a wide variety of electronic information resources for the purpose of simple cross-disciplinary resource discovery [DCM04]. It is to be noted that each Dublin Core element is defined using a set of ten attributes from the ISO/IEC 11179 standard for the description of data elements. These 15 DC elements have been listed in Appendix B.

These 15 elements of DC and the relationships defined between the resource-of-interest and other resources define the basic DC data model [KGJ98]. Aside from these elements, a completely abstracted DC data model is also required to include its two types of qualifiers: Encoding Scheme[16] and Element Refinement[17]. The data model is depicted below:

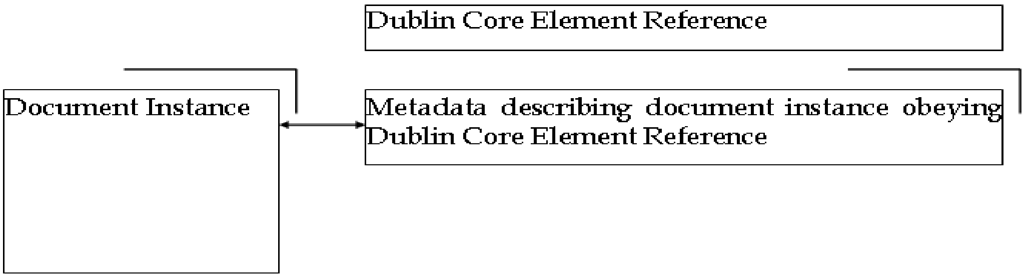

Dublin Core Element Reference
Document Instance
Metadata describing document instance obeying Dublin Core Element Reference

Figure 3.1: Data model for Dublin Core elements [KGJ98]

It should also be noted that these qualifiers were issued in July 2000 as a list of recommended Dublin Core Qualifiers. In addition, DC metadata initiative associates a limited set of consistently used and carefully defined terms, known as "controlled vocabulary", which are used to select content data for some elements of DC.

3.2.1 Dublin Core Assessed

This research involved studying the DC elements, its qualifiers and vocabularies and assessing them on the basis of their significance and potential effectiveness for long-term metadata management. The results of that assessment are as follows:

a) Element Set

The Dublin Core element set is intended to be as simple as possible to provide easy and inexpensive creation of metadata records to describe resources. In addition, the element set provides for effective retrieval of those resources in the networked environment. This simplicity of DC element set should enable relatively easy maintenance of the metadata records. Besides, DC element set provides users with the flexibility to use fields that are specific to their needs, e.g. describing data that is not part of the element set. Moreover, the syntax independency of DC was aimed to provide interoperability for metadata records within a heterogeneous data environment (see 2.4).

Although, the simplicity and comparatively low number (LOM – 80 elements, CSDGM around 400 elements etc.) of the DC metadata elements may seem to be ideal for easy

[16] Store an identifier for the vocabulary, encoding or language of the value.

[17] Used to further refine the semantic meaning of an element.

management of metadata, the optional[18] and repeatable[19] nature [RSM01] (i.e. no restriction on the maximum or minimum number of elements), of the element sets may well lead to potentially incorrect description of the resource object.

Research has proven that this syntax-independence, optional, extensible and repeatable natures of DC element set have resulted in some major weaknesses. In the words of the authors of the Warwick Framework, "The authors of the Dublin Core readily admit that the definition is extremely loose. With no definition of syntax, and the principles that 'everything is optional, everything is extensible, everything is modifiable' the Dublin Core definition does not even approach the requirements of a standard for interoperability. ..." In addition, the simplicity of the DC elements also has a downside since the description cannot be as accurate as a more complex resource description [RCD97]. This drawback of DC elements makes them inadequately comprehensive for complex information domains such as, Scientific Domain, e.g. the Council for the Central Laboratory of the Research Councils (CCLRC).

Furthermore, various researches have detected inefficiency of DC standard in resource discovery and retrieval (one of the main objectives of the DC element set), caused mainly due to lack of guidance provided by DC for system designers and implementers of web crawlers and spiders that may use the Dublin Core as the source for resource discovery and indexing. Research by Lloyd Sokvitne, State Library of Tasmania, states that the DC standard will have questionable value as a discovery tool unless the elements can be populated and used correctly [LSMNA]. However, the ability of DC elements to be mapped to other metadata standards, such as GILS (see Section 3.5) etc. might solve the problem of over simplicity.

b) Controlled Vocabulary

Generally, controlled vocabularies are intended to preserve the metadata quality by reducing the likelihood of spelling errors when recording metadata. However, controlled vocabularies, provided by DC, require an administrative body to review, update and disseminate the vocabulary, and may prove to be expensive. For example, the US Library of Congress Subject Headings (LCSH) and the US National Library of Medicine Medical Subject Headings (MeSH) are formal vocabularies, indispensable for searching rigorously catalogued collections. However, both require significant support organizations. Another cost is having to train searchers and creators of metadata so that they know the appropriate information to be entered into a file. This will, consequently, add to the total cost of metadata management [DIH00].

c) Qualifiers

The DC qualifiers are intended to promote interoperability among applications that use element refinements and encoding schemes to increase the semantic precision of metadata. The use of qualifiers simplifies mapping the DC elements to other information systems, thus enables a more precise description of the resource [LDR97].

However, these qualifiers may also introduce complexity of processing data, and the difficulties for interoperability. For example, the Author element name does not distinguish the form of author (e.g. personal, corporate, meeting etc.). It would be possible to use qualifiers to make these more precise distinctions, but the Dublin Core documentation does

[18] All DC elements are optional.

[19] Every element may be repeated without any constraint.

not attempt to make comprehensive recommendations in that regard. The following example comprising the Dublin Core suggested qualifiers for the Author element, illustrates this shortcoming of the Dublin Core specification:

Author (scheme=USMARC) =100 1 Shaon, Arif $c Mr, $d 1982.

Furthermore, the specification is incomplete and preliminary, in that, even for data-items that clearly need to be tightly defined on a particular domain, very limited guidance is provided. In particular, the Scheme qualifier enables a domain-definition to be nominated, but the values that the qualifier can take appear to be as yet undefined [RCD97].

d) Metadata Versioning

The DC metadata standard does not reflect the relationships among the data-elements. The only apparent means of expressing relationships among different metadata is the Relation element. One of the most serious concerns that arises in this regard is the failure to reflect the existence of multiple (e.g. in different languages, and in different formats) or successive versions of metadata records, and multiple instances of objects (commonly referred to as replication or mirroring). In other words, DC elements are unable to capture syntactical changes, making it impossible to provide common metadata version control features (e.g. roll-backs), which consequently restrict the semantic understanding of versioning change to the comparison of metadata records [CJJ03].

e) Data Formats

It is vital that metadata standard encompasses all potential forms that metadata records may acquire, including vector-graphics, sound, video etc. It is not clear that DC does so [RCD97].

f) Addressing Object-Identity

Identifiers will be highly valuable means of both finding and referring to documents and other objects, especially in a scientific domain in the distant future. The core elements of DC do not provide clear guidance regarding the methods of expressing versions of an object as well as distinguishing between logical and physical document identifiers and mapping from logical to physical identifiers and vice versa.

g) Multiple Instances of Metadata

Multiple (or alternative) instances of metadata records may be generated and stored in a data preservation environment. The purpose of this is, perhaps, to have at least one instance of metadata available with its quality intact, even though, other instances may have been altered or modified or even corrupted, hence to ensure the metadata quality as well as the data quality. The DC metadata fails to address this issue.

h) Metadata Storage

The DC specification does not address storage of metadata. However, Dublin Core metadata is often stored as name-value pairs within META tags, which are placed within the HEAD elements of an HTML document. It can also be located in an external document or loaded into a database enabling it to be indexed and manipulated from within a propriety application.

In summary, notwithstanding being simple, extensible, modifiable, and syntax-independent hence, widely adopted especially for cross-domain resource discovery, with several drawbacks, especially in regard to metadata versioning, the Dublin Core elements alone may not be sufficiently comprehensive for addressing the complex issues of long-term metadata management and quality assurance.

3.3 Content Standards for Digital Geospatial Metadata

The Content Standards for Digital Geospatial Metadata (CSDGM) was developed in June 1992 by the Federal Geographic Data Committee (FGDC) in response to the growing need to establish a broadly based and widely accepted content procedure definitions for the documentation of geospatial[20] data sets. The standard was developed from the perspective of defining the information required by a prospective user to determine the availability, fitness for intended use and means of accessing and successfully transferring the geospatial data. Although the standard was approved in June 1994 as CSDGM, it is commonly referred to as FGDC. Since its emergence, the CSDGM standard has been adopted by many public and private organizations [CSD02].

With about 334 different elements in the standard, latest approved version[21] (1998) of the CSDGM standard defines seven major categories of metadata (Figure 3.2). Within each of these are subcategories and layers of increasingly complex detail. While the full CSDGM metadata standard is too extensive to list and analyse, the minimum set, known as Metadata-lite, has been given in Appendix B. The following figure is a sketch of the structure of a document containing Metadata that conforms to the CSDGM standard [HSS97].

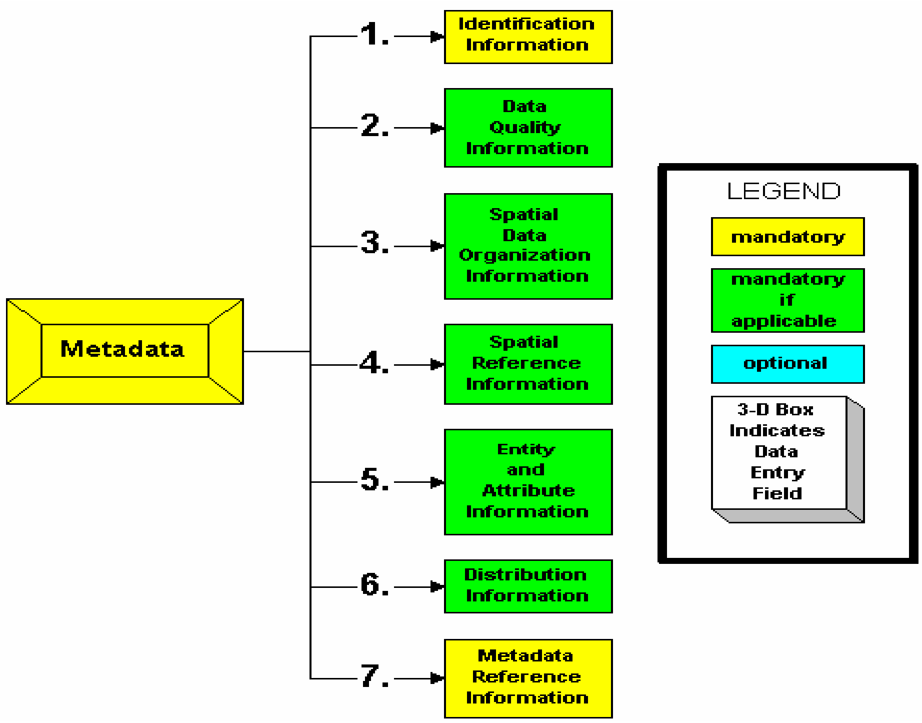

Metadata
1. Identification Information
2. Data Quality Information
3. Spatial Data Organization Information
4. Spatial Reference Information
5. Entity and Attribute Information
6. Distribution Information
7. Metadata Reference Information
LEGEND
mandatory
mandatory if applicable
optional
3-D Box Indicates Data Entry Field

Figure 3.2: Content Standards for Digital Geo-spatial Metadata (CSDGM) [JMS97]

[20] A term used to describe a class of data that has a geographic or spatial nature.

[21] Another version of CSDGM was to be released in 2001/20002, although no information was found in that regard.

3.3.1 CSDGM Standard Assessed

To begin with, the standard provides a very detailed content description for digital geospatial data sets with various specialised descriptive elements e.g. percentage cloud cover [LDR97]. However, as it has been mentioned before, the CSDGM standard is a complex format with over 300 data elements, 119 of which exist only to contain other elements. Some CSDGM metadata elements are mandatory; many more are "mandatory if applicable" or optional. Some fields have specifically enumerated values or require index terms to be drawn from an explicit thesaurus in order to improve machine readability and search-ability of these records [HSS97].

However, the structural overhead of the standard is the main reason for its complexity, resulting in lack of understanding of the CSDGM compliant metadata without firm knowledge of the standard. Moreover, due to the complexity of the standard, attention to format may easily overtake attention to actual content, resulting in metadata that are inefficient in describing the resource appropriately.

Although, the full standard allows for the maintenance of higher quality metadata by those data centres that desire it, special tool is required to assist with the creation of CSDGM compliant metadata. This may imply added time for editing existing metadata, which may result in higher cost for the overall management of metadata. On the contrary, the use of these tools for creating and editing metadata may prevent inadvertent error while performing those operations, hence ensuring the quality of metadata. The CSDGM home[22] page has a selection of such tools, developed by different agencies, available [LDR97].

The full set of CSDGM elements, while quite comprehensive, is far too onerous to adhere to in a setting of limited budgets. Especially, the length and number of highly scientific terms within the specification makes them rather daunting to implement for researchers whose area of expertise is outside the CSDGM information domain. Keeping this in view, the minimum set of CSDGM elements, i.e. Metadata-lite calls for an acceptable quantity of metadata enabling many data centres to participate [HSS97]. However, there are no provisions made within the standard for the description/use of other languages. This may be seen as a drawback when managing multi-lingual metadata.

Moreover, the CSDGM does not allow metadata versioning in direct terms. Metadata version management is addressed purely by the CSDGM mandatory element *Metadata_Date*. It is to be noted that the same field is used for both the metadata creation date and the date of any updates to the metadata. In other words, legacy metadata records are not maintained unless the legacy dataset itself is maintained. If the legacy dataset is maintained, a new metadata record is created for the updated version. This presents two versioning problems:

- If no naming convention is implemented, connectivity between the versions is lost.
- If the legacy data set is renamed to provide that connectivity, e.g., *'Bridges_old'* then the metadata record must be updated to reflect the new title and, technically, the *Metadata_Date* should be updated and the original date of the metadata is lost.

[22] :http://www.fgdc.gov/clearinghouse/mitre/task2/tools.html

Furthermore, the CSDGM does not make difference between date of birth and update date. While this may simplify the creation of date elements, at the same time it may cause confusion in regard to the purpose of the elements, i.e. whether this is a date of birth or an update date.

Quality assurance of the CSDGM metadata elements is limited to a validation tool (metadata parser, known as 'mp') that checks for:

- Compliance with Production Rules: mandatory elements, fixed domains, element format requirements (bounding coordinates in LAT/LON decimal degrees, date formats, and URLs), and record format requirements (order of compound and simple elements).
- Limited Logical Consistency: bounding coordinates (North value greater the South Value, East vs. West) and time ranges (start date earlier than end date).

Unfortunately, there is no currently available technique for assessment as to robustness (number of elements provided beyond the mandatory), spelling and grammar or validity of content.

Although, the standard specifies the information content of metadata for a set of digital geo-spatial data, it does not specify how this metadata should be encoded. It should be noted that metadata encoding is required to check its syntax against a metadata complier. Therefore, it may be necessary to devise the specification for metadata encoding in order to develop and use a metadata compiler. However, it has been proposed that the standard uses SGML[23] to support metadata loading, exchange and presentation.

The standard provides excellent documentation of a data set from the geospatial perspective. However, this excellence fails to prevail from the perspective of other information domain, such as Biology, Geology etc, where it is limited and, in some aspects, inadequate, for describing data [SMT96]. Geologists, for example, may require specific, keyword-searchable information about the types of rock strata that might otherwise be described in a free text field. Biologists might need specific information on species or habitat associations.

The essence of this specification is to let potential users identify the suitability of a data set to their purpose, obtain the information, and contact the creators of the data for further information if necessary. However, the original creators of the data may not be available over the long life cycle of preserved data, so it will be important to complete metadata documentation sufficiently to ensure utility of the data over decades or centuries. Therefore, from this perspective, even the full set of CSDGM metadata standards may not be sufficient.

Nonetheless, in its generic form, the CSDGM is fairly flexible for describing different types of data. Furthermore, it is possible to map CSDGM standard to many other existing recognised standards such as, DIF (the NASA Directory Interchange Format), GILS (Government Information Locator Service) and the Dublin Core. This may be useful in creating a specialized set of metadata elements for a specific information domain. In addition, unlike Dublin core, the elements of CSDGM standard, although may be too

[23] Standard Generalized Markup Language: (SGML) is a standard for how to specify a document markup language or tag set. E.g. HTML, XML – SGML based language.

numerous for long-term maintenance, will certainly prove to be sufficiently comprehensive for documenting data sets.

3.4 Data Documentation Initiative

The Data Documentation Initiative (DDI) is an XML-based standard for the content, presentation, transport, and preservation of metadata for the social and behavioural sciences data resources. Originally, DDI was an endeavour to cater to the need for well structured, both machine and human readable scientific documentation by providing a more modern and Web-aware specification than the existing and widely used OSIRIS Codebook/data dictionary. Consequently, this new specification could be used to structure the description of the content of social science data archives. Having originated in 1994 in the Inter-university Consortium for Political and Social Research, the current version (2000) of DDI is still a specification and yet to become a formal ISO approved standard [DDINA].

The DDI specification has been designed to fully encompass all kinds of data originating from empirical observations of the social and behavioural sciences derive from surveys, censuses, administrative records, experiments, direct observation, and other systematic methodologies for generating empirical measurements within its around 300 elements. These elements are represented as hierarchical tree-like structure, which is divided in to five main branches or sections with various sub-sections. Appendix B provides brief description of these sections and subsections [JRD02].

3.4.1 DDI Assessed

The DDI specification compartmentalizes the metadata elements in five major sections in order to support the modular development of system functionality and ease the task of improving and maintaining the metadata and associated data sets. In addition, the DDI contains 50 suitable elements [KAD01] among its 300 elements to describe qualitative data sets and many of these elements do not require any special adjustments [RED01].

The DDI is a very rich specification with defined placeholders for almost any piece of information that a data producer or distributor might find appropriate to associate with a dataset. However, this richness is the main reason for only one element of the specification, *Abstract*, declared as strictly obligatory. This is creating problems for application providers that need more predictability as to the type of information they can expect to find in a DDI instance as well as making the specification inefficient from the interoperability point of view [JRD02].

The DDI provides controlled vocabularies for a number of attributes. An example may be, *"the type-attribute of the "file structure" element which might take the values: (rectangular|hierarchical|relational)"*. However, the specification fails to provide any controlled vocabularies that guide (if not actually govern) the use of key type and subject attributes that are permitted throughout the DDI [JRD02].

The DDI facilitates the production of multilingual metadata instances, by associating an *xml-lang* attribute with every DDI-element contains. However, this may requires every element of the DDI to be repeatable, which may consequently corrupt the cardinality of the metadata structure creating severe difficulties for any processing software.

The specification provides “a strategic component of the infrastructure necessary to support the exchange of structured social research survey data” [RED01], consequently making it survey-data biased. This implies that the DDI is less complete for other kinds of scientific data, specifically; time series data and aggregate data (such as census tables) are treated much less thoroughly than survey data. However, work is in progress to boost DDIs ability to move beyond its original domain and to bridge the gap between the different data oriented communities.

Within the current version of the specification, there are no ways to add local extensions without compromising the interoperability of the core specification. This implies the inability of the specification to facilitate the specific needs of a given application or resource type. This limitation is a consequence of the inherent limitations of the XML DTD framework, which also accounts for the lack of modularity in the specification [RED01].

Furthermore, there are currently no tools available for generating DDI compliant metadata for those new datasets that have not yet been documented as well as legacy data for which codebooks and documentation has already been prepared. In addition, unlike CSDGM lite (see section 3.3), no lightweight version of the DDI that covers application of high-use elements (and probably 85% of the datasets for which the DDI actually applies [RED01]) is available; consequently may prove to be error-prone and cumbersome for long-term metadata management.

Aside from the limitations as mentioned above, the DDI specification is a potent standard for metadata with features, such as interoperability, increased search-ability etc. Besides, it is expected that the development of the future version of the specification will focus on eliminating these limitations and presenting a more flexible, modular and extensible specification.

3.5 Global Information Locator Service Metadata Standards

The Global Information Locator Service (GILS) metadata standards were developed pursuant to U. S. Public Law 44 USC 3511[24], to describe government agency information resources, serve as surrogates for those resources, and support networked information discovery and retrieval. The main goal of the GILS is to make it easier for people to find all of the information they need. Fundamentally, GILS is about managing information content, not just collecting new information technologies [GLS04].

The GILS provide 67 elements in total, 23 of which are core elements, consisting of mandatory and optional elements with fields to indicate whether the element is mandatory, repeatable, or controlled (i.e., if only a limited set of values may be used to record data). Some of the elements are compound, consisting of other sub-elements. A list of these elements with brief descriptions has been given in Appendix B.

[24] Requires establishment of "a distributed agency-based electronic Government Information Locator Service, which shall identify the major information systems, holdings, and dissemination products of each agency."

3.5.1 GILS Assessed

The GILS Core elements describe three different types of information resources. The first type of information resource that Federal agencies must describe with a GILS Core entry is locator to information dissemination products. These locators (not to be confused with the GILS itself) catalogue or describe information dissemination products, such as books, CD-ROMs, publications, studies, reports, and patents, regardless of medium. The second type of information resource that must be described in the GILS Core are automated information systems[25] that may be used for the collection, processing, maintenance, transmission or dissemination of information. The third type of information resource that is described by GILS Core elements is Privacy Act systems of records in electronic, paper or mixed formats [GLS04].

Unlike Dublin core, the GILS metadata standards are best described as being fairly *high* on the scale of fullness and complexity, i.e. much more comprehensive than Dublin Core. For example, in addition to the core elements, it also contains a number of elements subsets for dealing with simple geospatial and temporal metadata. However it was not specifically designed for high-level geospatial data [LDR97].

In order to provide extensibility, it is also permissible to use locally defined elements within GILS records in addition to the GILS Core Element set. Besides this, the GILS Data Element set contains two elements called *Language of Resource* and *Language of Record* to describe multi-lingual data. Also, the *Cross Reference* elements of the GILS Element Set provides for the ability to describe relationships between metadata records. The *Cross Reference* element subsets are also intended to be used inside *Controlled Subject Index Subject Thesaurus* structures for the purpose of describing where to acquire and reference the thesaurus. This, consequently, ensures the consistency of the description [GLD00].

The use of the *Controlled Vocabulary* element of the specification enables more search efficiency than normal text based search. In addition, if the promise of effective searching in a distributed computing environment is to be met, proper use of controlled vocabulary must be made as it enables the description of the resources to be as complete, accurate, current, and consistent as possible.

Similar to the CSDGM metadata standard, one potential shortcoming of GILS specification (even the full set of elements) is insufficiency in completing metadata documentation adequately to ensure utility of the data over decades or centuries. It should be noted that sufficient documentation of metadata is vital to address any query or problem related to the actual resource, as the original creators of the resource may not be available over the long life cycle of preserved data to address such queries.

In general, GILS define an open, low-cost, and scalable standard so that governments, companies, or other organizations can help searchers find collections of information, as well as specific information in the collections [GLS04]. Despite being fairly complex, the standard provides very comprehensive metadata format with enhanced search efficiency and interoperability. Therefore, the GILS standard, though does not directly address long-term metadata management issues, may prove to be useful in creating an appropriate metadata format for the job by mapping it to other useful standards.

[25] An automated information system is a discrete set of information resources organized using information technology as defined in OMB Circular No. A-130.

3.6 Directory Interchange Format

The Directory Interchange Format (DIF) of Global Change Master Directory (GCMD)[26] evolved in the late 1980s as a de-facto standard used to create directory entries, which describes all types of Earth science data sets. Rather than competing with other metadata standards, the DIF simply provides a "container" for the metadata elements that are maintained in the International Directory Network (IDN)[27] database of GCMD, where validation for mandatory fields, keywords, personnel, etc. takes place [GML04].

In order to detail specific information about the data, DIF provides a collection of fields, six of which are required in the DIF; the others expand upon and clarify the information. Some of the fields are text fields; others require the use of valid values. In addition, some of these fields are high-level fields, are composed of other child fields [GCM04]. A list of DIF fields has been given in Appendix B.

3.6.1 DIF Assessed

Similar to Dublin core standard, the DIF allows users of data to understand the contents of a data set. The DIF fields permit a complete set of descriptors that allow the researcher to make more informed choice among data sets. In essence, the DIF metadata elements aim to demonstrate the need for metadata requirements to be flexible and to continually evolve. However, unlike Dublin core, all of DIF fields are not repeatable, thus restricting the possibility of inadvertently added additional fields and minimizing efforts required for management.

In order to provide the flexibility to be able to catalogue the different types of data, the DIF offers the small amount of required or core metadata elements, known as Skinny DIF. Skinny DIFs are put into a directory to alert users of the existence of a particular data set, and may be modified at a later time to provide additional information. This also ensures interoperability among these data centres, which would be extremely difficult to obtain when using all of the GCMD fields. However, the full list of metadata elements is available as needed [GCM04].

The DIF metadata outside the core elements are deemed critical, meaning these elements are crucially important for data set selection (*i.e.*, searching), user understanding of the data, or data access. This also enables the creator of the metadata to focus on search-ability of the metadata. For example, if a user conducts a search by the fields critical for searching (i.e. *Parameters, Temporal_Coverage, Spatial_Coverage* and *Location*) and the DIF does not contain the information, the DIF will not be found in the search. In addition, in order to provide increased search efficiency of the metadata, the GCMD actively maintains controlled keyword lists or vocabulary for use with fields found within the DIF (Directory Interchange Format) document. While a full text search will produce higher recall, a controlled vocabulary search will result in higher precision.

[26] Operated By National Aeronautics and Space Administration (NASA).

[27] Consists of three coordinating nodes representing the international science community such as, these are the American node, the Global Change Master Directory at NASA/Goddard Space Flight Center, in Greenbelt, Md.; the Asian node at the National Space Development Agency of Japan in Saitama, Japan; and the European node at the European Space Agency/European Space Research Institute in Frascati, Italy, to share standardized data set descriptions.

The DIF Specification addresses the quality issues of metadata by allowing the users to perform different validity checks on both metadata syntax and semantics. In addition to providing a minimum number of mandatory fields, the DIF requires its two of the searchable fields, spatial and temporal coverage, to be expressed in the same format. The DIF also requires the terminologies used to describe a scientific concept in different records to be the same. This effectively enables user to write additional validation checks into appropriate software to ensure that both syntax and semantics are correct [JCD01].

Furthermore, the specification ensures the quality of the actual resources within four of its optional but critical fields. The field "quality" provides Information about the accuracy of the data or any quality procedures followed in producing the data described in the DIF. This may be deemed critical for verifying the accuracy of the data substantially long period of time after the data has been created. Moreover, two fields "access constraints and use constraints" restricts any unauthorized access to the actual data and helps the users become aware of any placed constraints on the data. Besides, the specification supports multi-lingual metadata through its "Data Set Language" element/field.

As far as the different versions of same metadata records are concerned, it is not explicit how the specification addresses this issue through its "Metadata_Version" core/mandatory element/field. In addition, its ability to be customized to cater for specialised needs of any other generic information domain rather than scientific domain is also questionable. For example, one of the mandatory fields, "Scientific Keywords", may not be quite appropriate for describing business-oriented resources.

Over the years, with every new version of DIF, new metadata fields have been added to address increasing the complexity and robustness of both the metadata and metadata management system. In general, the specification provides a number of desirable features, such as, minimum number of required fields, ability to be mapped to other standards (e.g. CSDGM) etc. that may be deemed useful for long-term metadata management. Although it may not be sensible to reach a concrete conclusion regarding the effectiveness of the specification in terms of long-term metadata management, with a number of desirable and potentially useful features, the DIF is certainly a competent candidate for the job.

3.7 CLRC Scientific Metadata Model, version 1 [BSK01]

The CLRC Scientific Metadata Model, version 1, as its name implies, is intended to provide a high-level generic framework for describing scientific data from any discipline. Developed at the Central Laboratory of the Research Councils (CLRC) in the UK, one of Europe's largest multidisciplinary research support organizations and Influenced by the CIP (see section 3.8) metadata catalogue for Earth Observation and the DDI (see section 3.4) metadata description for Social Science data, the model is based on a framework that consists of three main categories of metadata: Schematic, Navigational and Associative, providing variety of useful information about the resource. One example may be machine view of the resource, close to its physical representation, e.g. data formats, fields etc.

The framework is the basis of six abstract-level categories of metadata, the properties of which may be inherited by other domain specific metadata using an *object inheritance* mechanism. This explains the model's ability to provide specialized services to specific scientific disciplines. These abstract-level categories are formed by six major data areas contained within the CLRC scientific metadata record. The diagram below reflects this:

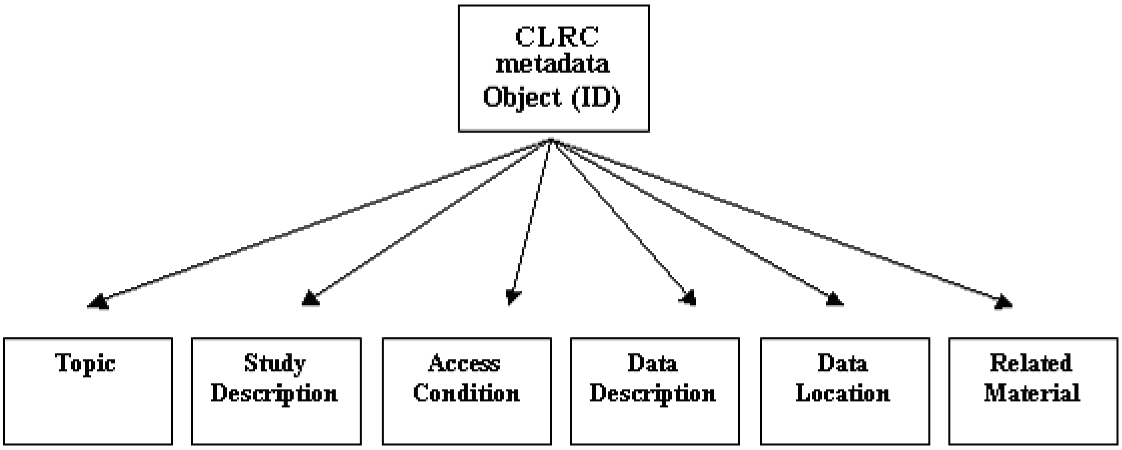


Figure 3.3: CLRC Scientific Metadata Model

Brief descriptions of these six metadata categories have been given in Appendix B. As the above diagram depicts, each CLRC metadata record is provided with its unique identifier for its reference.

3.7.1 CLRC Assessed

The scientific data described by the CLRC metadata model may be originated from general scientific data holdings, the contents of which may include information about different scientific experiments, raw data generated by scientific investigations, tools for processing such raw data and a set of files with physical location used to store processed data. The model provides sufficient metadata to access all layers of the data holdings either together or separately.

The standard attempts to overcome inappropriate search results being returned to the user by qualifying searches in accordance with different disciplines. This is done by issuing three fields (i.e. discipline, source, and keyword) to each keyword of its Topic metadata.

As in the "Time" field of the *Study Information* category of Study Description metadata, the standard addresses both date and time in the same field. Other well-known standards, such as Dublin Core, FGDC etc. usually have separate fields for these two attributes, hence it may be convenient for the users to have separate fields for time and date, thus eliminating the probability for any inadvertent errors.

The standard has another field within *Study Information* section, i.e. "Data manager" that provides a description of the primary organization(s) responsible for curating the data. This may prove to be useful in long-term data preservation as the contact information for forwarding any query in reference to different aspects of the data (e.g. quality, error etc.) will be available.

Currently, the "Units"[28] field of Parameters and conditions measured by an experiment or simulation in the Study hierarchy[29], presumably, accepts any scientific unit. Now if this scenario were to change, where the field would automatically convert the unit entered to any other specific unit, it would require appropriate conversion formulae to be built in, which would use up significant amount of storage. On the contrary, to many users, it might be desirable to have a unit conversation facility. Alternatively, if a list of built in units were provided for the users to select the required unit from, the list would need to be updated to keep pace with the exponential growth in the scientific world where new units emerge often.

Although, the standard has proposed a model to provide generic metadata for providing access control features within the metadata model, at the time of writing this report, version 1 of the standard had no access control procedures in place. This is presumably to be addressed in the forthcoming version of CLRC metadata model.

The *data descriptions* metadata within the metadata model, cover all its data sets, each data containing a name, some both a logical and file description metadata, and a set of files, each file having a name, some both a logical and file description metadata. This aids in constructing a recursive hierarchy of descriptions, which enables searching for the parent data set's metadata, in case metadata for an item (e.g. a file) cannot be found.

Aside from aforementioned features, CLRC metadata model addresses[30] storage of data and provides the users with information regarding how to access the data. These features are missing in well-known metadata standard like Dublin Core.

Finally, the *Related Materials* metadata of the standards provide contextual information about the resource being described, such as references to literature relevant to the resource, references to controlled vocabularies describing the subject of the data etc. These metadata may prove to be useful in providing the users with better understanding of the data, hence aid in its appropriate and efficient use.

In general, the CLRC scientific metadata model does not address the long term metadata management issues (see section 2.4) in direct terms but proposes a model which necessitates an implementation (e.g. Relational, XML etc) to store such information for easy data mining; and/or a template for the categories of information (in whatever format), that should be stored to capture useful meta information and information produced by studies/experiments. At present, the CLRC scientific metadata model is undergoing further development and improvements. Therefore, the future version is expected to eliminate all probable drawbacks of the current standard, and provide a fully potent metadata standard for scientific information domain.

[28] The unit of measurement in which the value is recorded.

[29] Break down structure of the Study information section of Study Description metadata.

[30] E.g. Data holding location field of data Location metadata

3.8 Catalogue Interoperability Protocol

The Catalogue Interoperability Protocol (CIP) was developed by the Protocol Task Team within the Committee on Earth Observation Satellites[31] (CEOS) in 1997 as an endeavour to establish long-term concepts, guidelines and standards for interoperability. Essentially, the development of the CIP was an international collaborative effort led by the European Space Agency (ESA), NASA, and DLR with contributions from the other CEOS members [PSC97].

The main objective of the CIP is to facilitate the access, searching and retrieval of Earth observation data. In other words, CIP aims to enable the user to search many physically distributed Earth observation data catalogues (without having to separately interrogate each one and manually correlate different sets of search results), effectively allowing all the data archives to appear as one database [PSC97].

With about 300 items of metadata in a hierarchical manner, the CIP defines the method by which data retrieval (location, requesting and delivery) is to be performed. In addition, it defines the method of interpreting data requests, queries, and search results, and also provides the definition of a data dictionary used to specify the common attributes that describe the primary objects within a catalogue system [PRNA].

The CIP was developed as a three iterative phases with output from one phase feeding into to the next in order to implement a number of useful features within the specification. These features include general catalogue facilities, e.g. browse, inventory, guide etc. and introduce the concepts of semantic attributes and hierarchical 'collections' of metadata. Besides these, the CIP offers a number of other features, such as ordering, security and administration facilities for metadata as well as features of a more complex nature, e.g. invoicing, accounting, etc. The hierarchical 'collections' of metadata within the CIP are organized thematically over a number of physical databases. The figure 3.4 illustrates the concept of hierarchical 'collections' [PARNA].

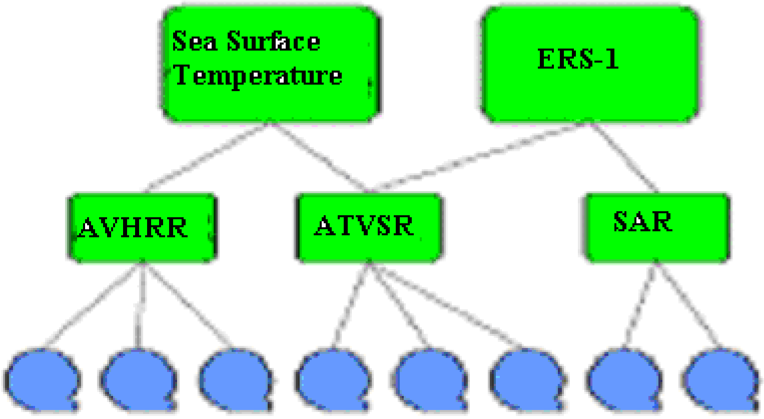


Figure 3.4: CIP Collection

[31] http://gds.esrin.esa.it/CCEOSinfo

3.8.1 CIP Assessed

The CIP specification offers some very useful features to facilitate interoperability, increased search-ability and ease of navigation for metadata. However, this information is not adequate to determine its usefulness for long-term maintenance of metadata. Furthermore, no information was found regarding its capability to be mapped onto other standards. Although, 300 elements may seem to be too numerous for maintenance, very limited information was obtained as to whether all of these elements are obligatory or optional or if any controlled vocabulary is used to add structure and predictability to the specification.

3.9 Open Archival Information System Reference Model

The International Standards Organization (ISO) Reference Model for an Open Archival Information System (OAIS) was developed, in May 1999, by the Consultative Committee for Space Data Systems (CCSDS) as a result of an effort to develop archive standards for the long-term storage of data in digital form. In essence, the OAIS reference model is a conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over substantially long period of time [OCS02].

The OAIS defines a range of functions, which are applicable to any archive, whether digital or not. These functions support the operations of the archive from receiving materials to archive (ingest), through storage, data management and administration, to the dissemination and release of the materials to those outside the archive (access). The figure below reflects these:

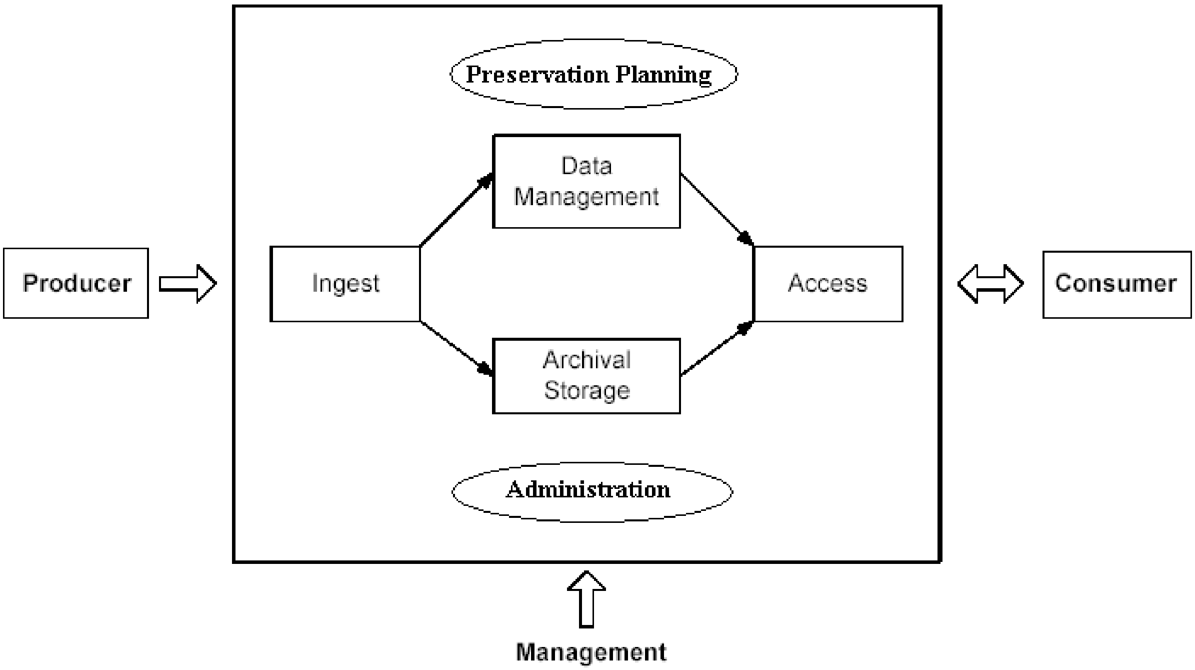


Figure 3.5: OAIS and its Environment [OCS02]

The OAIS model has identified and distinguished various types of metadata needed to support a digitally preserved resource. In accordance with the OAIS, each resource is packaged together with its metadata, as an ‘Information Package’. An Information Package combines two things: ‘Content Information’ and ‘Preservation Description Information’ (PDI). The Content Information groups the preserved digital resource, or *data object*, with ‘Representation Information’ (RI) metadata; the RI is the information needed to retain meaningful access to the preserved data object. The PDI groups different kinds of descriptive metadata, so that what the Content Information actually is can still be understood indefinitely.

3.9.1 OAIS Assessed

To begin with, the OAIS reference model is neither a metadata specification like Dublin Core, CSDGM etc. nor a metadata model in itself. The portion of the reference model that is of direct relevance to the issue of metadata in the context of long-term preservation is the information model embedded within the OAIS framework. The OAIS information model broadly describes the metadata requirements associated with retaining a digital object over long term.

Nevertheless, it focuses mainly on the functions and processes for preservation, not on metadata, which is however an essential part of the whole model. In short, the OAIS reference model does not, at least not in direct terms, address the issues of long-term metadata management. However, considering different services and functions[32] provided by the data management entity (figure 3.5) of the model, one may perceive that the entity of the model may well subsume different functionalities required for metadata management.

Although, the model focuses only superficially on the issues involved in perpetuation of digital information, it presents an archival model that ironically provides relatively little information about preservation. In fact, the model’s focal point lies in the processes of describing, packaging and manipulating stored information. This is mainly due to the fact that the model prefers migration as the only logical strategy for preserving information for long-term, despite recognizing the potential problems the strategy may pose. However, even with this strategy, it is not clear where the migration processes take place in the OAIS. Moreover, it dismisses emulation, a potential solution to the problems posed by migration, on the ground that it is “a major technical and economic risk” [EJR00].

It has been recognized that “update” operations are integral parts of any preservation process, in order to ensure that the information remains up-to-date. The OAIS reference model provides mechanism for updating the contents of document stored in Archival storage through the “Archive Information Update” function of its Administration entity. This update function operates by accessing the document, updating its content, and resubmitting it to Ingest entity (figure 3.5). However, the reference model does not clarify if and/or in what way this function belongs to a preservation process [OPM02].

[32] This entity provides the services and functions for populating, maintaining, and accessing both Descriptive Information, which identifies and documents archive holdings and administrative data used to manage the archive. Its functions include administering the archive database functions (maintaining schema and view definitions, and referential integrity), performing database updates (loading new descriptive information or archive administrative data) etc.

In spite of the arguable “deficiencies” as mentioned above, the OAIS reference model has proliferated rapidly through the digital preservation community and been explicitly adopted by, or at least informed, many prominent digital preservation initiatives. This is mainly due to the fact that the OAIS information model represents a high-level description of the types of information generated by and managed within the functional components of a complete archiving system. It makes no presuppositions either about the type of digital object managed by the archive, or about the specifics of the technology employed by the archive to achieve its goal of preserving and maintaining access to the digital object over long term. As such, the model provides a useful foundation for developing a preservation metadata framework of wide applicability.

From the perspective of long-term metadata management, the CEDARS project (see 6.1) is probably the most relevant of those endeavors. The metadata specification proposed by CEDARS project will allow for different manifestations or version of the same data object within the archive via reference links to previous and subsequent versions.

However, in the context of long-term data preservation, the NEDLIB (see 6.2) proposed to extend the OAIS model with a main function for long-term preservation, which was later accepted by CCSDS. In addition, the Working Group on preservation metadata of OCLC/RLG (see 6.3) published sets of metadata for preservation purposes; which uses the OAIS reference model as foundation. Generally, this published metadata sets attempt to identify more precisely what metadata are necessary to preserve (certain types of) digital objects.

All of these aforementioned adaptations of the OAIS reference model offer metadata specifications that are comprehensive and adequate for long-term successful data preservation to certain extent. However, the effectiveness and suitability of these specifications are yet to be proven, as, till date, they are more or less proposals rather than ISO approved standards. Nonetheless, it will not be unwise to consider these specifications as reliable foundations for developing a metadata specification, which will address all aspects of long-term metadata management rather than only preservation.

3.10 Metadata Standards' Assessment Matrix

The table below provides a matrix summarizing what the comparison between eight above described metadata standards, on the basis of their assessments in the context of digital curation, has yielded.

Metadata Standard	DC	CSDGM	DDI	GILS	DIF	CLRC	OAIS
Comprehensiveness for Long-term Management	Poor	Good	Average	Good	Average	Average	Very Good
Simplicity	Simple	Complex	Adequ-ate	Complex	Complex	Adequ-ate	Compl-ex
Syntax Independent?	Yes	Not Known	Not Known	Not Known	Not Known	Yes	Yes
Metadata Interoperability	Poor	Good	Good	Good	Good	Average	Good
Customizability	Good	Average	Poor	Good	Poor	Good	Very Good
Resource Discovery	Average	Good	Good	Very Good	Very Good	Very Good	Very Good
Record Preservation Technique?	No	No	No	No	No	Yes	Yes
Metadata Encoding Method	No	No	Not Known	Yes	Not Known	Not Known	Not Known
Controlled Vocabulary	Yes	No	Yes	Yes	Yes	Yes	N/A
Has a Lite Version?	N/A	Yes	No	No	Yes	No	N/A
Support for Multiple Languages	No	No	Yes	Yes	Yes	No	N/A
Support for Metadata Versioning	None	Indirect Support	None	None	Direct Support	None	None
Support Metadata Validity Checks?	No	No	No	No	Yes	No	Yes
Toolset Support	No	Yes	Not Known	Yes	Yes	No	N/A
Map-able to Other Standards	Yes	Yes	Yes	Yes	Yes	Not Known	N/A

Table 3.3: Metadata Standards' Assessment Matrix

It is to be noted that the CIP standard has not been included in the above assessment matrix as it was not possible to acquire sufficient information to assess the effectiveness and suitability of the CIP standard in the context of digital curation.

Due to the limited scope of this chapter, it was not possible to include all metadata standards assessed during the research. However, a summary of those additional reviewed metadata standards has been provided in Appendix C.

Chapter 4

Review of Related Published Works

A major part of this literature survey has contributed towards locating and reviewing all related published research works, articles, and journals etc. that pertain to different issues of long-term Metadata management (see section 2.4). This chapter outlines a select few of those related published works, which hold the most relevance towards this project. In order to provide ease of reading, these research works have been divided into separate sub-sections based on different information domains covered by the publications.

4.1 Generic Metadata Management

The particular published work that is the most relevant in terms of management of generic metadata is a PhD research by Shien-Chiang Yu & Kun-Yung Lu of Institute of Information Management, National Chiao- Tung University of Taiwan [SKR03]. Their research report, titled **"Metadata Management System: Design & Implementation"**, written in collaboration with Ruey-Shun Chen, an associate professor of the same institute, describes the design and implementation of a Metadata management system using XML (Appendix C) framework with various Metadata schema. As the article claims, this system is capable of eliminating the weakness of traditional object-oriented languages in information sharing as well as the constraints of storage and management between heterogeneous Metadata, while processing different Metadata information.

From the reading of the article, it can be apparent to one that the proposed system is composed of four modules: schema constructor, catalogue, metadata import/export, and enquiry. The assessment of the functionality of these modules, in the context of long-term metadata management, shows that there is no distinct quality assurance procedure for the metadata to be maintained by the proposed system. Although, the Schema constructor module, which is employed to provide the function of importing XML schema and establishing the system schema, contains a verifying mechanism for XML schema, which involves manual examination of the data format, extra function, input length etc. by a human operator. Now, it may be argued, how reliable this human controlled verifying mechanism is. Moreover, the description of the *data input* function that provides capability of metadata editing, which is also controlled by human operators, however does not provide any clear indication as to how the metadata quality is assured when performing various editing (e.g. duplicating, deleting, adding, updating etc.) operations on them. Besides, the storage mechanism employed by the Schema Constructor module for storing metadata may not be suitable for preserving metadata over the long-term.

Nevertheless, the functionalities of the system enable it to offer different useful features, three of which have the potentials to aid in the long-term management of metadata. One of these features is the ability to manage and store heterogeneous (i.e. originated from different sources) metadata. As technologies change rapidly, various new sources of metadata are emerging; hence this feature is definitely needed to cope with increasing heterogeneity of metadata. In addition, the system's ability to allow the users to retrieve metadata in various formats might also be useful for metadata management in the distant future when newer metadata formats will be in use. Moreover, in order to prevent inadvertent or malicious modification to metadata (i.e. XML schema), thus ensure the overall integrity of the metadata, the system provides access control facility, which includes audit control facility (i.e. who changed what) through the Authority Control function of its Catalog module.

4.2 Scientific Metadata Management

In 2002 Ruixin Yang, Menas Kafatos, and X. Sean Wang of George Mason University wrote an article titled, **"Managing Scientific Metadata Using XML",** presenting an XML based Distributed Metadata Server (Dimes), to manage scientific metadata in various formats and support sophisticated search and interactive data-access capabilities. The system described, comprises a flexible metadata model, search software, and a Web-based interface to support multilevel metadata access [RMX02].

The Metadata model, employed by the system, entails XML to integrate user-provided metadata mostly in its original form and to make all metadata searchable. Within the model, XML based metadata is wrapped by XML elements and this wrapping produces complete XML documents (with options to point to other XML documents) so that all the metadata are uniformly searchable. These XML documents that separately describe each data object; have tree structures, consisting elements, each of which has a unique parent element. Each of these elements with an ID attribute is called a node. Figure 4.1 reflects the relationship between the nodes of the metadata model.

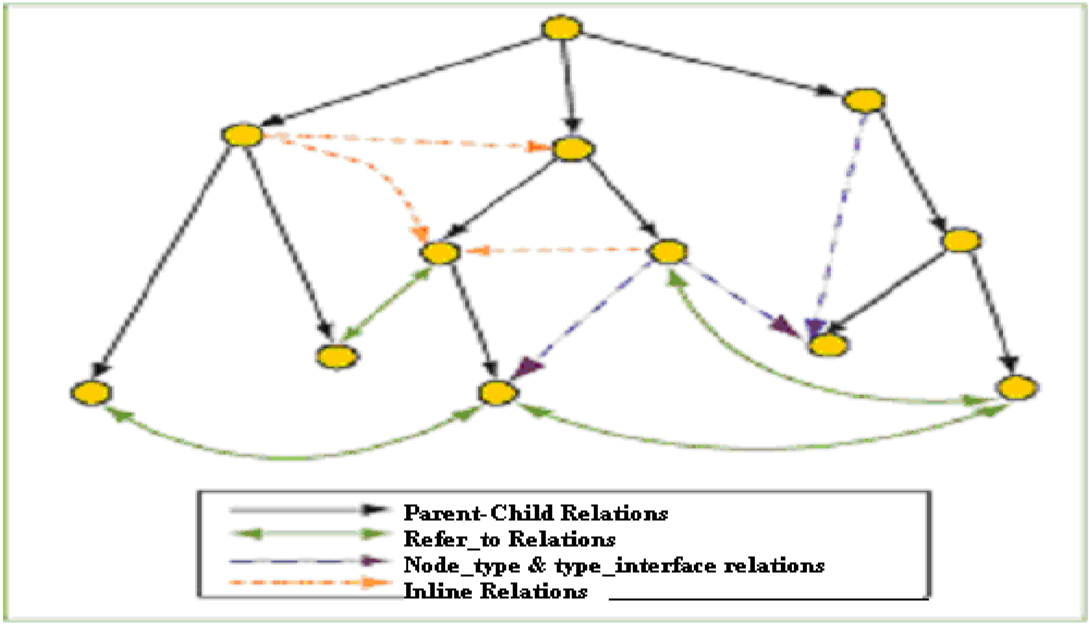


Figure 4.1.The node relation graph, where relationships between the structural parent-child are many-to-many [RMX02]

In the figure 4.1, **refer_to** relations are established by a type of attributes called **refer_to**, that is, if node A refers to node B, B also refers back to A. This attribute assumes a minimum semantics to link a pair of related nodes together. Node-type attributes are used to record a node's additional parents, and **type_instances** to record the reverse relationship.

This metadata model outlined in this article is claimed to be scalable as well as flexible. This implies that users are allowed to add new nodes and new links to the model to satisfy their metadata requirements.

As stated in this article, query engine is the key component of Dimes system, which is capable of handling new information without modification. Based on the XML4J package[33] with document-type definitions, this software answers queries against Dimes metamodel[34] compliant metadata by evaluating on each nodes of the metamodel using a breadth-first searching technique. The simplest queries handled by the engine are finding a node in an XML document by its ID attribute and any other queries for Earth science data involve finding data sets based on spatial or temporal resolution, spatial or temporal coverage, and textual conditions like keywords or free text. It is also capable of handling complex queries formed by combining the fundamental queries.

In addition, the article describes two prototype web interfaces for exploring Dimes' capabilities. The first prototype interface is for a regular search against Dimes, to allow users to search for Earth science data based on several criteria as mentioned before. The second is the Web-based Dimes metadata navigation interface that allows users to browse a metadata tree that groups nodes by various categories, each of which could be considered a dimension in a multidimensional database. This metadata navigation interface may serve one of the core requirements of scientific data curation - enabling exploration of related metadata. Nevertheless, article does not provide any information regarding how the access to the metadata is controlled (e.g. security etc.).

In order to solve the consistency problem for metadata search, Dimes is integrated with GDS[35] to create a Java based Scientific Data and Information Super Server (SDISS) that solves accurate data-search and outdated data-link problems by integrating metadata with the data systems. In order to ensure the quality of the new metadata being added to the repository, article describes several SDISS software components, including metadata-ingestion, metadata- merging, and cleaning modules. The ingestion component simply converts metadata from various sources into XML suitable for Dimes. To maintain the consistency and efficiency of the metadata repository after each SDISS update, merging module merges new metadata into current metadata, the resultant metadata then goes through a cleaning process in the cleaning module, where redundant nodes are deleted and all paired and symmetric links are fixed.

[33] www.alphaworks.ibm.com/tech/xml4j

[34] A tool's view of its underlying metadata or the details behind the metadata. Also, the graphical representation of an organised set of metadata requirements. Metamodels are depicted based on an underlying modelling methodology, e.g. object-oriented versus entity-relationship [ATM01].

[35] B. Doty et al., "GRADS and DODS," *Proc. 17th Int'l Conf. Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Am. Meteorological Soc., Boston, 2001.

In addition, the system employs XSL to ensure the conformance of existing XML documents or imported external XML files into our XML metadata repository, with the metadata model. This is a three-step process:

1. Extensible Stylesheet Language (XSL)[36] is used to define the elements and attributes to be compared and extract them.
2. XSL normalization is used on the extracted information so that strings can be compared to determine if the two nodes are the same.
3. Resultant metadata goes through the cleaning module for further checks, as mentioned above, to become ready for use.

However, it is not clear from the reading of the article how the interoperability of the metadata is ensured. As stated in the article, considering fact that this approach can only be applied to scientific communities such as bio informatics and space science, it may be inferred that this system may not be useful in any other scientific domain. In other words, Dimes system may not be customized to cater to specific needs of any other scientific communities than those mentioned in the article.

4.3 Educational Metadata Management

In 2001, Gyo Sik Moon, Taegu National University of Education, Department of Computer Education, Korea, attempted to address the issues of Educational Metadata Management in his research report, **"Design and Implementation of Metadata Management System for WWW Coursewares"** [GSM01]**.** As the title implies, this paper proposes a metadata management system with an aim to help search appropriate coursewares and shows that utilizing metadata for search can facilitate obtaining right information on the Web. The metadata in the context of this paper originates from four information domains with each domain comprising a set of elements, which characterizes the specific information domain. These domains are, document information domain[37], Web technology information domain[38], presentation information domain[39], and instruction information domain[40].

The metadata management system proposed by this paper comprises three components, which are Input Interface, Output Interface and Metadata Database. The input interface is divided into four sub-components in order to cater to the needs of four different types of users (i.e. developer, site manager, teacher, and student), who submit their different metadata related requests or searches through appropriate interfaces. The output interface, on the other hand, is designed to produce response to these requests or searches by the users. As the paper

[36] XSL is a language for creating a style sheet that describes how data sent over the Web using the Extensible Markup Language (XML) is to be presented to the user.

[37] This domain of information describes the overall picture of a courseware and consists of different elements, such as courseware information, author information, location information etc.

[38] This domain describes how technological features are incorporated into Web courses and comprises elements such as, accessibility to the course, ease of use, link-related characteristics, course management techniques etc.

[39] The domain focuses on the appropriateness of presentation of courseware contents. The elements of the domain consist of human interfaces, textual presentation, multimedia presentation, and types of presentation structure.

[40] The domain tells what instructional goal is and what strategy is used for achieving the goal. Elements of the domain are the following; instructional goal, instructional strategy, and instructional model.

claims, the search can be performed individually or collectively by one or more of the following search variables; topic, subject, courseware title, grade, author, and description of multimedia. Metadata database, which was implemented in the Microsoft ASP and built on an NT server, stores metadata extracted from the four aforementioned domains of information.

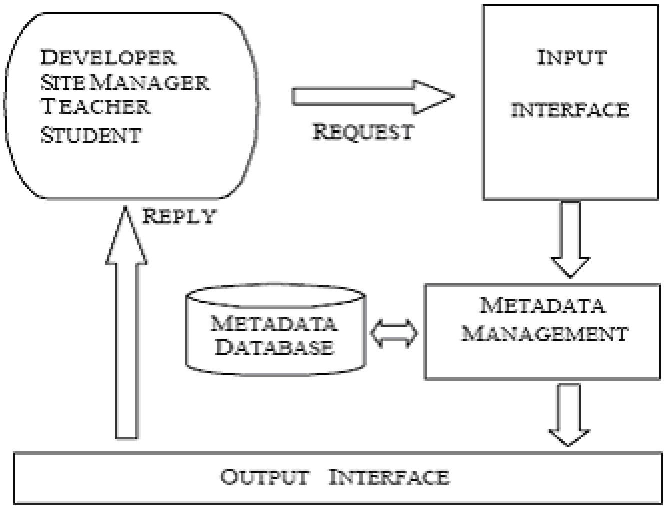


Figure 4.2: Metadata Management System Diagram [GMS01]

As the above diagram for the proposed metadata management system depicts, metadata from the users request as well as from the metadata database go through this metadata management component or, perhaps process, represented by a rectangle. Although, in the conclusion, the paper states that *"the management of metadata should be performed regularly by experts",* which involves addition of new materials, deletion of obsolete ones, and modifications to existing ones, it provides no indication as to whether this statement refers to the rectangle denoted as "Metadata Management" in the diagram. Furthermore, if the statement does refer to the metadata management rectangle, no information as to whether the metadata is managed on a long-term basis or how the quality of the metadata (e.g. access control etc.) is assured when it undergoes various operations such as, deletion, addition etc. is provided in the paper.

Another Research paper**, **titled**, "An Educational Metadata Management System Using a Deductive Object Oriented Database Approach"** by D. Sampson, V. Papaioannou, .N. Bassiliades, and I. Vlahavas of Aristotle University of Thessaloniki Greece, proposes an educational metadata management system, using a deductive, object-oriented database approach [DVN03]. The proposed solution is based on the architecture of the EM2 tool for providing the graphical interface for the interaction with the user. Typically, user interactions subsume creating a new educational metadata file based on metadata specifications, opening or edit/update data on an existing one, converting metadata files between specifications as well as creating maps for these conversions. In order to ensure the quality of the created or modified XML documents, they undergo structure and data validation, whenever possible.

Generally, all created or modified metadata files are stored in the associated XML repository of EM2. Rather than a database system, this XML Repository is a system folder storing the XML documents as files. In addition the DTD and XML Schema files are stored in their associated repositories respectively. The figure 4.3 depicts the architecture of the proposed system.

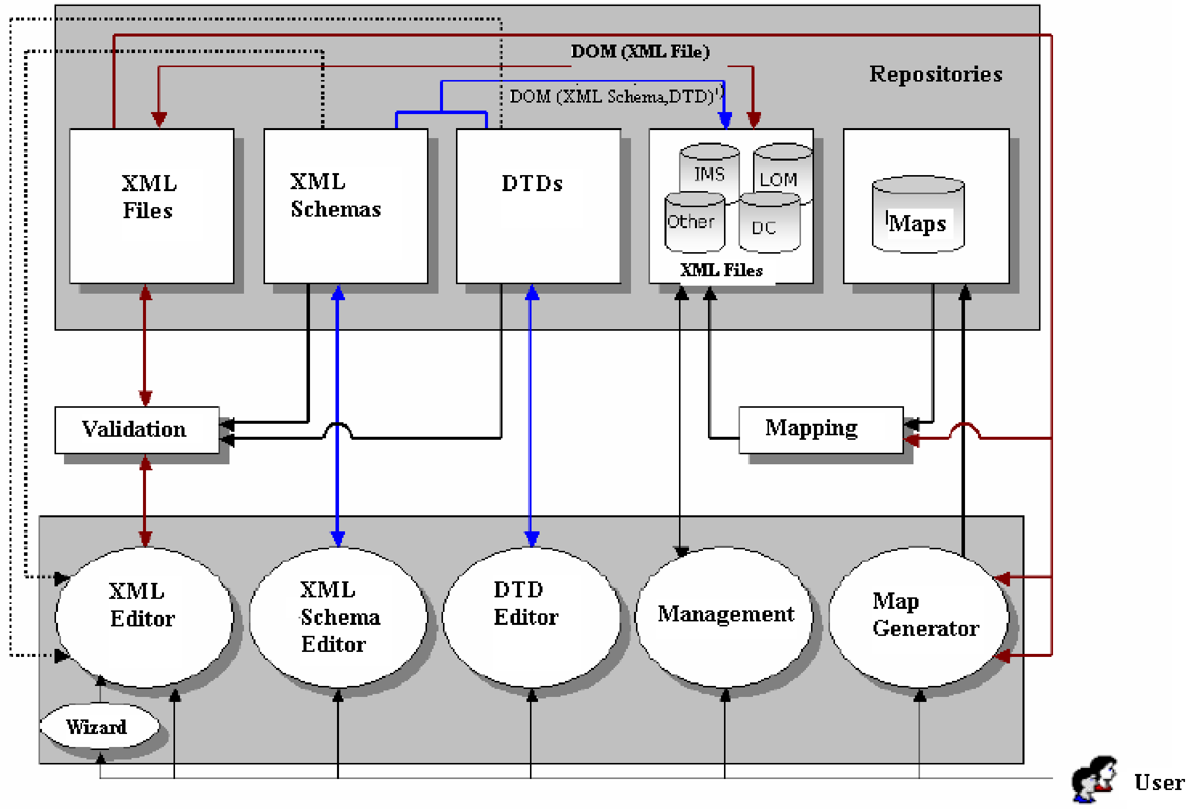


Figure 4.3: Architecture of an Educational metadata management tool [DVN03]

Furthermore, the EM2 tool may be used to provide database storage and data retrieval based on the user's queries by integrating ×-DEVICE to the EM2 tool. It is done by passing each stored XML document, together with its associated DTD to the ×-DEVICE system where the mapping to an Object Oriented Database (OODB) system takes place. Here, the OODB holds the data of every education metadata file that has been created, updated or stored in the EM2 XML repository. In addition to the graphical interface mentioned above, the proposed solution aims to provide another interface to facilitate users query submission.

The approach highlighted in the paper is an interesting approach for metadata management; that employs the architecture of an existing metadata tool with different features (e.g. authoring, editing etc.) integrated in to a repository type system for enhanced functionality. However, the reading of this research paper does raise a few questions. For instance, the paper does not provide any information as to how the access to the metadata within repository or OODB is controlled. Besides, no information is provided to explain the structure and data validation techniques to ensure metadata quality. Furthermore, it is not clear whether the proposed solution addresses the issues of metadata versioning within the two metadata storage facilities mentioned.

4.4 Data Warehouse Metadata Management

During the literature survey conducted for this project a significant number of research papers or articles that address metadata management from the perspective of data warehouse environments, were acquired. The most informative of them all, is a research paper written by Hong Hai Do, Erhard Rahm of University of Leipzig. The main focus of the paper titled, "**On Metadata Interoperability in Data Warehouses**", is on the problem of insufficient support for a consistent and comprehensive metadata management in current data warehouse environments to ensure a high quality of the warehouse data and provide sufficient flexibility to extend the scope of the warehouse to new information sources [HEW00].

In an attempt to address the problem stated above, the paper describes three main architectural approaches for metadata management in data warehouse environments, as outlined below:

1) **Centralised Approach:** this approach employs the central repository (see Appendix D) to manage shared[41] as well as tool/DBMS-specific metadata, instead of storing and maintaining metadata locally. The main advantage of this approach is that a non-replicated and consistent management of all metadata can be achieved. However, this approach poses two potential problems pertaining to compatibility between components from multiple vendors and overall system performance. This is mainly due to the dependence of all tools and their operations on central repository, which results in a loss of autonomy.

2) **Decentralised Approach:** In this approach, all tools and Database Management Systems (DBMS) possess their own (local) metadata repository and communicate with each other to exchange metadata. This supports maximal autonomy and performance for tool/DBMS-specific metadata.

3) **Shared Approach**: This approach tries to combine the advantages of two previous approaches. Each tool/DBMS contains its own repository for its local metadata thus supporting autonomy and fast access for this metadata. In addition, each component supports a metadata exchange interface to a common repository managing all shared metadata. In contrast to the decentralized approach, the number of tool-to-tool connections and the mapping overhead can be significantly reduced and metadata replication can be tracked and controlled centrally. Figure 4.4 depicts these three approaches.

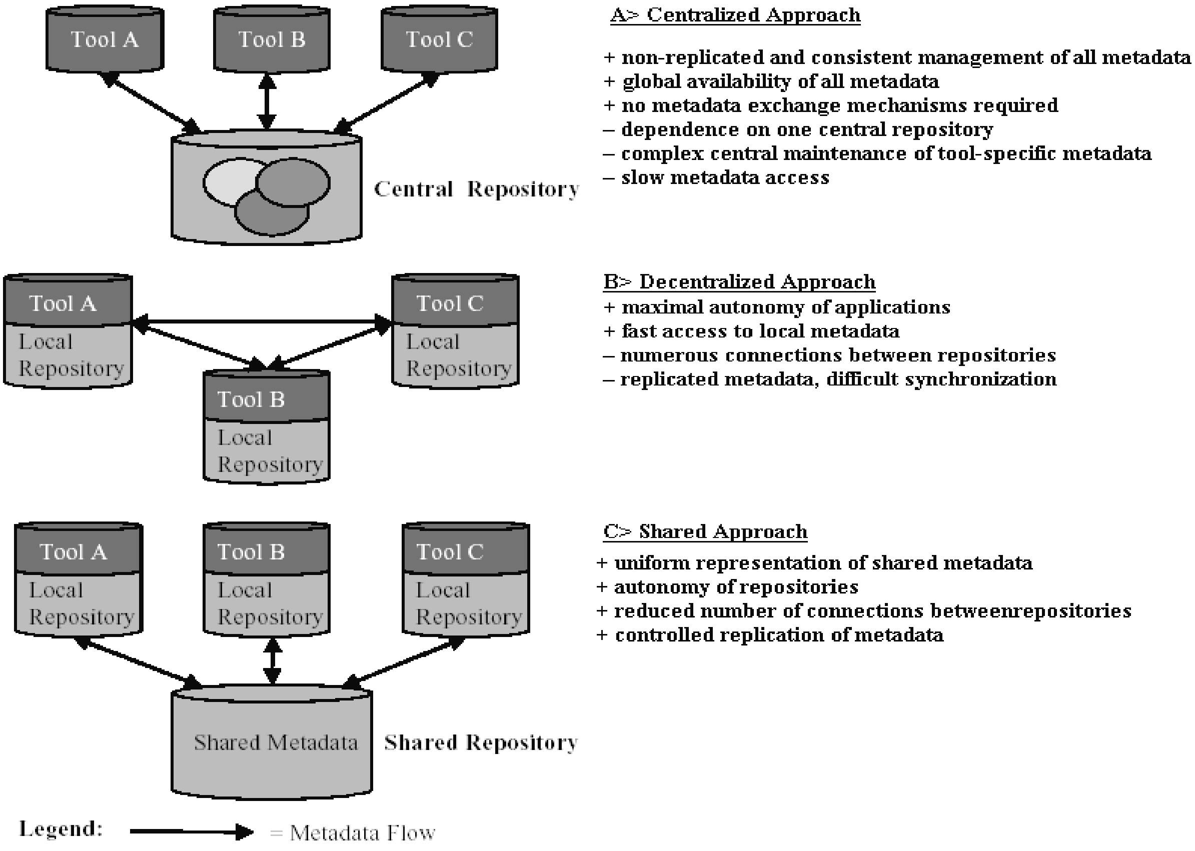
Aside from these three general architectural approaches, there is another approach mentioned in the paper. Effectively, this approach is just combination of the aforementioned approaches with a mixed or hybrid architecture, hence called **Mixed Approach**. This approach employs a shared repository for managing globally relevant metadata in order to achieve a controlled flow of metadata from the tools being used.

In order to provide better support for metadata interoperability within data warehouse environment, the paper proposes a federated architecture utilizing aforementioned shared approach in addition to the tool- and DBMS-specific repositories. In addition, the paper identifies the inability of commercially available standard metadata models for data

[41] Metadata required by more than one components/tools.

warehouses, such as OIM and CWM (see Appendix C) to cover all kinds of relevant metadata, despite being widely supported in commercial tools.

In a typical data warehouse environment, shared metadata is required to flow between different components, resulting in metadata replication or different version of the same metadata. In order to ensure the consistency between these versions and their relationships between each other, automatic detection and propagation of updates on these metadata, followed by application of the updated metadata within a repository are required to control different versions of metadata. The paper attempts to address these issues by discussing major alternatives and proposing the use of a lazy replication control with deferred update propagation.



A> Centralized Approach

+ non-replicated and consistent management of all metadata
+ global availability of all metadata
+ no metadata exchange mechanisms required
– dependence on one central repository
– complex central maintenance of tool-specific metadata
– slow metadata access

B> Decentralized Approach

+ maximal autonomy of applications
+ fast access to local metadata
– numerous connections between repositories
– replicated metadata, difficult synchronization

C> Shared Approach

+ uniform representation of shared metadata
+ autonomy of repositories
+ reduced number of connections betweenrepositories
+ controlled replication of metadata

Figure 4.4: Architectural Approaches for Metadata Management [HEW00]

The “lazy” approach claims to eliminate a major problem with currently employed version control approach. In short, this major problem lies in the fact that a given repository (subscriber) may obtain metadata updates from multiple sources (publishers) with different propagation methods and different timing approaches. As a result, even the most recent metadata objects can refer to different points in time making it difficult to group them together so as to obtain transaction-consistent versions of the metadata. The proposed approach controls this problem by not independently propagating every metadata change but by only periodically performing such updates. For instance, modifications of source schemata may be propagated in batches, e.g. weekly or together with changes of the warehouse schema.

In summary, it is reasonable to state that the techniques and approaches described in this article may be useful in ensuring metadata interoperability and controlling different metadata versions within a data warehouse environment. However, should these concepts and principles be applied in the context of long-term metadata management with the objective of efficient and proper re-use of good quality data over long time, further examinations will need to be conducted on them to determine the degree of their effectiveness in such complex area.

Another publication [GEM02] that attempts to address the issues of metadata management within data warehouse environment is a project report written by Gunnar Auth, Eitel von Maur and Markus Helfert of Institute of Information Management, University of St. Gallen. The paper mainly presents a software architecture that was developed for metadata management within data warehouse environment at a leading Swiss financial services group, based on a common shared meta-model (see Appendix E).

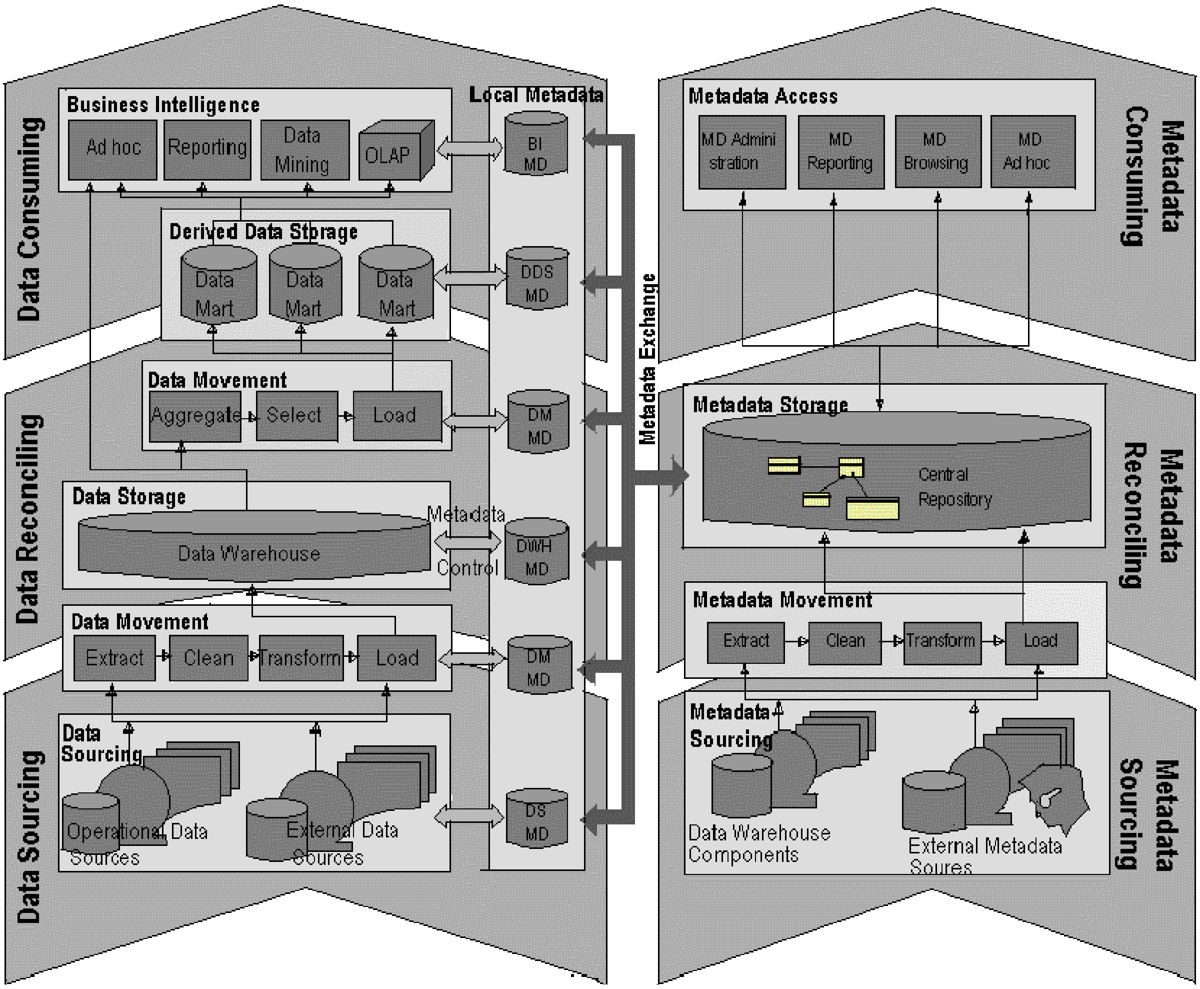


Figure 4.5: Software Architecture for Metadata Management [GEM02]

The software architecture presented in this paper titled "**A Model-based Software Architecture for Metadata Management in Data Warehouse Systems**", centres around a common shared metamodel based on OMG's Common Warehouse Metadata model (See Appendix C) and contains a number of layers, each providing different functionality, which contribute toward the total metadata management. As figure 4.5 reflects, the architecture's bottom layer, which is called 'Metadata Sourcing', comprises all kinds of metadata sources, such as data warehouse database, data marts etc. The next layer called ‘Metadata Movement’ provides the abstract functionality for extracting, cleaning, transforming and loading metadata into the repository. It corresponds to the according layer of the data warehouse system and has rich potential for reuse of code and data artifacts proved in the system.

The subsequent layer “Local Metadata” stored Metadata that is constantly produced and consumed within the data warehouse system during design, operation, and implementation phase of the data warehousing process. Examples include data structures, ETL mappings, and field descriptions etc. The architecture contains a repository located at the central ‘Metadata Storage’ layer to store and maintain global and shared metadata. This repository also serves as a hub for metadata interchange using a common metamodel-based CWM model. It is to be noted that Metadata and metamodels are exchanged between the central repository and local metadata stores utilizing XMI[42] as a standard interchange mechanism.

In order to enable different software components within the architecture to export metadata from other sources, an adapter that understands both the common metamodel and the internal metadata representation, is provided. This adapter facilitates exporting metadata in XML format. The top layer of the architecture, “Metadata Access” provides means (i.e. components) for accessing metadata by users. This layer also facilitates metadata administering by developers for editing and updating metadata as well as managing security by granting user privileges and managing versions and configurations of metadata.

In summary, the software architecture presented in this paper appears to be detailed enough for addressing the major issues associated with metadata management within data warehouse environment, and should be able to suit user requirements ranging from end users to database administrators. The approach is also cost-effective and reduces complexity, as the paper claims. However, there are a few aspects of the approach that are not explained in the paper. For example, although the paper mentions that the top layer of the architecture deals with metadata versioning, it does not quite clarify on the actual metadata versioning technique. Furthermore, the paper emphasizes on the fact that metadata available within the repository must be reliable, consistent and up-to-date. Nevertheless, it fails to provide any clear indication as to how the described software architecture addresses these issues.

4.5 Approaches for Metadata Quality Assurance

In 1998, William E. Moen, Erin L. Stewart[43] and Charles R. McClure[44] wrote a research paper that discusses application of qualitative and quantitative content analysis techniques to adequately assess the quality of the Government Information Locator Service (GILS) metadata records (see 3.5). The main objective of this assessment was to examine whether GILS is helping agencies fulfil information dissemination and management responsibilities and the extent to which GILS is meeting users’ expectations [WEC98].

[42] XML Metadata Interchange Format, mainly a format for file storage of UML models, Specifies an open information interchange model giving developers the ability to leverage the web to exchange data between tools, applications, and repositories.

[43] School of Library and Information Sciences, University of North Texas Denton, TX 72603.

[44] Distinguished Professor, School of Information Studies, Syracuse University Syracuse, NY 13244.

A primary objective of this paper, titled, **"Assessing Metadata Quality: Findings and Methodological Considerations from an Evaluation of the U.S. Government Information Locator Service (GILS)"**, is to demonstrate the utility of metadata assessment for identifying systemic problems and for developing recommendations to improve record quality in support of networked information discovery and retrieval. In addition, it also aims to identify conceptual and methodological issues in metadata assessment that require additional research attention. The paper regards systematic methods for evaluating metadata to be intrinsic for system designers and implementers to refine metadata and improve their quality.

The paper identifies the fact that no consensus has been reached on conceptual and operational definitions of metadata quality. Therefore, for the qualitative content analysis for the metadata records, it defines a number of criteria: Access, Accuracy, Availability, Compactness, Comprehensiveness, Content, Consistency, Cost, Data structure, Ease of creation, Ease of use, Economy, Flexibility, Fitness for use, Informativeness, Quantity, Reliability, Standard, Timeliness, Transfer, Usability. These criterions were defined on the basis of traditional practices of bibliographic description, ongoing development of metadata schemes, and digital library initiatives.

In order to assess the metadata, the paper suggests two levels of quality assessments for metadata records. First level conducts to determine compliance (the extent to which records are free from errors, complete, current, etc.) based on provided documented requirements for metadata composition. The second level involves assessing the outcome of metadata records for utility and appropriateness of elements in terms of whether they support the purpose and goals of the metadata scheme.

As stated in this publication, the first of the two aforementioned metadata quality assessment levels was used to assess the quality of the GILS metadata records. The entire assessment procedure was divided in two phases. Both of these phases involved examination of about 80 GILS metadata records and comparing them against four assessment criterions, such as accuracy, completeness, profile and serviceability.

According to the paper, to ensure the accuracy of the metadata records, the number of "visible" errors in each record (e.g., spelling or typographical errors, file formatting errors, or incorrect date formats) was counted. The importance of accuracy was confirmed by another component in the overall evaluation study, a scripted online user assessment, which revealed users' poor tolerance of formatting errors. Besides these, the fullness of sampled records was also addressed in terms of inclusion of elements in the record.

In general, the study described in this research report, focused mainly on an assessment of metadata records in terms of their alignment with GILS standards and record creation guidelines. On the basis of the understanding of the article, it is reasonable to say that the procedures for metadata assessment described was successful in indicating major flaws in the GILS metadata records (e.g. ambiguity of element semantics and uneven levels of description.), mainly due to an uneven understanding or appreciation among GILS implementers of the value of metadata to support a distributed information service.

As this report claims, the metadata quality assessment procedures used may be modifiable to support more comprehensive exercise. Nonetheless, the procedures were very time-consuming and labour-intensive in terms of the development of criteria, their operationalization, actual examination, coding, and data entry for thousands of instances of metadata. However, as suggested in the paper, machine processing (e.g., for element counts, incidence of hypertext, etc.) may be used to reduce these drawbacks. Therefore, these procedures might prove to be useful for metadata quality assurance in the context of long-term metadata management.

Aside from the research work mentioned above, the research for this Project located much work that has been done within the learning technology community to assure metadata quality, focused on the development of metadata standards, specifications and vocabularies, and their implementation within repositories. One such work, a research paper by Sarah Currier[45], Jane Barton[46], Rónán O'Beirne[47] and Ben Ryan[48] attempts to address the issues of metadata quality assurance from the perspective of metadata creation process; an issue that has been largely overlooked within the learning technology community thus far [SJR03].

In general, the research paper, titled, **"Quality Assurance for Digital Learning Object Repositories: Issues for the Metadata Creation Process",** only investigates the creation of metadata necessary for resource discovery via searching and browsing within digital learning object repositories. The paper emphasizes on metadata quality assurance, presenting three cases of UK repositories whose experiences have raised issues for debate and further investigation. Although the emphasis is not directly on the metadata quality issues in the context of long-term metadata management, information provided in the paper may prove to be useful from a wider perspective of data curation (see Appendix A).

As mentioned above, the paper presents outcomes of the survey, conducted on three UK repositories, for metadata quality related issues. These three repositories are the Scottish electronic Staff Development Library (SeSDL); The Bolton Woods Local History Project and the Higher Level Skills for Industry Repository (HLSI). Below outlines a number of areas, highlighted from this survey, where quality of the metadata may impact on the discovery of resources in this economy:

1. **Error Management**: The HLSI case study identified that the issue of errors was significant in their repository with a large number of records. This is mainly due to the creation of metadata by untrained resource authors; therefore the paper emphasizes on the necessity for checking of metadata irrespective of its creator(s).

2. **Authors' and Other Contributors' Names:** Many libraries, archives and museums offer management of authors' names by using centralized name authority records, in order to ensure the search efficiency of metadata in case the names do change (e.g. after marriage etc.). The paper regards this as time consuming and costly exercise and recognizes the lack of a viable solution for this problem within learning object repositories.

[45] Centre for Academic Practice, University of Strathclyde, 50 George Street, Glasgow G1 1QE sarah.currier@strath.ac.uk

[46] Centre for Digital Library Research, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH jane.barton@strath.ac.uk

[47] City of Bradford Libraries, Archives and Information Service, Central Library, Prince's Way, Bradford BD1 1NN ronan.obeirne@bradford.gov.uk

[48] High Level Skills for Industry, Learning and Teaching Innovation Unit, University Of Huddersfield, Queensgate, Huddersfield, HD1 3DH b.ryan@hud.ac.uk

3. **Subject Area:** The paper identifies this to be the most complex areas of both metadata creation and resource discovery as all three case studies showed significant problems when untrained resource authors attempted to create subject metadata. Therefore, the major issues in this area mainly pertain to maximum resource discoverability by a heterogeneous population of searchers and confusion in deciding upon the creator of the subject metadata. While a resource author may know their subject area and its terminology well, a metadata specialist may know the specific area less well. However, a metadata specialist may be better placed to step back and think about all the potential users of a resource and about consistency of key words and classifications across a repository or network.

4. **Accessibility Metadata:** With the new SENDA (Special Educational Needs and Disability Act, 2001) legislation in the UK there has been some interesting recent work around developing metadata to describe the accessibility properties of a resource. However, the paper perceives this to be problematic for metadata creators who are not experts in accessibility.

In light of the aforementioned metadata quality related issues, the paper suggests the following three models for creating metadata while ensuring its quality: resource author or contributor only; metadata specialist only; and collaborative.

In the first model, the issues of metadata quality assurance are addressed in the design of metadata tools and user support and training. In addition, Metadata quality in all four of the above named specific metadata issues may be impacted by inadequate provision here. The second model requires the trained metadata specialist carrying out the task to eliminate any lack of knowledge about the pedagogical context, history or subject area of the resource, by in-depth research on the actual resource. Finally, the collaborative model may consist of a number of possible scenarios. One of these scenarios may be the metadata specialist having to check the data entered by the author of the resource, in certain fields for accuracy and conformance, and add other selected fields such as subject classification, keywords and accessibility information. This scenario, therefore, requires true collaboration between the metadata specialist and the author to ensure the metadata quality.

4.6 Approaches for Metadata Versioning

Although the research for this project has located a significant number of efforts that address the issues of metadata management and its quality assurance, only a limited number of those publications attempt to deal with the issues related to metadata versioning (see 2.4.4). One such attempt is a Whitepaper by Joe Futrelle[49] and Jeff Gaynor that outlines metadata versioning techniques as one of the capabilities of a distributed Metadata Service of Earthquake Engineering Simulation (NEES) Program of the National Science Foundation, USA, called NEESgrid[50] [JJG02]. The capabilities described, in the paper titled "**The NEESgrid Metadata Service API: Overview**", are in terms of the client API, which is an implementation of a generic object access API[51] that can be interfaced to arbitrary back-ends. In general, the service is designed to allow remote clients to browse, update, and otherwise manage metadata objects representing these entities of interest.

[49] National Center for Supercomputing Applications, Urbana-Champaign, IL 61820

[50] http://www.neesgrid.org

[51] Application Programming Interface.

The most interesting and relevant feature (to the main interest of this project) of the services of the NEESgrid is its metadata version management techniques. In generic terms, each metadata object within NEESgrid is represented in the repository and associated with versioning information that identifies the sequence and timing of the version. When an object is updated, the old version of the object is retained and the new version, with its time of creation (i.e. when the update was received by the Metadata Service), creator, and version number, is linked to it. Following the links between different versions, it is possible to obtain references to all versions of an object in order to find out which version existed at any given time. At any time, most recent version of any metadata object may be obtained.

Within NEESgrid updating a metadata object creates a new version of it with modified attributes. However, when updating a set of metadata objects with each of them referring to each other, it is required to perform the updates on all of them together. In order to prevent other clients from creating new versions of the objects, the users are enabled to place locks on the objects they are making changes to. When the client releases the lock, the objects will all appear to be modified simultaneously. If the client fails to release a lock (due to failure, for instance), the lock will expire, and the affected objects will revert to their previous versions. This consequently, helps prevent inconsistencies in metadata versions from appearing in the database. Furthermore, deleting an object in effect creates a special, final version of the object, which is marked as deleted. However, an object can be rolled back to an earlier version, which in effect creates a new version, which is in its attribute values identical to the earlier version.

In general, the metadata version management techniques outlined in this paper is quite an interesting approach to address such issues and might prove to be effective within a repository managing metadata over long time. However, these techniques will need to be subject to further examination for determining their effectiveness in terms of long-term metadata management.

Another effort for metadata versioning is a research paper, titled "**Metadata Efficiency in a Comprehensive Versioning File System"** by Craig A.N. Soules, Garth R. Goodson, John D. Strunk, Gregory and R. Ganger of Carnegie Mellon University; that evaluates mechanisms for encoding metadata versions more efficiently than conventional versioning systems by describing specifically two methods for storing metadata versions more compactly. These two methods are journal-based metadata and multiversion b-trees. The paper also describes the integration of these two space-efficient metadata structures for versioning file systems into the Comprehensive Versioning File System (CVFS) [CGJ02]. It is to be noted that Comprehensive Versioning implies retention of every version or every file and creation of a new version from every modification of a file.

Generally, Journal-based metadata records metadata changes in a journal by maintaining a full copy of the current version's metadata and a journal of each previous metadata change. Effectively, Journal-based metadata encodes each version of a file's metadata in a journal entry with each entry describing the difference between two versions, allowing the system to recreate old versions of the metadata by undoing each change in the metadata backward through the journal until the desired version is recreated. This process of undoing metadata changes is referred to as journal rollback in the paper. The figure 4.6 illustrates how journal-based metadata works.

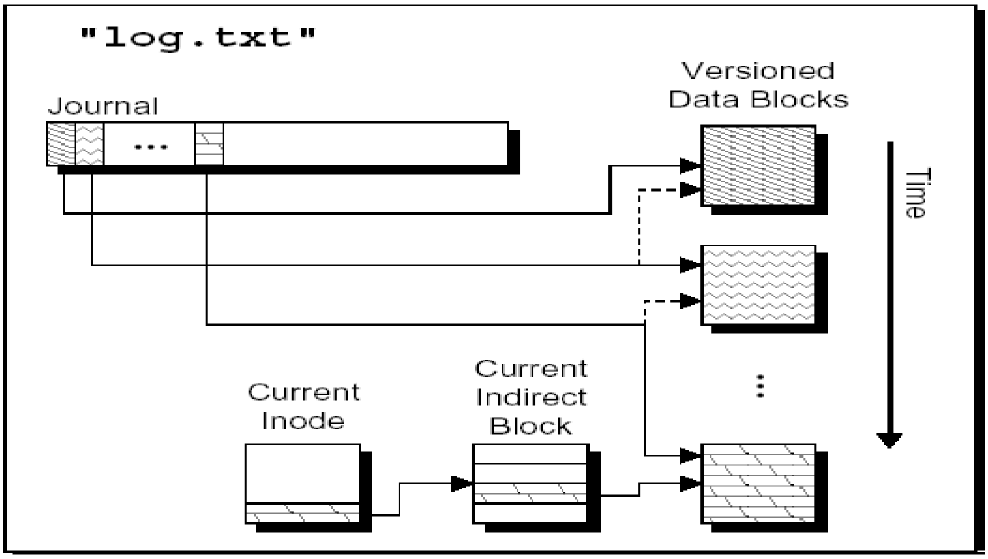


Figure 4.6: Journal-based metadata system [CGK02]

On the other hand, as a variation on standard b-trees, multiversion b-trees maintain all versions of a metadata structure within a single tree. Each entry in the tree contains unique user-defined key and is marked with timestamps indicating the time over which the entry is valid. Having unique keys means that entries within the tree are never overwritten; therefore, multiversion b-trees can have the same basic structure and operations as a standard b-tree. To facilitate current version lookups, entries are sorted first by the user-defined key and then by the timestamp. The figure 4.7 depicts an example of a multi-version b-tree.

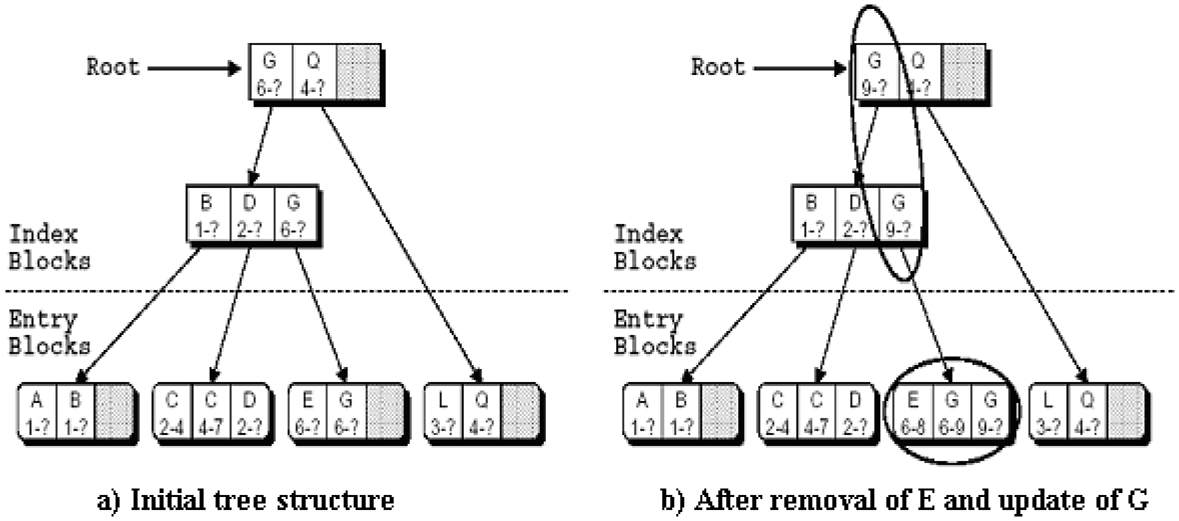


Figure 4.7: The layout of a multiversion b-tree [CGK02]

In essence, a multiversion b-tree keeps old versions of entries in the tree. As in a standard b-tree, an entry in a multiversion b-tree contains a key/data pair; however, the key consists of both a user-defined key and the time at which the entry was written. With the addition of this time-stamp, each key becomes unique.

According to the paper, the advantage of using these two metadata versioning solutions lies in the space utilization of versioning. In essence, both of these solutions are more space efficient than conventional versioning. As everything has both positive and negative sides, both of these two mechanisms, however, do have some drawbacks. For instance, journal-based technique incurs performance penalty for recreating old versions of the metadata with rollback process. One solution to this problem, as stated in the paper, is to checkpoint a full copy of a file's metadata to the disk occasionally.

In multiversion b-trees, on the contrary, accesses to old and current versions have the same performance since both current and history entries are stored in the same tree. Due to this reason, large numbers of history entries can decrease the performance of accessing current entries. Nevertheless, when assessing the effectiveness of these two mechanisms in a wider context of long-term metadata management, it is only reasonable to say that their suitability will largely depend on the file or storage system (i.e. Comprehensive Versioning or Other) used to store metadata for long time. To elaborate, as proven by the experiment in this paper, these two versioning mechanisms are space-efficient, i.e. capable of doubling the duration of time (generally limited by finite storage capacities) over which comprehensive versioning is possible, when they are integrated in to the CVFS file system. However, without any concrete information, it will not be wise to come to a conclusion in regard to their effectiveness in the context of any other file system than the CVFS.

Another publication that is worth mentioning is a combined effort [CJJ03] by Christopher Brooks, John Cooke, and Julita Vassileva of ARIES Laboratory, Computer Science Department University of Saskatchewa, Canada; that tried to deal with metadata versioning in education information domain. The paper, titled **"Versioning of Learning Objects"**, mainly introduces a metadata model with an aim to facilitate the maintenance of consistent version information about learning object[52]. In regard to the context of this paper, it should be noted that due to the distributed nature and highly mutable nature of learning objects, keeping consistent version information is an extremely difficult task.

Based on the current e-learning metadata specification, such LOM (see Appendix C) and Dublin Core (see 3.2), the metadata model for versioning described in the paper, specifically allows for agents to better reason about versioning changes between learning objects even if the vocabularies being used to describe the objects are not known. In addition, it allows learning object repositories to provide a higher level of versioning services (e.g. roll-backs, branching, etc). These are done by enabling the metadata model to map given changes in learning objects, captured as a set collection of syntactical operations, more generally to the metadata that describes a learning object.

In essence, the main objective of this model is to eliminate the inability of current metadata specifications to capture both syntactical and the semantic changes that occur when learning objects are versioned. The underlying concepts of this metadata model for versioning may well be used to address the versioning requirements of long-term metadata management.

[52] Learning objects are reusable pieces of educational material that are intended to be strung together to form larger educational units such as activities, lessons, or whole courses.

4.7 Long-term Metadata Preservation

The extensive literature survey performed for this MSc. Project came across a significant number of publications that address the issues associated with long-term preservation of digital information to different extent. Oddly enough, most of these publications (e.g. research papers, articles etc.) tend to focus only on the metadata requirements for preserving information for long time, completely overlooking the fact that these metadata also need to be preserved along with the resources that they are describing, in order to ensure the longevity of these resources.

One such publication is a research report, titled "**Metadata for Long Term Preservation**", written by Catherine Lupovici and Julien Masanès of NEDLIB (Networked European Deposit Library) Consortium. This document defines the core minimum metadata that are mandatory for long-term data preservation within NEDLIB's Deposit System for Electronic Publications (DSEP), in order to handle large amounts of data items in a changing technological environment. The report also describes the main concepts behind the development of the DSEP, which are largely based on the OAIS reference model (see 3.9) with the main focus being on the storage and preservation functions. This report does touch upon the metadata management issues associated with data preservation within the specific problem domain of the DSEP [CLJ00]. However, the discussion in that regard, mainly attempts to shed light on the locations of the metadata rather than providing definite and useful solution for metadata management. As stated in that report, the best possible solution for archiving purposes is to use both of the two possible locations for metadata (see 2.4.5) by duplicating metadata from the archive's item to more practical databases.

However, one exception in the aforementioned trend of publication addressing long-term data preservation is another research-oriented publication, written by Jeff Rothenberg in 2000. This report, titled, **"An Experiment in Using Emulation to Preserve Digital Publication"** presents the results of a small study[53], which was intended to test and evaluate the feasibility of using emulation as a means of preserving digital publications in accessible, authentic, and usable form within a deposit library. The report presents the analysis and synthesis required for that study in the context of the increasingly accepted Open Archival Information System (OAIS) as well as the NEDLIB adaptation of the OAIS, the Deposit System for Electronic Publications (DSEP) [EJR00].

In term of the relevance for the context of this MSc project, the information provided in the aforementioned report may be divided in two different segments: advantages & disadvantages for existing metadata management and preservation approaches of the DSEP & recommendations for using emulation approach for metadata preservation in the DSEP.

[53] Undertaken by RAND-Europe for the National Library of the Netherlands (the Koninklijke Bibliotheek, or "KB") in connection with their work on the NEDLIB (Networked European Deposit Library) effort jointly funded by the European Commission's Telematics for Libraries Programme.

To begin with, in the first segment, this report briefly discusses how the aspects of metadata management are addressed in the DSEP. In short, the DESP separates from the AIP[54] all metadata elements except those that are considered part of the original publication. These separated metadata are to be preserved in its administrative computer system (i.e. the metadata store) in some convenient form, such as in a database or document management system, which is not intended to be a long-lived preservation format. This allows the Data Management (figure 4.8) process to "own" and control all other metadata elements (such as those describing a document's current format, condition, location, usage, etc.). In particular, such information can be updated on a frequent basis and migrated to new data management software or representations as necessary, without having to modify the preserved publication itself and incurring the risk of corrupting it or the remainder of the AIP in the process.

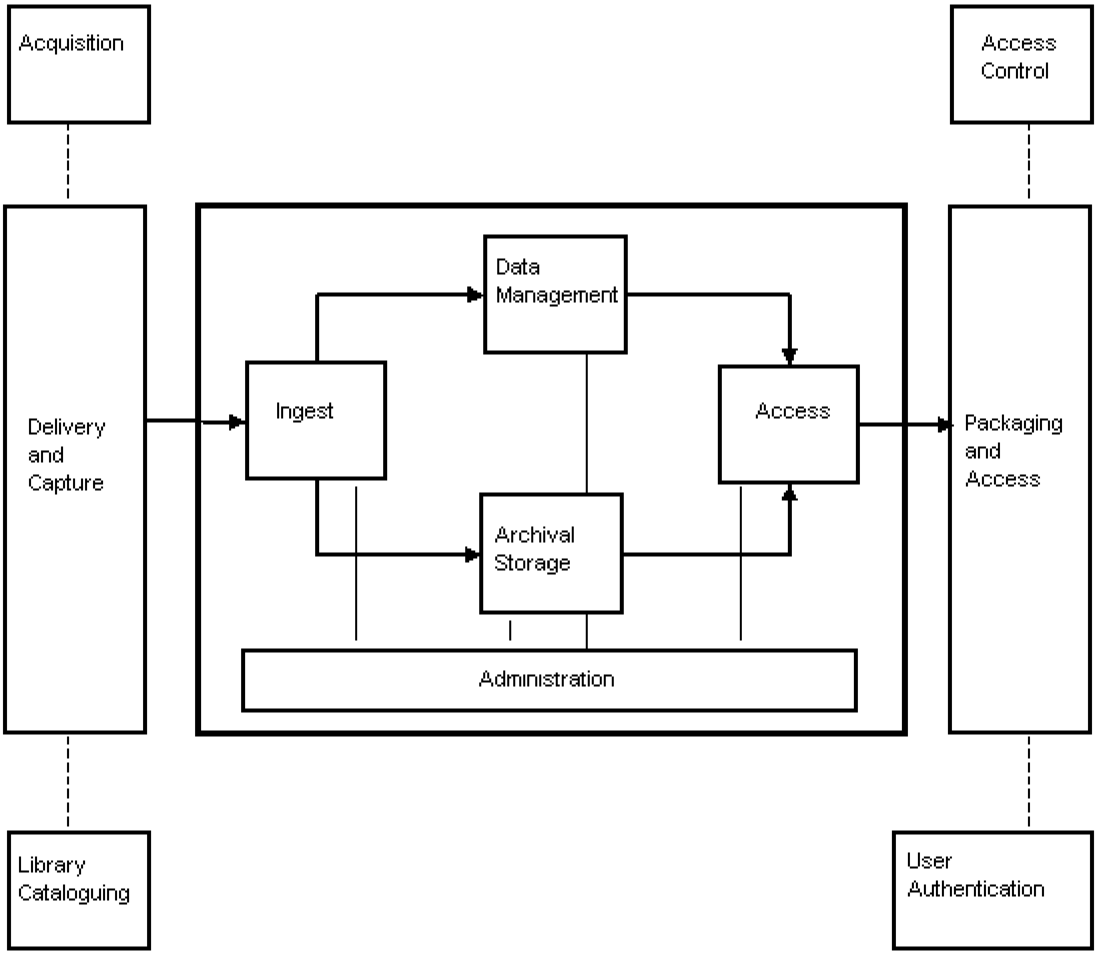


Figure 4.8: OAIS functional entities scoped to DSEP processes [NDO00]

As mentioned in the report, there are two sides to this above approach. The positive side is that it eliminates the problem with including metadata in AIPs, of course, - tendency of the content of AIPs becoming unintelligible over time, as representations change. The report regards this problem as the crux of the preservation problem, which the emulation schemes described here are designed to address. One alternative solution to this problem is subjecting

[54] **Archival Information Package (AIP)**: An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS (see 3.9).

metadata to the same preservation procedures, such as emulation or migration, as the documents they are associated with. Yet it seems unnecessarily cumbersome to use emulation just to access metadata about a preserved document, whereas using migration to keep metadata comprehensible reintroduces the danger of corruption, which emulation was intended to remove. Therefore, the report suggests that AIPs should ideally be considered immutable, to protect them from inadvertent corruption.

The disadvantage of separating metadata from AIPs is that important information about an AIP may be lost if access to the metadata store is compromised. AIPs, being designed for long-term preservation, will presumably be subject to rigorous access controls and backup procedures, which will be all the more robust if AIPs are kept immutable. The metadata store, on the other hand, being part of an administrative system that is in daily use, may be far more vulnerable to misuse (whether inadvertent or intentional) and cannot be kept immutable, since it must be constantly updated. Therefore, to the extent that the DSEP relies on information in the metadata store that is not replicated in the AIPs themselves, it may be vulnerable to loss. One classic solution, as described by the author of this report, is to treat this volatile information as a "cache" that can be corrupted or discarded without serious loss as it can always be rebuilt from information stored in the persistent entities themselves.

Now, in the second segment, the report offers the following recommendations for the DSEP system for applying the emulation approach to both data and metadata preservation:

- The data management entity of the DSEP system needs to ensure that any required explanatory metadata associated with digital documents remains understandable. This may require converting at least the topmost level of such explanations (sufficient to explain how to read lower levels of explanation) into successive "explanation formats" as previous such formats become obsolete and vice versa.

- The data management entity should also have necessary functionality to maintain linkages between metadata references to named emulator specification languages, emulator specification interpreters, emulator specifications, and emulation virtual machine specifications and implementations.

- The DSEP should offer publishers a remotely accessible online "Validation Service" that would allow them to validate their SIPs[55] themselves before submitting them. This might subsume facilities for obtaining (and including in the SIP) the necessary standard identifiers for the relevant software versions and platform configurations, and it might offer an emulation testing capability that would allow publishers to verify that their publications will be usable under emulation before they submit them. As the report perceives, making such a facility available to publishers could greatly reduce the validation effort required during Ingest (see 3.9) while simultaneously ensuring that submitted publications will be properly preserved.

[55] **Submission Information Package (SIP)**: An Information Package that is delivered by the Producer to the system for use in the construction of one or more AIPs.

- It is necessary to continue to refine the metadata required to describe digital documents and to link them to the software and emulated hardware environments required to render them in the future. Moreover, it is necessary to ensure that these descriptions can themselves be maintained in human-readable form indefinitely.

It is to be noted that a summary of the emulation approach described in the above mentioned report has not been provided, as it does not hold sufficient relevance to the main subject of this MSc. project. Besides, it would not have been possible to summarize such lengthy description of the approach within the limited scope of this thesis. However, detailed information in regard to this approach may be acquired from the reading of that report.

As it was not possible to include all reviewed efforts (i.e. published works) for metadata management and its different aspects within such limited scope of this thesis, summary of other studied research materials has been provided in Appendix E.

Chapter 5

Assessment of Existing Metadata Management Systems

This chapter presents the results of the assessment and evaluation conducted on the functionalities and features offered by a range of different existing systems (or tools) that pertain to the main principles and issues of (long-term) metadata management (see 2.4). These systems were systematically evaluated on the basis of a number of criteria, such as, metadata creation/updates facility, search facility, metadata quality assurance, versioning, preservation technique, security, platform independence etc. of the tools in terms of how their efficiency for long-term metadata management may be perceived in the context of digital curation.

5.1 MetaStar Digital Library

The MetaStar Digital Library (MetaStar DL) is one of the metadata related products, developed and distributed by Blue Angel Technologies, a web-based software solution provider. In general, the MetaStar DL, which may well be regarded as a complete solution, allows libraries to create, capture, describe, publish and discover all types of digital objects, including images, video, audio, traditional documents such as PDF's and MS Office, and Web content from throughout the library, library consortia, or institution.

5.1.1 MetaStar DL Assessed [MDB03]

The outcomes of the assessment (in the context of data curation) performed on the MetaStar DL are detailed as follows:

a) Metadata Creation/ Updates/ Editing

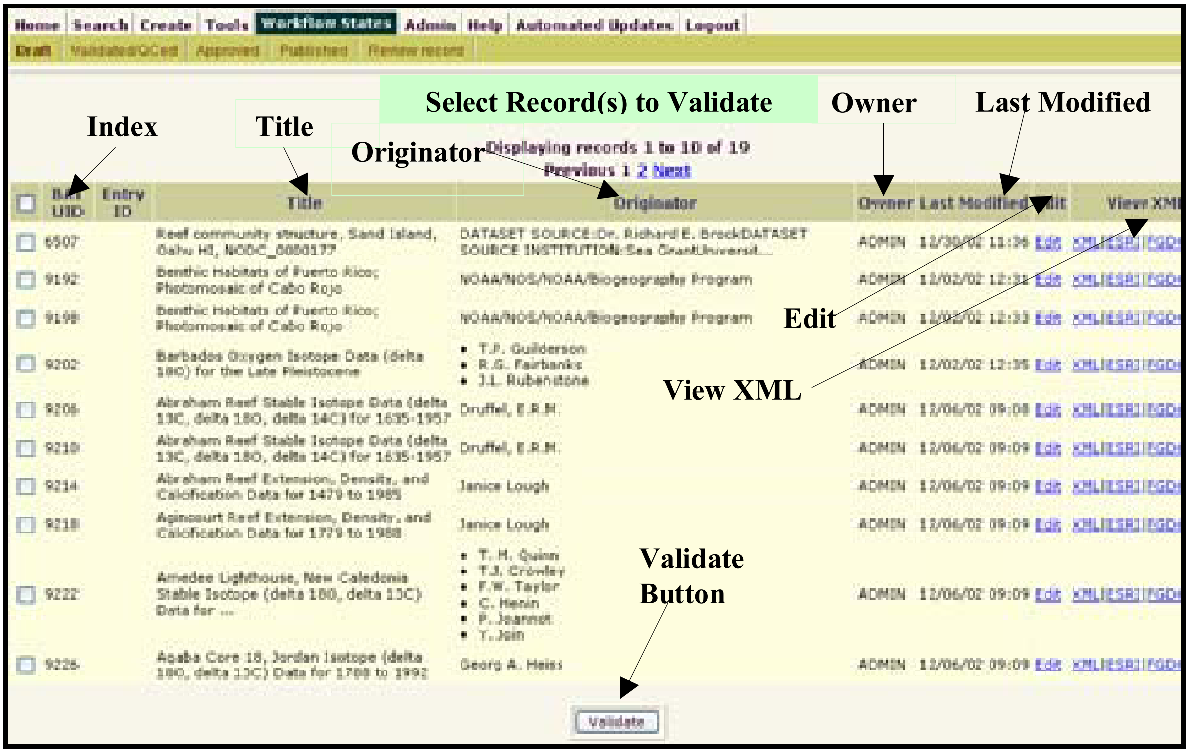
MetaStar DL supports Web-based access to the repository for adding and editing metadata about each digital object or an entire collection of objects. In addition, it enables automatic metadata generation for information such as file type, size, and date created. The tool also provides interoperability between metadata conforming to different metadata standards by supporting a range of metadata standard including Dublin Core, MARC, EAD, TEI, GILS, FGDC (see chapter 3) etc. as well as any customized metadata standard. This is quite a useful feature from the perspective long-term metadata management, which may well involve maintaining metadata originated from different sources and conforming to different metadata formats.

b) Search Facility

MetaStar DL provides 100% configurable search and retrieval interface with the ability to deploy multiple search screens (Simple, Advanced etc.). The system also facilitates searching metadata in multiple languages. Furthermore, MetaStar DLS is fully Z39.50[56] compliant at both the client and server level, thus the search interface of MetaStar DLS can easily search other Z39.50 compliant repositories. Besides, any Z39.50 client can query the underlying digital repository of the MetaStar DLS as well.

c) Metadata Quality Assurance

In order to ensure accuracy and consistency of metadata during its creation, the system provides data dictionaries and controlled vocabularies for the users. In addition, the metadata created undergoes a validation process to further ensure its quality before being stored and becoming searchable. However, the conclusion reached from very limited information that was available to clarify how this validation process is performed; indicates that it is definitely a human operated procedure and the creator of the metadata is responsible (figure 5.1) for accuracy of the validation. Therefore, bearing in mind the importance of good quality metadata in the context of long-term data preservation, this validation process may not be quite the feature that a system managing metadata over long periods of time should offer.



Displaying records 1 to 10 of 19
Previous 1 2 Next

□	[illegible] UID	Entry ID	Title	Originator	Owner	Last Modified	Edit	View XML
□	6507		Reef community structure, Sand Island, Oahu HI, NODC_0000177	DATASET SOURCE:Dr. Richard E. BrockDATASET SOURCE INSTITUTION:Sea GrantUniversit...	ADMIN	12/30/02 11:36	Edit	XML|ESRI|FGD
□	9192		Benthic Habitats of Puerto Rico; Photomosaic of Cabo Rojo	NOAA/NOS/NOAA/Biogeography Program	ADMIN	12/02/02 12:31	Edit	XML|ESRI|FGD
□	9198		Benthic Habitats of Puerto Rico; Photomosaic of Cabo Rojo	NOAA/NOS/NOAA/Biogeography Program	ADMIN	12/02/02 12:33	Edit	XML|ESRI|FGD
□	9202		Barbados Oxygen Isotope Data (delta 18O) for the Late Pleistocene	• T.P. Guilderson • R.G. Fairbanks • J.L. Rubenstone	ADMIN	12/02/02 12:35	Edit	XML|ESRI|FGD
□	9206		Abraham Reef Stable Isotope Data (delta 13C, delta 18O, delta 14C) for 1635-1957	Druffel, E.R.M.	ADMIN	12/06/02 09:08	Edit	XML|ESRI|FGD
□	9210		Abraham Reef Stable Isotope Data (delta 13C, delta 18O, delta 14C) for 1635-1957	Druffel, E.R.M.	ADMIN	12/06/02 09:09	Edit	XML|ESRI|FGD
□	9214		Abraham Reef Extension, Density, and Calcification Data for 1479 to 1985	Janice Lough	ADMIN	12/06/02 09:09	Edit	XML|ESRI|FGD
□	9218		Agincourt Reef Extension, Density, and Calcification Data for 1779 to 1988	Janice Lough	ADMIN	12/06/02 09:09	Edit	XML|ESRI|FGD
□	9222		Amedee Lighthouse, New Caledonia Stable Isotope (delta 18O, delta 13C) Data for ...	• T. H. Quinn • T.J. Crowley • F.W. Taylor • C. Henin • P. Joannot • Y. Join	ADMIN	12/06/02 09:09	Edit	XML|ESRI|FGD
□	9226		Aqaba Core 18, Jordan Isotope (delta 18O, delta 13C) Data for 1788 to 1992	Georg A. Heiss	ADMIN	12/06/02 09:09	Edit	XML|ESRI|FGD

Figure 5.1: Metadata Validation Interface of MetaStar DL [MDB03]

[56] Z39.50 is an American National Standard that specifies an open network application protocol for information retrieval, which enables interoperability between disparate information systems over a heterogeneous network (National Information Standards Organization, 1995).

d) Metadata Versioning

Although the system does not appear to address the issues associated with metadata versioning directly, its ability to support parent child (hierarchical) relationships between metadata records may possibly serve as the foundation for devising efficient and intelligent metadata versioning techniques.

e) Metadata Preservation

The MetaStar DL stores both data and metadata in its underlying repository. However, although the repository is Z39.50 compliant, thus facilitates increased search-ability for the stored metadata, it is not intended for perpetuation of metadata to ensure longevity of the actual resources.

f) Security/Access Control

The MetaStar DL facilitates user and user group management to control/restrict access to metadata records for editing purposes. This, in theory, should prevent unauthorized access, consequently possible modification to metadata - a core requirement for effective data curation. The security is generally ensured by the system administrator with the help of the administrator module associated with the MetaStar DL.

g) Platform Independence

The MetaStar DL is compatible with any java enabled environment, thus making it platform independent, which is a sought after feature from the viewpoint of users of any commercial software. In addition, the system supports most commonly used web servers (IIS, apache, Netscape) and databases (Oracle, MS SQL server).

5.1.2 Concluding Remarks

Aside from a weak quality assurance procedure and lack of long-term metadata preservation facility, the MetaStarDLS is an easy to use, easy to administer and easy to afford tool that possesses almost all sought after features to provide a standards-based, scalable and extensible solution for long-term metadata management.

5.2 MetaMatrix MetaBase™

The MetaBase™, developed by the MetaMatrix - a leading provider of Enterprise Metadata Management and Enterprise Information Integration solutions based in New York, USA, is a Metadata management system designed to meet the needs of the enterprise, providing both departmental and enterprise-wide Metadata management. It serves as a powerful Metadata repository, allowing groups to streamline application development [MEM04].

To provide a brief overview of the system architecture, the MetaBase™ metadata management system is comprised of the MetaBase™ Modeler, the MetaMatrix Console, the MetaBase™ Server, the MetaBase™ Repository the MetaViewer and the SearchBase or MetaBase™ Reporter [MIM04].

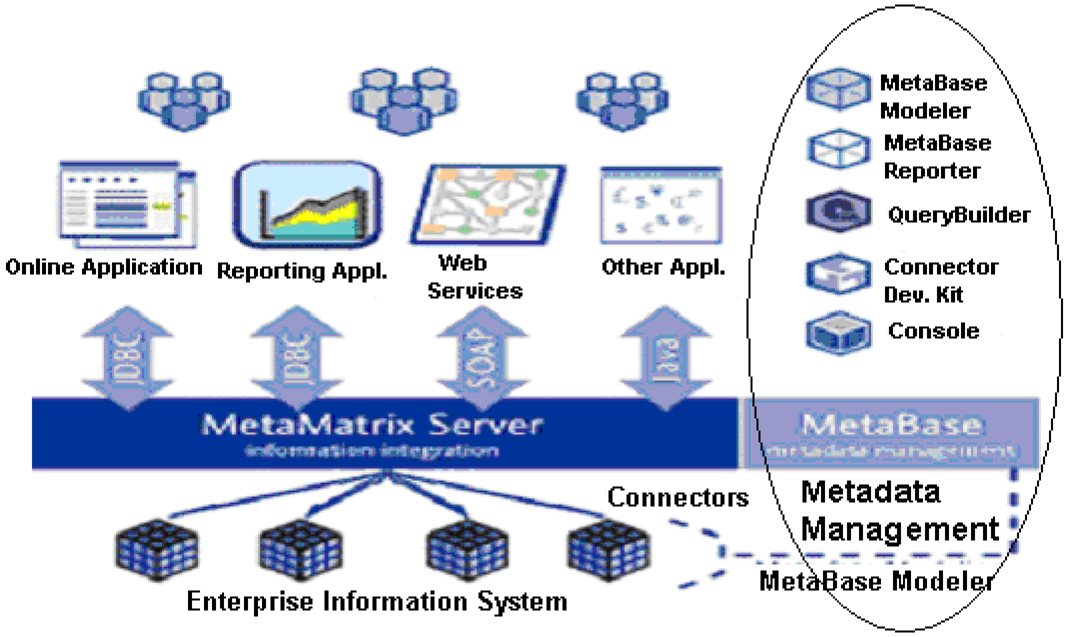


Figure 5.2: Architectural Overview of the MetaBase™ [MIM04]

5.2.1 MetaBase™ Assessed

The most significant outcomes of the evaluation performed (in the context of long-term data curation) on the MetaBase™ are presented in details below:

a) Metadata Creation/Updates/Editing [MEM04]

The MetaBase™ and its metadata management tools import and facilitate the creation of metadata from enterprise information systems. The MetaBase™ Modeler, an UML-based graphical modelling tool, is used to import or create metadata and to then build a metamodel. The MetaBase™ Modeler also eases the creation of metadata for sources that do not explicitly expose metadata such as text files and legacy systems. All metadata are stored into the MetaBase™'s standard and scalable metadata repository.

Furthermore, the MetaBase™ Modeler enables users to define relationships between disparate information sources. Using this component of the MetaBase™, data modellers can create virtual metadata models of physical data sources to join, transform, and otherwise relate disparate information sources. Also, using the XMI specification, MetaBase™ can exchange metadata models with common modelling tools as well as importing metadata from databases. This certainly ensures interoperability between metadata originated from different sources.

One potential drawback may be the MetaBase™'s inability to support no other metadata standards than four Object Management Group's (OMG) metadata standards, such as MOF, CWM, XMI and UML (see Appendix C). Therefore, this may be seen as a deficiency for a metadata management system, which should preferably facilitate maintenance of metadata irrespective of their standards over long time.

b) Search Facility

The MetaBase™ employs MetaViewer, a Web browser-based tool in order to facilitate browsing and searching metadata models, providing access to details for all entities of the published metadata models. In addition, through the MetaViewer, the SearchBase or MetaBase™ Reporter enables users and developers to search the metadata descriptions in the MetaBase™ Repository to find and analyse the data assets they need across multiple departmentally or geographically dispersed repositories [MCMNA].

c) Metadata Quality Assurance

Only metadata quality assurance procedure that the MetaBase™ appears to provide is Metamodel validation and reconciliation through the MetaBase™ Modeller [MIM04]. This only ensures that any metamodel being created conforms to set standard(s) or rules. Although this may partly perform the task of metadata quality assurance, the tool does not appear to have any automated procedure to validate the metadata semantically as well as syntactically. Furthermore, the tool does not provide any controlled keywords or vocabulary to prevent any inadvertent error during metadata creation, hence, may not ensure the proper and accurate description of the data source.

d) Metadata Versioning

According to [MCMNA], the MetaBase™ repository is capable of versioning metamodel and these versions can be controlled and managed through the MetaBase™ modeller. However, as neither [MCMNA] nor the World Wide Web provides sufficient (and relevant) information to explain how these tasks are performed, several attempts were made to contact MetaMatrix personnel seeking answers to these questions. Unfortunately, no response from the MetaMatrix has been received till date. Therefore, it was not possible to assess the efficiency of the version control technique(s) employed by the MetaBase™ in the context of long-term metadata management.

e) Metadata Preservation

As per the information acquired on the MetaBase™ repository, it does not employ any special technique (e.g. emulation, migration etc.) for preserving the stored metadata for long periods of time. Therefore, these metadata may face corruption and be lost over long-term, resulting in probable extinction of the actual resources.

f) Security

The MetaBase™ Repository is a secure meta-database providing single sign-on security, and a single access point to the disparity of information sources to prevent unauthorised access to stored metadata [MEM04].

g) Platform Independence

The MetaBase™ is a platform independent tool that supports all major operating systems and databases.

5.2.2 Concluding Remarks

Notwithstanding, the lack of suitable long-term metadata preservation facility, the MetaMatrix MetaBase™ provides some potentially desirable features such as, metadata interoperability, security, intelligent search facility etc. However, on the basis of the limited information acquired in reference to its metadata quality assurance and version control facilities (two of the most critical issues associated with metadata management, see 2.4), it may not be sensible to reach a conclusion in regard to its effectiveness or usefulness for the purposes of long-term metadata management.

5.3 Spatial Metadata Management System (SMMS™) Version 5.1

The SMMS™ version 5.1 was developed by the Intergraph, a leading provider of products and services, open technology and data integration, partners and people to help customers implement successful geo-spatial information-based solutions. These solutions can be deployed on the desktop, the Web, or with mobile technology. In short, the SMMS™ can be best described as the industry-leading desktop metadata management system, including auto-capture and viewing of spatial data in popular data formats. [INT03, GEO98]

5.3.1 The SMMS™ Assessed

This sub-section details the results of the assessment and evaluation performed on the SMMS™ in order to determine its suitability in the context of long-term metadata management [ISM03, INT03].

a) Metadata Creation/Updates/Editing

In general, SMMS™ makes the metadata maintenance process as efficient as possible by implementing a relational data model that makes many common “sections” of information (contact information, citations, etc.) re-usable. It allows users to create, edit, view and publish only CSDGM (see 3.3) standardized spatial metadata. Therefore, it may not be compatible with metadata with any other standard, such as DIF or a customised standard etc. Nevertheless, the SMMS™ utilizes GeoMedia[57] technology to automatically extract from various spatial data formats certain metadata content elements which are inherent in the data: bounding box, attribute names, etc.

In addition, any CSDGM-compliant metadata record can be loaded from ASCII text, CSDGM-standard XML, or SGML format into the SMMS™. The tool also facilitates sharing SMMS metadata records with other metadata software users by exporting the CSDGM-standard interoperable metadata in ASCII text, XML, or SGML format. Besides, in order to reduce the overall creation time, the tool provides Metadata template, which is an efficient way of using metadata that is already stored in the SMMS™ to create new records. The figure 5.3 reflects the workflow within the SMMS 5.1.

[57] GeoMedia is a member of the SMMS family; that provides all the features of SMMS, plus it integrates into GeoMedia workflows. GeoMedia SMMS allows users to query their enterprise metadata catalogue, and automatically load the data they need right into their GeoMedia map window [INT03].

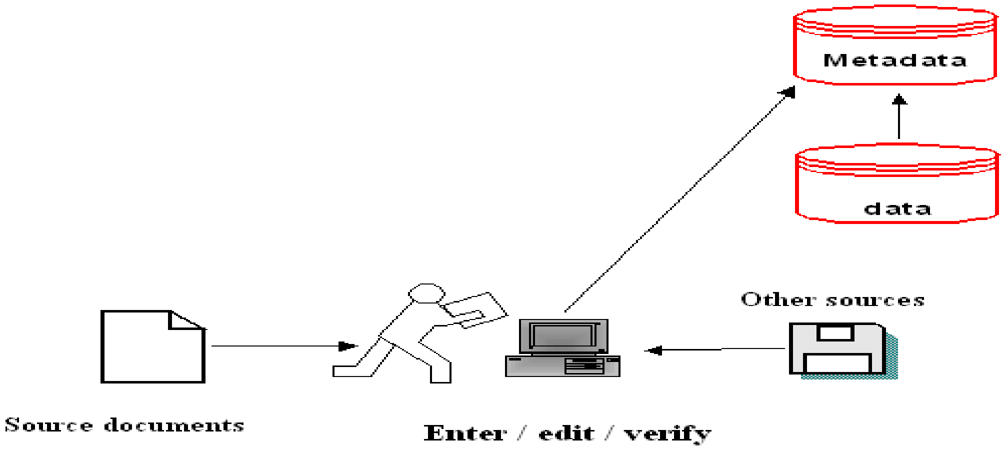


Figure 5.3: Workflow within the SMMS™ 5.1[INT03]

b) Search Facility

The SMMS™ has a meta-database, named the MetaGate Data Catalog that enables users to view metadata side by side with the GIS data layer it describes, open SMMS records using a Windows™-style tree view, and quickly locate metadata records using a powerful yet simple search interface. In addition, necessary facility is provided to enable users to search for metadata by using keywords, attributes, time-period of the data layer, and extent of the data layer.

c) Metadata Quality Assurance

With an aim to make up for its incompatibility with other metadata standards than the CSDGM, the SMMS™ provides users with the ability to create custom metadata profiles and thus ensures accurate and adequate documentation for information resources. In addition, the tool contains a number of features that aid in assuring metadata quality during its creation. For example, it provides context-sensitive help so that the users can quickly retrieve the CSDGM definition for each metadata field along with sample metadata representation. It also utilizes many pull-down lists of keywords, so as to make metadata content authorship easy and relatively error-free. However, what it lacks is an automated (or manual) procedure for checking both syntax and semantics of newly created or captured metadata before storing them into the repository.

d) Metadata Versioning

The SMMS™ does not address the issues of metadata versioning instead it relies on the versioning mechanism (see 3.3.1) of CSDGM metadata standard. As it has been mentioned in chapter 3 the CSDGM content standard allows users to indicate versions of spatial datasets and versioning of metadata associated with the datasets. This may well be viewed as a potential shortcoming from the perspective of long-term metadata management as a system dealing with such complex and critical issues should ideally incorporate automated and intelligent metadata version management procedures.

e) Metadata Preservation

In simple terms, the MetaGate Data Catalog, the metadata-database associated with the SMMS™ has no incorporated preservation technique/mechanism in order to ensure the longevity of the stored metadata with its quality intact. In other words, the SMMS™ offers no relevant facility to prevent metadata from becoming obsolete (or even corrupted) over the long periods of time with rapid evolvements of newer technology.

f) Security

The SMMS™ provides login facility to prevent unauthorised access to its metadata-database.

g) Platform Independence

The SMMS™ is compatible only with all major windows operating systems (not UNIX, Linux etc.) and a few select commonly used databases (Access, Oracle, or MS-SQL – not DB2, Sybase etc.).

5.3.2 Concluding Remarks

As a whole, the SMMS™ can be best described as a user-friendly, forms-based GUI for authoring CSDGM-standard spatial metadata, based on a relational data model (implemented in Access, Oracle, or MS-SQL.). However, although the tool appears to have efficient metadata creation/updates/editing facility, due to its incompatibility with other metadata standards than the CSDGM as well as lack of effective metadata quality assurance, versioning procedure and finally long-term preservation technique, it may not be considered as a prospective candidate for serving the purposes of the long-term metadata management.

5.4 The GCMD Metadata Management System

The Metadata Management System (MMS) developed by the NASA's Global Change Master Directory (GCMD), employs an interesting approach for maintaining the integrity and consistency of the records to ensure the quality and search efficiency of the 9800 metadata records contributed to date with additional metadata created and modified on a daily basis.

5.4.1 GCMD MMS Assessed [JCD01]

The assessment results of the GCMD metadata management system are provided as follows:

a) Metadata Creation/Updates/Editing

The system enables metadata creation, updates, editing etc. within an Oracle Relational database system. In order to ensure the quality of metadata during its creation, updates or editing, hence ensure overall data integrity, the system provides a list of controlled vocabulary or keyword, maintained by the GCMD, for use with fields found within the metadata records created using DIF (Directory Interchange Format) specification. With all maintenance activities and updating being done using this database, the system is capable of allowing massive updates to be made in a single transaction, with the aim to minimize maintenance. If this operation is performed on a long-term basis, it might lead to potential inconsistency, hence flaw in the actual quality of the metadata. Besides, the assessment failed to acquire any explicit indication as to whether the system allows creation in standards other than DIF. However, in order to handle metadata from disparate sources, the system makes use of relevant features of XML and XSLT.[58]

b) Search Facility

The system provides efficient search facility for the users using terms included in controlled vocabularies, which results in the increased retrieval of relevant documents. It also provides a search interface with an intelligent layout of the hierarchical science parameter keyword list, with each keyword being a link.

c) Metadata Quality Assurance

In order to ensure the quality of the metadata records, the system performs validation checks not only on the syntax of records, but the semantics as well. In essence, in order to test for semantically valid metadata the system utilizes several methods in conjunction with one another. Thus, the system ensures incorporation of only quality metadata into the database while eliminating the drawback of currently available validation techniques, which mainly focus on the syntax, leaving the semantics of the data untested

However, the basic method for metadata validation employed by the system involves the GCMD staff taking great efforts to ensure that information contained within the records are current and accurate. From the perspective of long-term metadata management, this sort of continual maintenance may be deemed costly in terms of time and resources.

In addition, the system validates the metadata contents for syntax as well as semantics by requiring the use of standard formats for spatial (e.g. whole degrees longitude or latitude) and temporal (i.e. YYYY-MM-DD) coverages, with only those values formatted in a specific manner passing internal validation. This subsequently, allows additional validation checks into appropriate software to ensure the semantics are correct. For example, with the temporal coverage, the software checks that the "stop date" is indeed later in time than the "start date" and warns the user if this is not the case. This enables the discovery and resolution of errors in the data set description prior to the document being committed to the system.

For further validation checks on the syntax of metadata, written in XML, the system employs a tool Xerces that contains advanced parser functionality for DOM[59] (Document Object Model), SAX[60] (Simple API for XML), and XML Schema. The system uses both the SAX and DOM APIs in its application to parse documents and traverse trees in processing of

[58] The Extensible Stylesheet Language Transformations is a templating markup language used to express how a processor creates a transformed result from an instance of XML information. In other words, XSL Transformations (XSLT) is a language for transforming XML documents into other XML documents. XSLT is designed for use as part of XSL, which is a stylesheet language for XML. Besides, XSLT is also designed to be used independently of XSL. However, XSLT is not intended as a completely general-purpose XML transformation language. Rather it is designed primarily for the kinds of transformations that are needed when XSLT is used as part of XSL.

[59] DOM is a platform- and language-neutral interface, that provides a standard model of how the objects in an XML object are put together, and a standard interface for accessing and manipulating these objects and their inter-relationships.

[60] An event-driven interface created specifically for XML parsers that are written in object-oriented applications, such as Java.

documents. In addition, the system entails java implementation of Xalan tool to convert metadata in original DIF format to any other standard formats, e.g. CSDGM.

Aside from the aforementioned techniques for validating metadata, the system employs another Master Directory software to perform the GCMD specific validation on the metadata. The validation process includes checks of the controlled keywords, personnel, spatial, and temporal coverage. The software systematically steps through the document, examining the contents of each one of the fields as it is encountered and extracts the contents of the field as well as comparing this against all valid or approved formats/structures of that type currently listed in the database. If the contents format of a field in the document cannot be located in the database, the user is notified and prompted for some action to ensure the consistency of the metadata.

d) Metadata Versioning

The GCMD site [JCD01] does not provide any information in regard to how the system addresses issues associated with metadata versioning, version management, updates etc. However, it may be assumed that in order to deal with metadata versioning the system relies on the ability of the DIF specification to address such issues through its field "Metadata_Version".

e) Metadata Preservation

As per the in-depth research conducted on this tool, it does not contain any special technique that is able to preserve metadata for substantially long periods of time ensuring that the overall quality of metadata remains unchanged.

a) Security/Access Control

The system relies on the science coordinators of the GCMD, who are very careful about controlling unauthorized access to updating the metadata from docBUILDER tools (DIF authoring tools). In effect, any submissions from the tools go into a queue that requires coordinator action before being loaded into the database. If the person submitting the update is not an authorised person (e.g. DIF author or Data Center contact), then the coordinators will initiate an email request to the contacts already specified in the DIF to verify that the requested update is legitimate.

b) Platform Independence

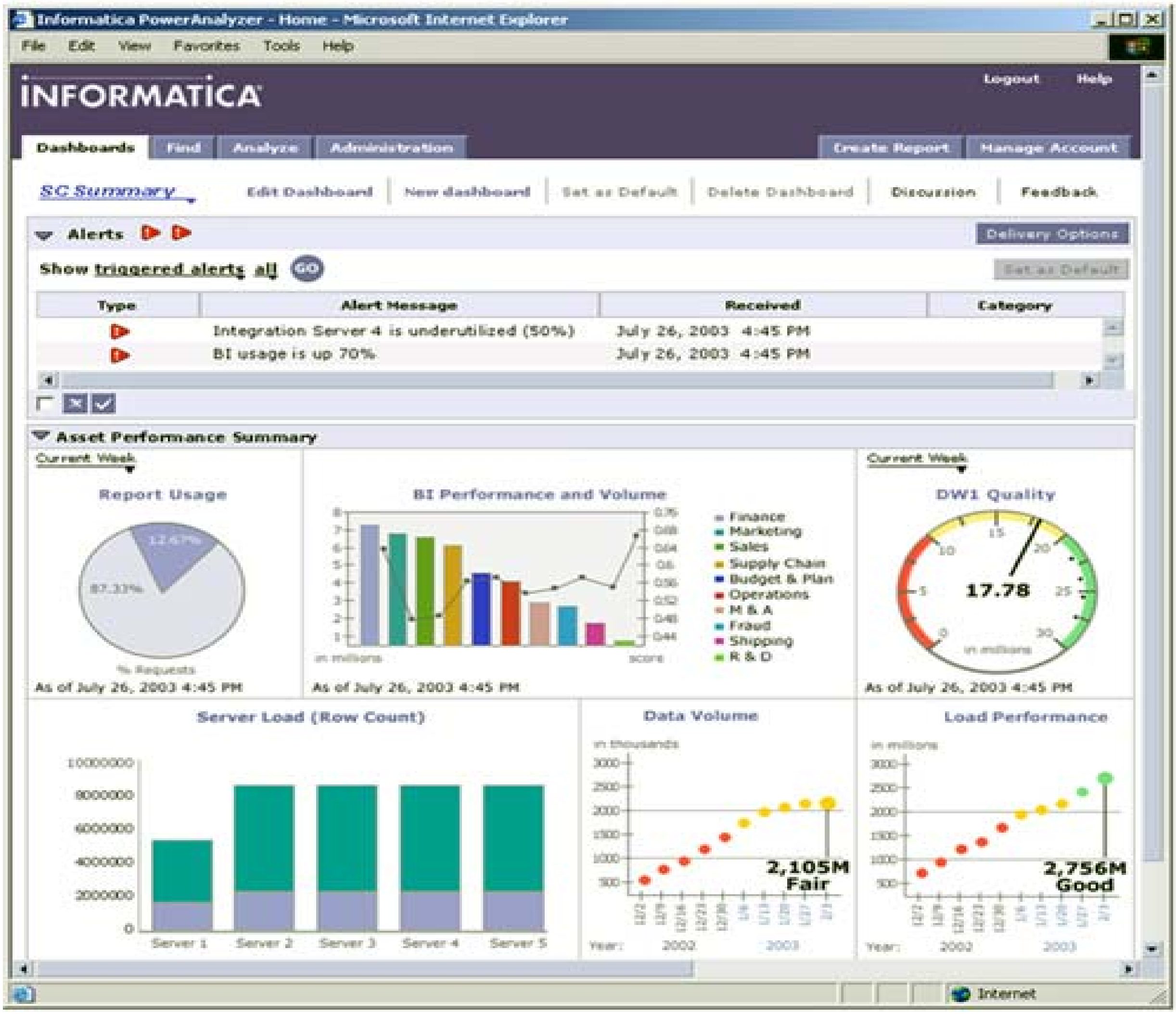
The software runs on all flavours of Linux, Sun's, SGI's etc., however is not compatible with any version of windows operating systems. In term of database, the GCMD metadata management tool was initially designed for Oracle but later improved to provide compatibility for PostgreSQL and McKoi. Besides, it supports all Internet browsers.

5.4.2 Concluding Remarks

The highlight of the metadata management system of the GCMD is its metadata quality assurance technique, which entails validation on both metadata syntax and semantics. In theory, this type of validation technique is a sought after feature that a system managing metadata over the long term should possess in order to ensure the overall quality of the metadata. Nevertheless, the validation technique utilized by the system, though seems effective is not infallible; therefore the possibility of erroneous data does exist. Furthermore, the GCMD site [JCD01] does not provide any relevant information as to how the system addresses some significant long-term metadata management issues, such as metadata versioning, version management, updates etc. Therefore, it was not possible to determine the efficiency of the system in addressing those issues.

5.5 Informatica SuperGlue™

With the aim to increase transparency of information assets and processes, this web-based Metadata management tool, developed by Informatica corp. Japan integrates Metadata from various systems as well as portraying and controlling the movement, lineage and utilization of information assets. According to an article by Denise Callaghan on EWeek, this tool provides analysis and reporting via Web-based dashboards (Figure 5.4) and other visualization techniques to gain insights into data quality and usage, as well as detecting redundancies and performing change impact analysis [IDC03].



Type	Alert Message	Received	Category
▶	Integration Server 4 is underutilized (50%)	July 26, 2003 4:45 PM	
▶	BI usage is up 70%	July 26, 2003 4:45 PM	

Figure 5.4: Web-based Dashboard of the Informatica SuperGlue™ [ISC04]

5.5.1 Infomatica SuperGlue™ Assessed [ISC04]

Detailed assessment results of the Informatica SuperGlue™ are as follows:

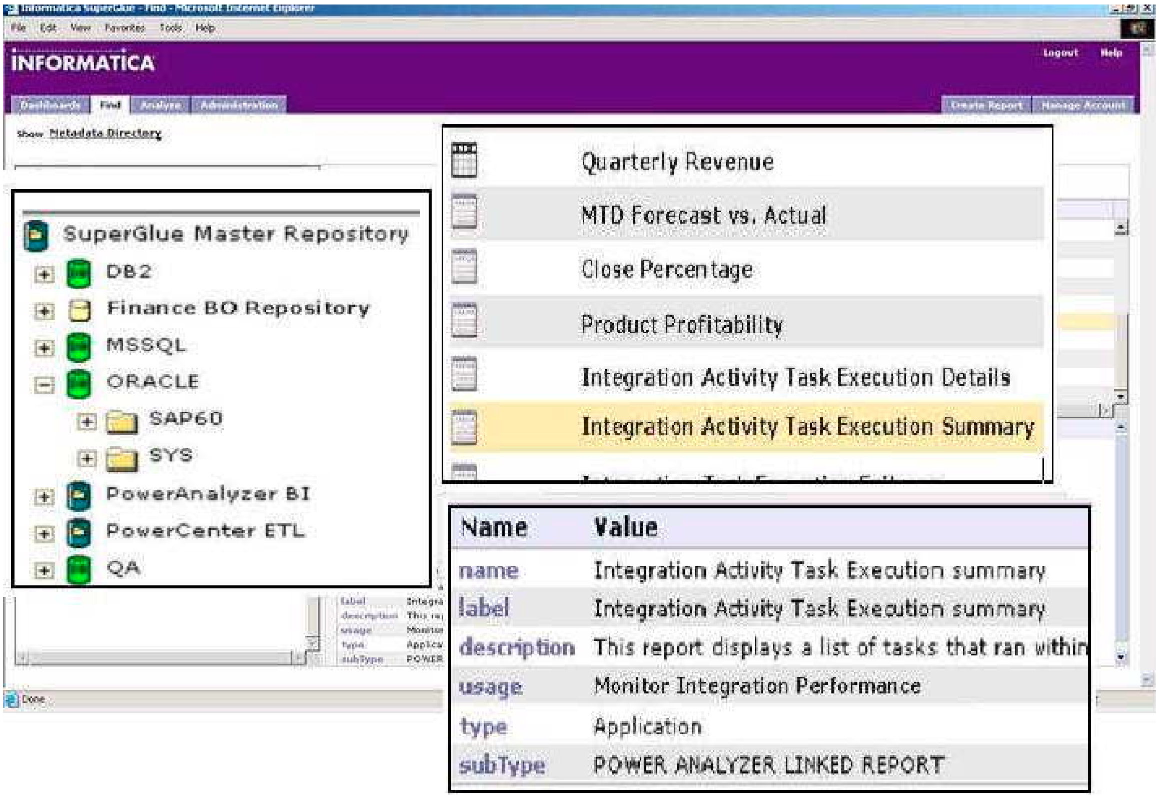
a) Metadata Creation/Updates/Editing

Based on an open and fully extensible metamodel architecture (a combination of MOF & CWM - see Appendix C), the Informatica SuperGlue™ offers a full complement of sophisticated metadata management capabilities. In general, metadata is collected in the Informatica SuperGlue™'s personalized information asset directory, which is an extensible, editable, and searchable catalogue of information assets. However, it is not clear how this metadata created or whether SuperGlue™ provides any metadata creation facility.

Nevertheless, the SuperGlue ™ ensures interoperability between metadata from disparate sources with the help of one of its most important components, PowerCetnre[61]. In addition, SuperGlue™ Xconnects, pre-built metadata adapters, are used to leverage the PowerCentre to link to and extract from specific metadata sources.

b) Search Facility

Metadata can be searched using the SuperGlue™'s personalised information asset directory (Figure 5.5). However, as per the information acquired in regards to this search interface as well as the SuperGlue™ as a whole, the tool does not appear to provide any controlled keywords or any other relevant feature in order to ensure accuracy and efficiency of the search results, thus facilitating enhanced search-ability of the metadata.


Informatica SuperGlue - Find - Microsoft Internet Explorer
File Edit View Favorites Tools Help
INFORMATICA
Logout Help
Dashboards Find Analyze Administration
Create Report Manage Account
Show Metadata Directory
SuperGlue Master Repository
DB2
Finance BO Repository
MSSQL
ORACLE
SAP60
SYS
PowerAnalyzer BI
PowerCenter ETL
QA
Quarterly Revenue
MTD Forecast vs. Actual
Close Percentage
Product Profitability
Integration Activity Task Execution Details
Integration Activity Task Execution Summary
Done

Name	Value
name	Integration Activity Task Execution summary
label	Integration Activity Task Execution summary
description	This report displays a list of tasks that ran within
usage	Monitor Integration Performance
type	Application
subType	POWER ANALYZER LINKED REPORT

Figure 5.5: Personalised Information Asset Directory [ISC04]

[61] A real-time data integration server that offers a complete set of integration capabilities.

c) Metadata Quality Assurance

One of the key benefits of using the SuperGlue™, as claimed by Informatica is the facility to identify information asset redundancies and opportunities for reuse. Now, in order to facilitate accurate reuse of data, it is absolutely vital to ensure its quality through efficient metadata quality assurance. However, it is not comprehensible whether and/or how the SuperGlue™, ensures quality of stored metadata, which is frequently subject to undergo different metadata management related operations, such as deletion, editing etc.

d) Metadata Versioning

As stated on its homepage, the SuperGlue™ aims to leverage traditional metadata management solutions that are focused mainly on metadata storage, version control etc. by providing a range of other business related facilities as mentioned above. Although such statement may well lead to the assumption that the SuperGlue™ addresses the issues of metadata versioning, this research was unable to acquire any information as to how and to what extent the tool handles issues associated with metadata versioning.

e) Metadata Preservation

The SuperGlue™ employs a repository (e.g. database) to store metadata that are created. However, it does not use any special technique for preserving these stored metadata over the long term.

f) Security

As a business oriented tool, the SuperGlue™ offers security facility to prevent unauthorised access to its information asset directory.

g) Platform Independence

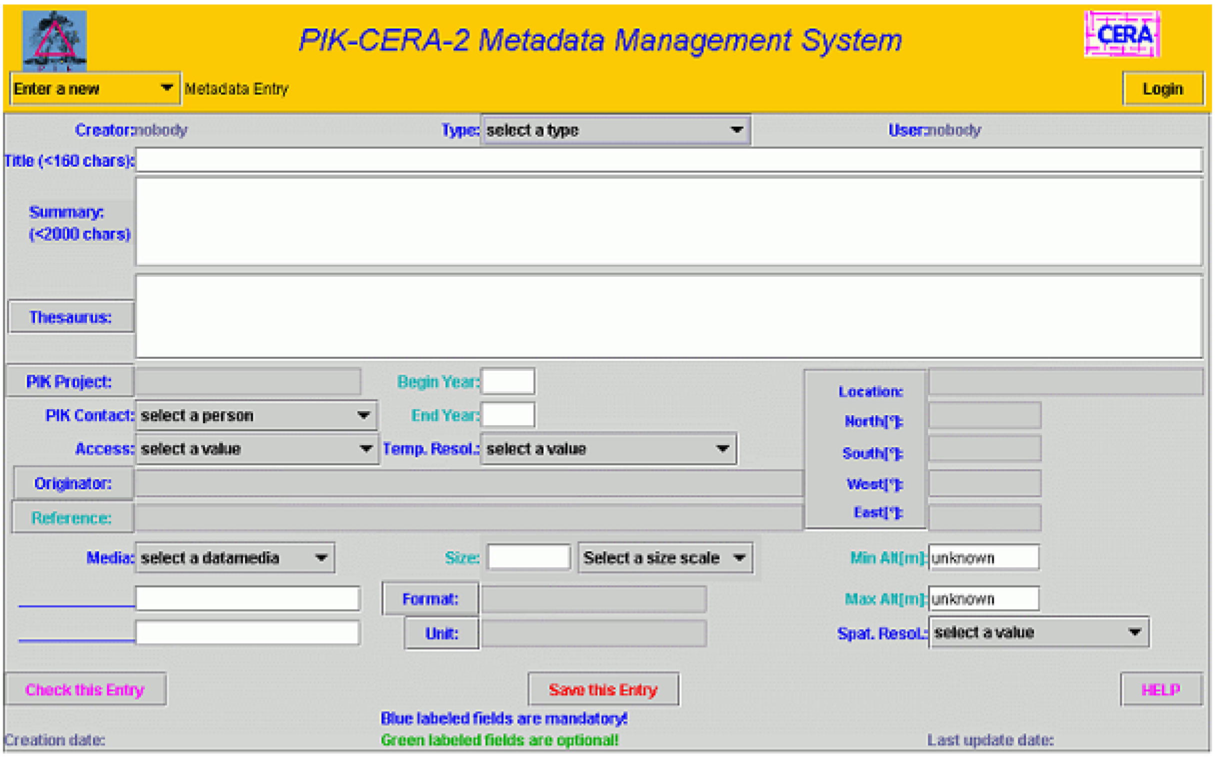
The Informatica's SuperGlue™ supports all major operating systems (e.g. Windows, Unix) and databases (e.g. DB2, Oracle, SQL Server).

5.5.2 Concluding Remarks

On the basis of the assessment results as presented above, the SuperGlue™, will probably obtain an average (or even low) score for its suitability for providing efficient metadata management service over the long term; mainly due to unavailability of sufficient information regarding some of its features, such as metadata quality assurance, versioning technique and metadata creation facility. However, it may not be sensible to reach any final verdict about this tool without thorough assessment of those features as they hold very high significance in the context of data curation. In order to clarify these aspects of the SuperGlue™, and how one would perceive its use for the purpose of long-term metadata management, several attempts were made to contact Informatica. Unfortunately, no reply has been received till the time of writing of this thesis. Therefore, in that regard, the assessment of the Informatica SuperGlue™ may appear somewhat incomplete.

5.6 The Java based PIK-CERA2 Metadata Management Tool MMT

The PIK-CERA2 MMT is the operational service for Metadata Management of Potsdam Institute for Climate Impact Research (PIK); that allows creating, updating, deriving[62] and deleting Metadata entries in the PIK-CERA2 meta-database. This meta-database is a web-accessible relational ORACLE™ database, developed on the basis of a subset of CERA2 (Climate and Environmental Data Retrieval and Archiving). The main interface of the tool is depicted in figure 5.6.


PIK-CERA-2 Metadata Management System

Enter a new: Metadata Entry

Creator: nobody

Type: select a type

User: nobody

PIK Contact: select a person

Access: select a value

Temp. Resol.: select a value

Media: select a datamedia

Size: Select a size scale

Min Alt[m]: unknown

Max Alt[m]: unknown

Spat. Resol.: select a value

Blue labeled fields are mandatory!

Green labeled fields are optional!

Creation date:

Last update date:

Figure 5.6: Main Interface of the PIK-CERA2 MMT [MMP03]

5.6.1 The PIK-CERA2 MMT Assessed

This sub-section provides in details the results of the assessment and evaluation conducted on different functionalities and features offered by the PIK-CERA2 MMT.

a) Metadata Creation/Updates/Editing

From the perspective of metadata management, this tool enables metadata creation, updates etc. through a PIK contact person who holds the responsibility of the data, which is described in a metadata entry. As it has been mentioned before, the all metadata are stored in the PIK meta-database associated with the system. In general, the underlying PIK meta-database, (MDB) PIK-CERA2, contains information about data available for or used by scientific projects at PIK. The data, mainly available in digital form, refer to earth and social sciences and have mainly been acquired from scientific, governmental and private institutions. However, it is not explicit from the information provided on [MMP03] if and how interoperability between metadata originated from disparate sources and conforming to different standards is ensured.

[62] Derive a new metadata entry from an existing entry.

b) Search Facility

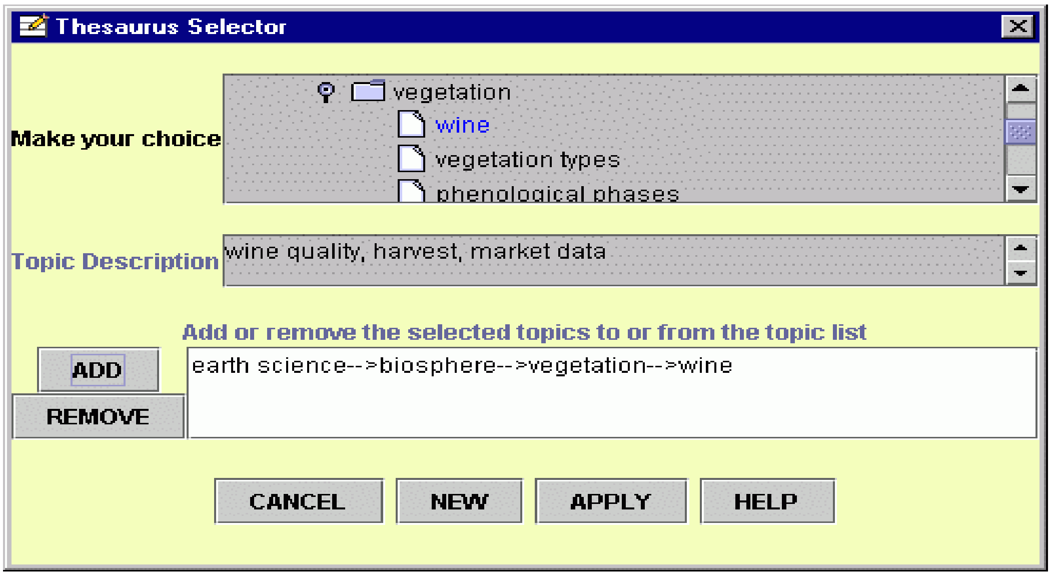
The system also incorporates a Java based, platform independent tool, xDat[63] to browse, retrieve, visualize and download metadata entries to database. However, the [MMP03] does not provide sufficient information to assess the efficiency of this xDat tool in performing those operations in data curation environment.

c) Metadata Quality Assurance

The quality of the data described in the MDB is determined on the basis of the geographical region they refer to, the way they have been created, the year they were taken, the distributor of the data, their mode of digital storage, a PIK contact person that has already used them and many others. In addition, the structures of the metadata stored in the database conform to first levels of the CSDGM and the DIF metadata standards.

However, the tool does not contain any automated mechanism or process for ensuring the quality of the metadata being entered into the database. Only the person, making entries into the database, is responsible for the correctness of the metadata. Besides, this person also has the right to update or delete a metadata record or to trust another person with this task. This may be regarded as a drawback from the viewpoint of long-term metadata management.

Nevertheless, the tool allows attaching keywords from an extendable multi-levelled hierarchical thesaurus (stored in the database: Figure 5.7) to a data entry, where higher levels prove a greater amount of detail in describing the record than lower levels. This facility may be beneficial to users for describing data accurately.



Topic Description: wine quality, harvest, market data

earth science-->biosphere-->vegetation-->wine

Figure 5.7 Thesaurus Selection Interface of the PIK-CERA2 MMT [MMP03]

In addition, the entered metadata undergoes further checks for inconsistency, errors etc. by PIK-CERA2 administrator, before being stored into the database and becoming accessible by the xDat retrieval tool. Nevertheless, on a long-term basis, the lack of automated metadata quality assurance functionality may pose certain threats, in terms of inconsistency and inadvertent errors, to the overall quality of the metadata.

[63] Extensible Database Access Tool

d) Metadata Versioning

As an added drawback for managing metadata over long-term, the tool does not provide any specific functionality to address the issues of metadata versioning.

e) Metadata Preservation

Neither the web-accessible relational ORACLE^{TM} database associated the tool nor the PIK MMT as a whole incorporates any appropriate preservation mechanism/technique to ensure perpetuation of the metadata.

f) Security/ Access Control

In order to ensure the consistency in metadata entries into the database, the tool has access control facility (e.g. mandatory log on functionality) to prevent unauthorised access to the database.

g) Platform Independence

As the PIK MMT is a Java-based software, it should be compatible with all major commercially available operating systems, such as Windows, Linux, UNIX etc. In addition it should also support all popular Internet browser, such as Microsoft Internet Explorer, Netscape etc. Nevertheless, very limited information was available in reference to the tool's platform independency.

5.6.2 Concluding Remarks

In an attempt to test the functionality of the tool, efforts were made to create new username and password for the system. However, the attempt was not successful as the link on the PIK website to such facility appears to be only for internal use and prohibits any outside access. A request was sent to the contact email address provided on the [MMP03], in order to obtain further detailed information with regards to the functionality of the tool and possibly an evaluation copy. In reply, Michael Flechsig of the PIK - Potsdam Institute for Climate Impact Research Dept. Data & Computation, confirmed that due to lack of resources there is currently no (semi-)automated service available to keep track with changes in the data in the meta-database at the PIK.

5.7 Metadata Management Systems' Assessment Matrix

This section presents a matrix in the form of table 5.1 highlighting the most significant and relevant (for long-term metadata management/data curation) outcomes that have yielded from the comparison between the results of the assessment and evaluation performed on six above described metadata management systems/tools.

Tool/System Name		**MetaStar DL**	**MetaMatrix MetaBase™**	**SMMS™**	**GCMD**	**Super Glue™**	**PIK MMT**
Metadata Creation/ Updates/ Imports	**Ensures Interoperability?**	Yes	Yes	No	No	Yes	Not Known
	Controlled Vocabulary	Yes	No	No	Yes	No	Yes
	Other Documentation?	Adequate	Adequate	Adequate	Adequate	Not Known	Adequate
Search Facility		Very Good	Very Good	Good	Very Good	Average	Average
Metadata Quality Assurance (Validation etc.)		Average	Average	Average	Good	Not Known	Average
Support Metadata Versioning		No	Yes	No	Not Known	Yes	No
Support Long-term Metadata Preservation?		No	No	No	No	No	No
Scalability/Customizability		Very Good	Average	Good	Not Known	Good	Average
Security/Access Control		Very Good	Good	Good	Very Good	Very Good	Good
Platform Independent?		Fully	Fully	Partly	Partly	Fully	Fully

Table 5.1: Metadata Management Systems' Assessment Matrix

Aside from the six metadata management systems as described above, a number of other systems that pertain to long-term metadata management or data curation were also studied and assessed on the basis of the same criteria as mentioned before. A summary of the results of their assessments has been provided in Appendix E.

Chapter 6

A List of Potential Collaborators

One of the main requirements for this MSc. project was to conduct an exhaustive research in order to locate and assemble a list of experts, research groups etc. who are working or have worked erstwhile to achieve similar or related objectives to those of the future portion of this project (in the coming 2.5 years) and are most likely to act as potential collaborators. Expectantly, these collaborators will lend necessary expertise and knowledge to the future project, which will aid in developing a working prototype of a system that will manage high quality metadata over the long-term in order to ensure effective perpetuation of the actual resources. This chapter presents a list of such potential collaborators as resulted from this project.

6.1 The CEDARS Project

On the April 1st, 1998 the CEDARS project (*CURL Exemplars in Digital Archives*) was officially launched as collaboration between three *Consortium of University Research Libraries* (CURL) institutions, the universities of Leeds, Cambridge and Oxford. Being a higher education-initiative, the project was funded by the *Joint Information Systems Committee* (JISC) of the UK higher education funding councils under Phase III of its Electronic Libraries (eLib) Programme [CED02].

In terms of the main objectives, the CEDARS intended to explore the challenges posed by the archival storage and long-term preservation of digital information with significant emphasis on Emulation as a long-term data preservation approach. More importantly to this MSc. Project (and the future work), the CEDARS perceived the role of metadata as pivotal for both long-term preservation strategy and the collection management. Inspired by such perception and strongly influenced by the OAIS reference model, the CEDARS proposed and implemented a basic set of preservation metadata elements, which were tested in a demonstrator archive to determine their effectiveness in successful perpetuation of data.

As a whole, the CEDARS' main interest coincides with the broader objective of this MSc. Project - successful long-term good quality data preservation. Although the project ended in March 2002, it may still be seen as promotion of awareness about the importance of digital preservation. Therefore, this mutuality in ultimate goals along with the CEDARS' strong interest in Metadata, certainly make its three partner institutes desirable sources of collaboration to provide aid in the future work for this MSc. Project. Relevant contact details of CEDARS personnel have been provided in appendix G.

6.2 The NEDLIB Project

Launched on the January 1st, 1998 and funded by the *European Commission*'s *Telematics Application Programme,* Networked European Deposit Library (NEDLIB) is a collaborative project of eight European national libraries, which mainly focuses on long-term data preservation. Having initiated by the *Conference of European National Libraries* (CENL), the *Koninklijke Bibliotheek* of the Netherlands headed the project with participations from other national libraries, such as France, Norway, Finland, Germany, Portugal, Switzerland, and Italy. Further partners include a national archive and three major publishers, namely *Kluwer Academic*, *Elsevier Science*, and *Springer-Verlag* [NED01].

The main objective of the project is to construct the basic infrastructure upon which a networked European deposit library can be built; which will consequently ensure long-term preservation of both, on-line and off-line digital publications. During the course of achieving this goal, the project proposed guidelines and technical standards (preservation metadata standards) to bring in a common basis enabling close cooperation and, hence, spreading research costs. In addition, on the basis of the OAIS standard, the project identified and formalised models that fundamentally cover all steps from the acquisition of the documents, via access provision, to their long-term archivation. The model is subsumed under the generic architecture of a *deposit system for electronic publications* (DSEP) process (see 4.7).

As a whole, the work of the NEDLIB project should provide enhanced insight into the pros and cons of different long-term preservation strategies as applied in digital deposit collections, therefore may well be useful for deciding upon an appropriate preservation strategy to be implemented for long-term metadata preservation. Therefore, although the project officially completed in December 2000, the experiences and expertise of the project partners in terms of long-term data preservation grant them a place in the list of potential collaborators for the project to be undertaken in the coming 2.5 years. Relevant contact details of the project partners have been provided in Appendix G.

6.3 The OCLC & RLG Working Group

The Online Computer Library Centre (OCLC) & the Research Libraries Group (RLG), both are non-profit, membership organisations, with the former being a computer library service and research organization whose computer network and services link more than 36,000 libraries in 74 countries and territories and the latter being a corporation of over 160 universities, national libraries, archives, historical societies, and other institutions. The OCLC was founded in 1997, whereas The New York Public Library and Columbia, Harvard, and Yale universities initiated the RLG in 1974 [OCL04, RLG04].

Over the last few years, these two organizations jointly have embarked on mainly two ventures that aimed to address the issues of long-term preservation of digital information. First of these two endeavours occurred in March 2000, when OCLC and RLG announced their shared commitment to encourage the development of infrastructure to support the long-term preservation of digital materials. Subsequently, two working groups, jointly sponsored by OCLC and RLG and comprised of expert participants from a variety of institutional and geographical backgrounds, were created. While the first group was responsible for identifying key attributes and responsibilities of trusted digital repositories serving cultural heritage institutions, the second working group was set out to identify and describe metadata necessary to support the digital preservation process [PRM03].

Shortly after the completion of these working groups' work in May 2002, an opportunity and need emerged to extend the work of this group through some form of follow-on activity. This resulted in the OCLC and RLG convening an expert working group, namely Preservation Metadata Implementation Strategies (PREMIS), with the aim to focus on the practical aspects of implementing preservation metadata in digital preservation systems. The PREMIS were scheduled to finish around June 2004 [PRM03].

Considering these two organisations' organised and ambitious activities as well as significant experiences in the field of long-term data preservation and related metadata issues, it may be stated that both the OCLC & the RLG, at least the PREMIS working group, are fully capable of lending necessary expertise and help for the future work of this MSc. Project. Further assertion of such collaboration lies in one of the PREMIS's long-term goals – to continue its efforts to engage interested parties in the digital preservation community as work proceeds. Relevant contact information has been provided in Appendix G.

6.4 The NLA Working Groups

The National Library of Australia (NLA) is currently undertaking a project, titled "Digital Services Project", which is literally the library's key strategy for ensuring effective management of its digital collections and their preservation for future access as technologies change. The project encompasses a wide set of IT development and procurement activities which together support the overall framework and systems architecture for the NLA digital library. One of the main objectives of this project is to provide the infrastructure for long-term management of digital material in the Library's collection through provision of hardware and software systems supporting integrated collection management in a digital environment.

In pursing these goals, the project has already developed a number of prototype systems. Among these systems, the metadata repository & search system, Digital Collection Management System and Digital Object Management System hold the most relevance to this MSc. Project, as these systems addresses issues associated with long-term metadata management and storage within their functionalities. Therefore, on the basis of similarity in main goals and relevant experience in the field of long-term metadata management rather than the wider context of data preservation (unlike CEDARS, NEDLIB etc.), the members of the Digital services project of the NLA should be elected as potential collaborators for the future phase of this MSc. Project [ADD04].

In addition, the NLA incepted a preservation metadata-working group, around 1999, which possess extensive knowledge in regard to main metadata requirements for long-term data preservation and should be able to provide assistance in developing a suitable and efficient metadata standard for long-term metadata management [PWG99]. Relevant contact information for both of these working groups has been provided in Appendix G.

6.5 The NERC Data Grid Project

The Natural Environment Research Council (NERC) is one of the seven UK Research Councils[64] that fund and manage scientific research and training in the UK. The primary goals of the NERC are to promote and support, by any means, high quality basic, strategic and applied research, survey, carry out long-term environmental monitoring and provide advice on disseminate knowledge about as well as promote public understanding of environment.

The NERC consists of seven designated data centres established to carry out the NERC data policy, which has been instituted to encourage data sharing and curation (backed-up by a requirement for NERC funded researchers). Among these data centres, the BODC (British Oceanographic Data Centre) and the BADC (British Atmospheric Data Centre) are the most active in digital curation as their main objective essentially is long-term data preservation - one of the core activities of long-term digital curation. Currently, investigators or experts from the BADC and BODC as well as the CCLRC e-Science Centre are contributing to a digital curation related project, which is financed jointly by the NERC and National e-Science Core programme[65]. The project, titled *NERC Data Grid,* principally aims to build a grid, which makes data discovery, delivery and use (i.e. curation) much easier than it is now, facilitating better use of the existing investment in the curation and maintenance of quality data archives [NRC02].

Considering the highly intricate and crucial nature of the digital curation related issues that the aforementioned NERC project aims to handle, the researchers or investigators involved with this project are expected to possess significant expertise and insightful knowledge in various aspects of digital curation. These proficiency and knowledge are very much sought after for resolving any curation related problem that might arise during the course of development of a working prototype of long-term metadata management system. Relevant contact information for the NERC Data Grid project may be found in Appendix G.

6.6 The UK Data Archive (UKDA)

The UK Data Archive (UKDA), founded in 1967, is an internationally renowned centre of a range of expertise including data preservation. In addition, it is curator of the largest collection of digital data in the social sciences and humanities in the UK. Funded by the Economic and Social Research Council (ESRC), the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils and the University of Essex, currently, it accommodates several thousand datasets of interest to researchers in all sectors and from many different disciplines [UDA04].

As far as the extensive research activities of the UKDA are concerned, over the past four years it has incepted or affiliated itself with a couple of research groups or projects, such as Metadata Management and Production System for Surveys in Empirical Socio-economic Research and Cluster Of Systems of Metadata for Official Statistics (COSMOS), which centre on the issues associated with metadata management (see 2.4). While the former

[64] A strategic partnership set up to champion science, engineering and technology supported by the seven UK Research Councils - http://www.rcuk.ac.uk/

[65] http://www.rcuk.ac.uk/escience/

(currently ongoing) of these two projects dedicates research efforts primarily towards developing an appropriate metadata standard as well as a metadata management tool for large scale comparative surveys over space and time, the latter (ended successfully in August 2003), which is an accompanying cluster of five projects of the European Union, focused mainly on building metadata repositories by exchanging ideas and experiences in using metadata systems for the individual projects.

In light of the above discussion, it may not be inaccurate to say that the UKDA possess proven proficiency in relevant (to the main interests of this MSc. project) aspects of metadata management along with desirable curatorial expertise. Therefore, this organisation may be regarded as a commendable source of potential collaboration for developing a working prototype of long-term metadata management system in the context of digital curation. Relevant contact information of the UKDA has been provided in Appendix G.

6.7 Other Sources of Collaboration

Aside from the aforementioned sources of collaboration, there are a few other working or research groups, who have extensive knowledge and proven expertise relevant for the future works for this MSc. Project. Brief summaries of these potential collaborators have been given as follows.

6.7.1 The Digital Archiving Consultancy (DAC)

Established in March 2002, the Digital Archiving Consultancy (DAC), UK provides independent expert advice and consultancy services on data preservation, archiving and curation. The organisation is already a leading provider world wide of high quality advice on these issues over a broad range of application areas. The DAC brings distinctive skills to the specific challenge of archiving scientific, technical and medical data and their curation. In addition, they provide customized training courses and seminars on all aspects of digital archiving, curation and preservation [DAC03]. Contact information of the DAC may be found in Appendix G.

6.7.2 The National Information Standards Organization (USA) Working Group

In 1999, the National Information Standards Organization (USA) formed a working group to undertake a project that was intended to address the standardization need in the digital imaging community, as technical metadata was perceived as an essential component of any digitisation initiative for short-term and long-term management purposes. The group delivered a draft in July 2000, which presents a comprehensive list of technical data elements required to manage digital image collections. An important goal of the Standards Committee is to outreach various communities that will be interested in the development of such a standard. Therefore, the experiences and expertise gained by this working group during the course of this project may provide necessary aid in developing a standard for long-term metadata management [NIS03]. Relevant contact details are given in Appendix G.

6.7.3 The NEESgrid Working Groups

The NEESgrid project (see 4.6) is collaborative efforts of several US based working groups or experts for building a grid system, which will be capable of linking earthquake researchers across the U.S. with leading-edge computing resources and research equipment, allowing collaborative teams (including remote participants) to plan, perform, and publish their experiments. Metadata management and curation are very important part of the NEESgrid. In fact, NEESgrid has a designated team of experts for managing data and metadata within the project. In addition, the NEESgrid System Integration team, has recently held a summit in Chicago that brought together experts in library information science, earthquake engineering, data infrastructure, and data curation, to forge a forward-looking plan needed to improve the NEESgrid data usage and curation [NES04]. This clearly underscores its remarkable interest in digital curation. The working groups or experts involved with metadata management and curation within the NEESgrid project may be considered as estimable sources of collaboration for related issues associated with the forthcoming work of this MSc. project. Contact Information is given in Appendix G.

6.7.4 The DCMI Preservation Working Group

The Dublin Core Metadata Initiative has a currently active working group, namely DCMI preservation working group, aims to collect information on, and review, existing preservation metadata schemas, to investigate the need for domain specific preservation metadata schemas, and to liaise with other global preservation metadata projects (e.g.PREMIS) as appropriate. The working group was established at the Seatle meeting of the DCMI Advisory Board in 2003 [DPW04].

6.7.5 The Database Group of the University of Leipzig, Germany

The Database group within the computer science department of University of Leipzig, Germany are currently undertaking a number of metadata related projects. One of these projects is titled “Model Management”, which attempts to devise a new approach for generic metadata management that manipulates models and mappings between models using high-level operators. Vital information in regard to generic metadata management techniques may be borrowed from the experts involved in this project. It should be noted that the model management project is carried out in collaboration with Microsoft’s database group and researchers from four other universities [ULP04]. Relevant contact information may be found in Appendix G.

6.5.6 The European Bioinformatics Institute

The European Bioinformatics Institute (EBI) is a non-profit academic organisation for research and services in bioinformatics; that forms part of the European Molecular Biology Laboratory (EMBL) [EBI04]. It has a data curation team who is responsible for performing various curatorial operations on data being submitted to one of the centre’s repositories or databases called Array express. The array express is, in fact, an international public repository for microarray gene expression data; which aims to store and provide access to well-annotated data from microarray experiments [EDB04]. Appendix G provides relevant contact information of the EBI data curation experts.

Chapter 7

Future Plan of Work

This chapter aims to present detailed plan of work for a project, which is to be undertaken over the coming 2.5 years in order to develop a working prototype of metadata management system in the context of data curation. The project plan outlined in this chapter is susceptible to adapt to any changes that may aid in accelerating the achievement of the project goal.

7.1 Project Phases/Tasks

This section provides the work breakdown structure of the future project in details, with recommendations in regards to different aspects of the project; constructed and complied on the basis of the principle and most relevant findings of this MSc. Project as presented in the previous chapters (3-5) of this dissertation.

7.1.1 Phase 1 - Requirements Gathering & Definitions

The traditional approach for requirements gathering for any project is user consultation, requirements workshops etc. However, in case of wide variety of end-users and unprecedented metadata creation procedures, this approach may be deemed inappropriate. In that case, alternative approach will be to perform an extensive survey within the potential users (e.g. organizations, individual users etc.) of the metadata as well as collecting information from similar projects, i.e. project collaborators that have been implemented for similar metadata domain.

Nevertheless, irrespective of the requirements gathering technique, this phase of the project should aim to obtain the followings:

- **Specification of the Metadata Source & Environment:** The very first step of this phase should be to specify (or identify) and study the metadata source and environment as well as the scope of the data to be dealt with. As it has been mentioned before, metadata environment may be an archiving system or a data warehouse or some other data storage environment. This first step of the plan is highly crucial as the decisions to be made on the above mentioned issues would considerably influence the operations of the Metadata Management System. Therefore, in order to develop an efficient approach for metadata management for the long term, it is absolutely vital to know and understand standard operations on metadata within a particular metadata environment, as different metadata environments require different management techniques.

- **Compatibility & Interoperability**: As different applications use different programming structures, syntax, and semantics to model their metadata, degree of metadata compatibility with, and interoperability between the applications in the metadata environment will be defined.

- **Extensibility**: It is deemed essential to define all the functional requirements relating to the exposure of metadata to any external services required.

- **Levels of Granularity**: Degree of granularity (e.g. a metadata element "type" which describes the nature or genre of the resource might be used to 'filter' search results.) in metadata description required to support efficient discovery, retrieval, use and preservation of information objects, should also be decided upon.

- **Requirements for a Metadata Management System:** Research and survey will be carried out to define the functional and non-functional requirements for an ideal metadata management system that is to manage metadata of required quality over substantially long periods of time. Requirements analysis will also take in to considerations a number of aspects such as, integration with other applications, hardware & software requirements, system security, number of users, response times and performance, volume of data etc. In addition, these requirements will reflect the general requirements as mentioned above.

7.1.2 Phase 2 - Feasibility Testing

Having collected all core requirements and specified the metadata source & environment, next phase of the project will conduct a feasibility testing to address the followings:

- Is long-term metadata management feasible within the metadata environment and for the requirements gathered?

- Is it feasible to develop a generic infrastructure for managing metadata over long periods of time? If so, what would it need to do, beyond what is offered in the best object-oriented databases and repositories?

The feasibility testing will require thorough research on the metadata environment and gathered requirements in relation to the principles and concepts for long-term metadata management (see 2.4) as well as relevant existing works presented in this thesis.

7.1.3 Phase 3 - Analysis & Design

This phase should have the following tasks to be performed:

Task 3.1 - Defining a Metadata Model or "Application Profile"

If the feasibility testing generates positive result, then the next step will be to define a common and standard way (e.g. XML Schema, Metadata Object Facility etc. see Appendix B) to represent or model metadata. A well-defined model will have very precise definitions of what the features and attributes of particular model instances mean. These precise definitions will then allow defining exact and unambiguous mappings of the model features to particular languages and interchange formats. This is often called "Application Profile" for metadata as it involves defining a list of metadata elements, encoding schemes and controlled vocabularies needed to support particular project requirements.

Defining a metadata model or application will require establishing an agreed and most appropriate **Metadata Standard**. As the research results of this MSc. Project have explicitly indicated that none of the recognised metadata standards addresses the issues associated long-term metadata management in direct terms, it may be necessary to formulate a metadata standard appropriate and suitable for this job, ideally with the most relevant features but without the drawbacks (in the context of data curation) of currently available recognised standards. On the basis of the relevant outcomes of this MSc. Project, the following features/attributes may be sought for a metadata standard especially designed to serve the purposes of long-term metadata management:

i. The most significant feature that the standard should possess is adequately comprehensiveness for addressing long-term preservation, management (e.g. preservation technique, certain structures of the bit stream of the digital objects etc.), and accessibility over long time etc. However, overall complexity of the standard should be restrained so that attention to format does not overtake attention to actual content.

ii. The metadata elements should be extensible, i.e. customisable to facilitate description of information specific to any dataset. Besides, the metadata structure should easily handle non-static datasets. Therefore, rather than defining the elements as either "obligatory" or "optional", they should be defined as "Very significant", ”Significant", ”Less Significant" or even ”Essential", " Essential If appropriate", ”Desirable" etc. depending on their usefulness for preservation or general resource discovery etc.

iii. As it is impossible to determine unequivocally what will be essential in order to manage digital preservation in the future, the metadata elements should reflect necessary assumptions about the future requirements in that regard. Therefore, the metadata set should ideally support both migration and emulation approaches for long-term preservation of digital information.

iv. The metadata elements should be capable of being mapped to the elements of other approved recognised standards in order to ensure metadata interoperability as well as reducing overall metadata creation time.

v. The standard should provide controlled vocabularies, keywords and any other features and/or documentations deemed necessary to ensure metadata quality and accessibility (or discovery).

vi. The standard should support different versions (i.e. version control/management) of both metadata and data.

vii. The standard should, preferably, support metadata in multiple languages.

Task 3.2 - Constructing an Approach for Metadata Management

The next step of the design & analysis phase will be to define and/or construct an approach for metadata management based on the metadata model or application profile defined in the preceding step. Different existing approaches to managing metadata, described in this thesis (chapter 4), may well serve as the foundation for deriving and defining the most suitable approach for managing metadata for the long-term within particular metadata environment. Besides, the collaborators of the project are also expected to provide aid in the successful completion of this task. The definition of metadata management approach, aiming to ensure high quality and well-managed metadata over long time, should incorporate the followings:

Quality Assurance Process

In general, determining whether the metadata involved is good enough to support the functional requirements defined above or conforms to the metadata model defined over long time will assess metadata quality. Therefore, it will be essential to define an efficient and intelligent quality assurance process to ensure high quality metadata for substantially long periods of time. Intrinsically, this quality assurance process should incorporate efficient metadata verification/validation (both syntactically and semantically) mechanisms.

Preservation

An appropriate and effective technique (migration or emulation etc.) will need to be devised in order to preserve metadata for long time. However, due to the long-term character of the task, the ultimate efficiency of the preservation strategy can only be roughly estimated at this point of time, therefore at least a viable near-optimal strategy has to be constituted. The assessment result of the OAIS reference model as well as different existing preservation approaches described in chapter 4 may aid in formulating a suitable preservation strategy. It should also be noted that metadata will ideally be stored in a repository and in order to provide storage space of a huge repository, hard-disk arrays are recommendable to keep the information directly accessible. In addition, metadata requires storage technologies capable of structured access modalities, such as queries against collections of objects, or partial updates to complex objects. Therefore, recommended storage technologies for metadata are the relational database management system (RDBMS) and the associated standard query language (SQL), as they are probably the most mature and robust technology of this sort [STN02].

As it has been mentioned before, manual long-term data preservation is very expensive due to labour intensity of the underlying processes/operations, even if the storage is minimal. Therefore, in order to reduce the total cost of preservation, all the operations (at least those that are deemed most critical) associated with data preservation should be highly automated.

Metadata Version Control

Suitable techniques for controlling & managing different metadata versions will also be defined on the basis of different existing metadata versioning mechanisms assessed by this MSc. Project.

Security

In order to guarantee that metadata stored in the repository is not corrupted inadvertently or altered maliciously, an efficient security infrastructure should be designed. The infrastructure and its underlying security mechanism should be easy to apply and durable over long periods of time. Therefore, an algorithmic solution, such as digital signature, digital watermarking, public key infrastructure etc. may be required. A recommendation may be to design a Rights Management System and/or Access Control System (integrated into the main system) containing the appropriate security technique to provide different access levels/privileges for users as appropriate. In addition to user authentication, audit trail or records (see 2.4.6) of metadata will need to be supported.

This approach will need to be approved by the project supervisor and adapt to any changes as required.

Task 3.3 - Designing a Metadata Management Architecture for the Working Prototype

A suitable architecture for metadata management will be designed in accordance with the metadata management approach defined in the preceding step of the project. This architecture will take into considerations, all potential interactions between the metadata management system, users, and other applications (e.g. database/repository etc.). Besides, at this stage of the project, all functional and non-functional requirements of the working prototype as gathered or acquired during phase 1 of the project will be analysed and transformed into specifications. As per the research results of this MSc. project, the working prototype should ideally have the following main features or functionalities [MRS99]:

- **Metadata Entry Functions**: The system should be capable of supporting Metadata collection management workflows, including insert, update and delete functions. Thus, the users will be able to update the underlying metadata repository in real time through a Web user interface (preferably). Ideally, Metadata insert and delete functions should have incorporated metadata validation or verification mechanisms to check metadata integrity (for both syntax and semantics) at the point of entry or update. In addition, these functions should have support for metadata versioning. In general, the system should be capable of supporting metadata for a wide variety of formats, including image, graphics, text, sound, and video.

- **Metadata Import/Export Functions:** The system should support the import/export of metadata created in approved metadata standards (standards that are not supported should ideally be converted or mapped on to the approved or accepted standard) and formats from other applications, systems etc. Metadata Import/Export Functions should also support metadata versioning.

- **Search/Browse Functions:** The system should offer supports for any combination of fielded metadata searching. Besides, support for searching based on digital object content is desirable, for example full text searching of text objects or searching by image or video content (Ideally, ability to navigate from the metadata to the digital object itself and to related metadata/digital objects is a sought after feature). There should be support for searching via a user-definable multi-level hierarchical thesauri or controlled vocabularies used to index digital objects. In addition, the users should be allowed to browse flexibly through result sets, switching between brief and full displays and exploiting links to related information.

- **<u>Management and Administration Functions:</u>** The working prototype should allow definition of additional metadata fields, management of the metamodel, user, group profiles and access levels as well as creation of programming interfaces to and reports from the System, through these functions. In addition, these functions should provide support of an audit trail (see 2.4.6) over the lifecycle of each metadata object.

- **<u>Authenticity</u>**: Ideally, the system should interface with a Rights Management and/or Access control system to allow or deny access to the stored metadata in the repository depending on access conditions specific to an object or set of related objects, user class, location etc. In addition, the system should have infrastructure in place to interface with appropriate signature and/or public key services to enable external users to ensure the authenticity of the metadata. Thus, the system will be able to provide sufficient security in order to prevent malicious or inadvertent modifications of the metadata, consequently, ensure that the overall metadata quality remains intact.

The design of the Metadata Management Architecture will reflect all specified functional requirements and will be submitted to the supervisor for approval. The design is susceptible to adapt to any change as suggested by the supervisor. The diagram below illustrates the aforementioned recommendations for metadata management architecture:

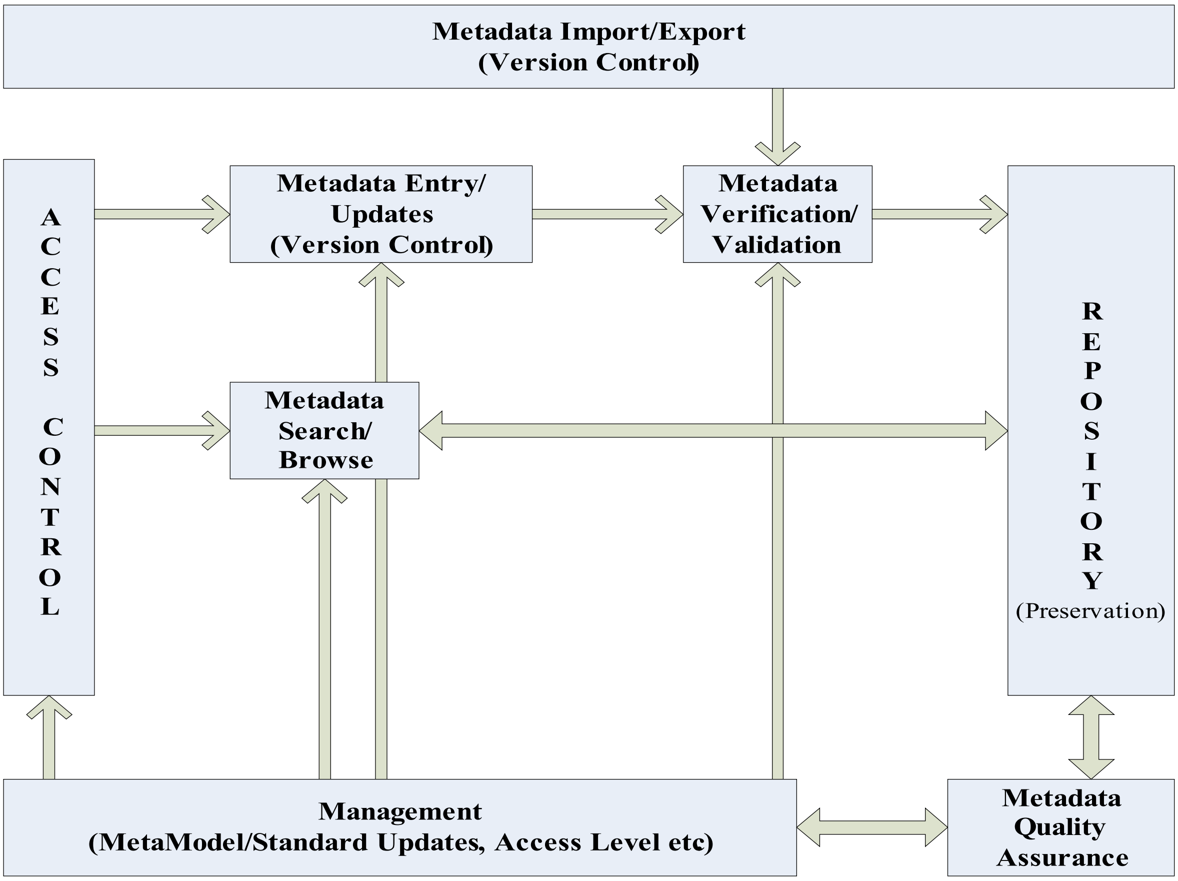

Metadata Import/Export
(Version Control)
ACCESS CONTROL
Metadata Entry/ Updates (Version Control)
Metadata Verification/ Validation
REPOSITORY
(Preservation)
Metadata Search/ Browse
Management
(MetaModel/Standard Updates, Access Level etc)
Metadata Quality Assurance

Figure 7.1: Recommended Architectural View of the Working Prototype

7.1.4 Phase 4 - Implementation, Testing & Re-Design of the Working Prototype

Once the design has been fully approved, the implementation of the working prototype will commence. After the implementation, the prototype developed will be thoroughly tested (e.g. Black Box, White Box etc.) in accordance with pre-developed test scenarios. If any flaws or inefficiency resulted from testing phase, design and requirements analysis phases will be repeated to ensure efficiency and usability of the prototype. Every repetition will result in a new version of the system.

7.1.5 Phase 5 - Deployment, User Manual, Training etc.

Having eliminated all the flaws & errors of the prototype, the final phase for this project will involve installing the prototype within the metadata environment. Also, a user manual detailing how to use the system will also be produced and made available to the users. In addition, training scheme may be designed to help train the users within the organisation concerned about the proper use of the prototype.

7.2 Estimated Time Scales for the Project

A Gantt chart detailing estimated completion times for above described different phases of the project and their constituting tasks have been given in Appendix I. It is to be noted that the time scale or duration for each project phase/task presented in the Gantt chart is estimated for one-person project and may well be less if more (than one) people are to work on the project. Furthermore, aside from the design documentations etc. progress reports of the project will be produced and submitted every 6 months of the project period.

Chapter 8

Conclusions

To summarise, this dissertation has attempted to discuss in details the main achievements of a MSc. Project, the main focal point of which lies in the subject of long-term metadata management and its quality assurance with a broader objective of successful long-term data preservation. The paper begins with explicit specification of the project objectives, followed by an insight into the main issues and criteria that have been considered for the project tasks. Then it thoroughly describes the main projects accomplishments in the subsequent chapters. The thesis ends by presenting a detailed work plan for the coming 2.5 years for developing a working prototype of a system that will serve the tasks of managing metadata over long periods of time in the context of digital curation.

Efficient and effective long-term metadata curation is a key component of successful preservation, enrichment and access of digital information in the long term. This paper identifies that a number of current relevant metadata standards, systems and approaches in existence do not address the full set of metadata curation requirements. The necessity of metadata standards, metadata management standards and system, which more fully address the needs of metadata curation is highly evident from the survey of existing systems, standards and approaches. Developing new standards for both the metadata and metadata management realm would not be an efficient strategy; therefore a specification of extensions needed to aid metadata curation for existing standards and systems is recommended and seen as a fruitful area of further work. Some of the results of this work will no doubt be based on a union of the best features of existing systems. However there are many things (e.g. versioning, migration, annotation) which are not dealt with effectively by any of the existing approaches mentioned.

Also of note is the fact that a significant number of publications available at present address the issues associated with long-term preservation of digital information to differing extents. However, ironically, most of these publications tend to focus only on the metadata requirements for preserving information for the long term, completely overlooking the fact that the metadata associated with the data also needs to be preserved along with the resources that they are describing; in order to ensure the longevity of these resources.

Existing metadata management systems offer some commendable features (e.g. advanced search facilities, interoperability support) for generic metadata management. None of them is intended for long-term metadata curation or management; for example they do not support versioning with annotation or migration from old formats to newer formats. Essentially, the principle outcomes of the assessments and evaluations presented in this paper converge to signify the need of standards, systems and further research to fully realise the needs of effective metadata management.

In conclusion, it may not be unreasonable to regard this MSc. Project as a complete success as all of its stages/phases and their constituting tasks were completed successfully within their allocated time scales (see Appendix H), meeting all of the project requirements in full and with credibly satisfactory results. Furthermore, considering the high significance of the main project subject, i.e. long-term metadata management in ensuring successful perpetuation of high quality data in environments where technologies change or evolve rapidly; and uniqueness or novelty of such endeavour as that of this project, this dissertation is expected to provide appropriate and useful guidance not only to the future work of this project, but also any future project or research that pertains to long-term metadata management and quality assurance in the context of digital curation. From the same perspective, this project may also be regarded as a very first or rather the only step till date towards developing an efficient and appropriate approach for long-term metadata management.

References & Bibliography

Sources from Books

[ATM01] Adrienne Tannenbaum, *"Metadata Solutions - Using Metamodels, Repositories, XML and Enterprise Portals to Generate Information on Demand"*, 2001, Addison-Wesley, page 145.

[CGK00] Cagle, K., "*XML Developer's Handbook"*, SYBEX Inc., San Francisco, 2000 CA, p. 272.

[GEG04] G E Gorman, *"International Yearbook of Library and Information Management, 2003-2004, metadata applications and management*", facet publishing, 2004, part 1, pages 1-17.

[JDG85] J. K. Ousterhout, H. D. Costa, D. Harrison, J. A. Kunze, M. Kupfer, and J. G. Thompson. *"A trace-driven analysis of the Unix 4.2 BSD file system. In Proceedings of the 10th ACM Symposium on Operating Systems Principles"* SOSP 1985, pages 15–24.

[MMC98] Murtha Baca, *"Introduction to Metadata, Pathways to Digital Information*", Getty Information Institute, 1998, Chapter 1 & 9.

[MBM00] D. Marco, "*Building and managing the metadata repository - a full lifecycle guide"*, Wiley, New York, 2000.

[PCM02] J. Poole, D. Chang, D. Tolbert, D. Mellor, "*Common warehouse metamodel - an introduction to the standard for data warehouse integration",* Wiley, New York, 2002.

[PDU97] Paul and Daniel Greenstein, *"Discovering Online resources Across the Humanities A Practical Implementation of the Dublin Core*", 1997, UKOLN.

Sources from World Wide Web

[AMP02] Alison Macdonald and Philip Lord, *"Digital Data Curation Task Force Report of the Task Force Strategy Discussion Day"*, November 2002 - http://www.jisc.ac.uk/uploaded_documents/CurationTaskForceFinal1

[ARC01] Bert Vermeij, *"Implementing European Metadata Using ArcCatalog"*, ESRI, Nederland, July 2001 – http://www.esri.com/news/arcuser/0701/metadata.html

[ASD04] *"ASDD geospatial metadata management: MetaStar Suite"*, 2004 – http://www.ga.gov.au/asdd/tech/manage-metastar.html

[ADA04] "*Activities – Digital Archiving"*, NLA, 2004 – http://www.nla.gov.au/initiatives/digarch.html#pres

[BSK01] Brian Matthews, Shoaib Sufi and Kerstin Kleese van Dam, *"The CLRC Scientific Metadata Model Version 1"*, CCLRC, February 2001, – http://www.dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf

[BLSNA] *"Best Practices in Metadata Management"*, Technical Paper, BellSouth: Metadata Services Group - http://www.wilshireconferences.com/award/2003/Submissions/BellSouth.pdf

[BJH03] Barton, J; Currier, S and Hey, J. "*Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice",* DC, 2003 - http://www.siderean.com/dc2003/201_paper60.pdf

[CPA96] Commission on Preservation and Access and Research Libraries Group Inc. (Commission) "*Preserving Digital Information: Report of the Task Force on Archiving of Digital Information",* 1996 - http://www.rlg.org/ArchTF/tfadi.index.htm.

[CMS02] Chuck Mosher, "*A New Specification for Managing Metadata"* April 2002 - http://java.sun.com/developer/technicalArticles/J2EE/JMI/

[CHN01] Canadian Heritage Information Network (CHIN),*"Metadata Standard*", 2001 http://www.chin.gc.ca/English/Standards/metadata_intro.html

[CLJ00] Catherine Lupovici and Julien Masanès, *"Metadata for long term-preservation*", Technical report, Nedlib, July 2000 - http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm

[CJJ03] Christopher Brooks, John Cooke, Julita Vassileva, *"Versioning of Learning Objects" Canada*, 2003 - http://csdl.computer.org/comp/proceedings/icalt/2003/1967/00/19670296.pdf

[CSD02] "*Content Standard for Digital Geospatial Metadata (CSDGM)", 2002* - http://www.fgdc.gov/metadata/contstan.html

[CGJ02] Craig A.N. Soules, Garth R. Goodson, John D. Strunk and Gregory R. Ganger, *"Metadata Efficiency in a Comprehensive Versioning File System",* May 2002 - www.pdl.cmu.edu/PDL-FTP/Secure/CMU-CS-02-145.pdf

[CMD02] Chad Berkley, Matthew Jones, Jivka Bojilova & Daniel Higgins, *"Metacat: a Schema-Independent XML Database System"*, 2002 - http://classweb.gmu.edu/kersch/inft864/Readings/SDBSystems/MetaCat.pdf

[CED02] "*Cedars Project"*, 2002 - http://www.leeds.ac.uk/cedars/

[DVN03] D. Sampson, V. Papaioannou,˜ N. Bassiliades, and I. Vlahavas, "*An Educational Metadata Management System using a deductive object oriented database approach",* 2003 - http://lpis.csd.auth.gr/publications/edmedia2002.pdf

[DBC01] *Scientific Data Curation and the Grid,* David Boyd CLRC e-Science Centre, 2001 – http://www.dpconline.org/graphics/ events/presentations/pdf/DavidBoyd.pdf

[DRB02] Denise R. Bleakly *"Long-Term Spatial Data Preservation and Archiving: What are the Issues?"* January 2002, Sandia National Laboratories - https://repository.lanl.gov/retrieve/141/bleakly-020107.pdf

[DCM04] DCMI Usage Board, "*Dublin Core Metadata Initiative Terms",* 2004 - http://dublincore.org/documents/dcmi-terms/

[DIH00] Diane I. Hillmann, "*Using Dublin Core"*, 2000 - http://dublincore.org/documents/2000/07/16/usageguide/

[DDINA] "*Data Documentation Initiative, About the specification"* http://www.icpsr.umich.edu/DDI/codebook/

[DAC03] "*The Digital Archiving Consultancy - archiving, long-term storage, preservation, curation of digital data and electronic information 21 CFR Part 11*", 2003 - http://www.d-archiving.com/index.htm

[DPW04] "*DCMI Preservation Working Group*", 2004 – http://dublincore.org/groups/preservation/

[EJR00] Jeff Rothenberg*, "An Experiment in using Emulation to Preserve Publication"*, The Koninklijke Bibliotheek Den Haag, April 2000 – http://www.kb.nl/nedlib/results/emulationpreservationreport.pdf

[ECO04] "*ASDD geospatial metadata management: Eco Companion document management service*", 2004 - http://www.ga.gov.au/asdd/tech/manage-ec.html

[EON01] Erik Oltmans, "*Metadata Interoperability*", 2001 – https://doc.telin.nl/dscgi/ds.py/Get/File-14690/interoperability.pdf

[EBI04] "*About the EBI*", 2004 - http://www.ebi.ac.uk/Information/index.html

[EDB04] "*EBI Databases - ArrayExpress Home*", 2004 - http://www.ebi.ac.uk/arrayexpress/

[GSM01] Gyo Sik Moon, "*Design and Implementation of Metadata Management System for WWW Coursewares", 2001* -http://fie.engrng.pitt.edu/fie2001/papers/1144.pdf

[GLS04] "*Describing Agency Information Resources Using Gils Core",* The Government Information Locator Service, 2004 - http://www.archives.gov/records_management/policy_and_guidance/gils.html

[GLD00] "*Definition for GILS mandatory core Elements*", Government Information Locator Service, 2000 - http://www.access.gpo.gov/su_docs/gils/fld.html

[GML04] Gene Major and Lola Olsen, "*A Short History of the DIF*", 2004 – http://helium.gsfc.nasa.gov/User/difguide/whatisadif.html

[GEM02] Gunnar Auth, Eitel von Maur and Markus Helfert, *"A Model-based Software Architecture for Metadata Management in Data Warehouse Systems",* Institute of Information Management, University of St. Gallen, 2002 - http://verdi.unisg.ch/org/iwi/iwi_pub.nsf/0/7D83210E1E92D8FBC1256BD7005A972D/$file/BIS02AuthHelfertvonMaur.pdf

[GEO98] "*New systems for Spatial Metadata Management*", GeoInfo, 1998 – http://www.govtech.net/magazine/gt/1998/nov/geoinfo/geoinfo.phtml

[GCM04] *Directory Interchange Format (DIF) Writer's Guide, Version 9*; 2004. Global Change Master Directory. National Aeronautics and Space Administration – http://gcmd.nasa.gov/User/difguide/.

[HEW00] Hong Hai Do, Erhard Rahm*, "On Metadata Interoperability in Data Warehouses",* March 2000 - http://dol.uni-leipzig.de/pub/2000-13

[HFH03] Heiner Stuckenschmidt and Frank van Harmelen, "*Generating and Managing Metadata for Web-Based Information Systems",* May 7, 2003 – www.cs.vu.nl/~frankh/postscript/KBS03.pdf

[HSS97] Marilyn Drewry, Helen Conover, Susan McCoy & Dr. Sara J. Graves, "*Metadata: Quality vs. Quantity*", 1997 NASA- http://www.computer.org/proceedings/meta97/papers/hconover/mdrewry.html?SMSESSION=NO

[IDC03] Dennis Callaghan, "*Informatica Adds Metadata Management*", Eweek, August 18, 2003, - http://www.eweek.com/article2/0,3959,1224825,00.asp

[ISC04] "*Informatica SuperGlue -- Visibility through Metadata Management*", Informatica Corporation, 2004 – http://www.informatica.com/products/superglue/superglue_overview_0104br1411b_lo.pdf

[ISM03] "*Intergraph-The SMMS Family of Metadata Solution Key Features*", Intergraph, 2003 – http://imgs.intergraph.com/smms/features.asp

[INT03] "*Intergraph-The SMMS Family of Metadata Solution*", Intergraph, 2003 – http://imgs.intergraph.com/smms/

[ILM02] "*IEEE Learning Object Metadata*", 2002 - http://ltsc.ieee.org/wg12/

[IGG02] David Lowe, "*The Principles of Good Metadata Management*", The IGGI Working Group on Metadata Implementation, January 2002 - *www.iggi.gov.uk/achievements_deliverables/pdf/Guide.pdf*

[ISO02] "*ISO 19115 - Geographic information – Metadata*", 2002 - http://metadata.dgiwg.org/standard/detail.htm

[IQM04] Marieke Guy and Michael Day, **"***Improving the Quality of Metadata in Eprint Archives*", Ariadne, January 2004 – http://www.ariadne.ac.uk/issue38/guy/

[JMS97] Jean-Pierre Kent and Maarten Schuerhoff, "*Some Thoughts About a Metadata Management System", 1997* Statistics Netherlands - http://www.vldb.org/archive/vldb2000/presentations/jarke.pdf

[JRB00] Jostein Ryssevik *"Bazaar Style Metadata in the Age of the Web - An 'Open Source' Approach to Metadata Development"*, Invited paper, 2000 - http://www.icpsr.umich.edu/DDI/papers/bazaar.pdf

[JCD01] J. Pollack, C. Gokey, D. Kendig, L. Olsen Goddard Space Flight Center, Greenbelt, MD, SSAI, Greenbelt, MD, *"Syntactic and Semantic Validation within a Metadata Management System",* GCMD, 2001 - http://helium.gsfc.nasa.gov/Aboutus/presentations/conferences/eogeo01/eogeo_01.html

[JSC03] *"Quality Assurance For Metadata",* QA Focus Document, QA Focus, a JISC-funded advisory service supporting JISC 5/.99 projects 2003 -http://www.ukoln.ac.uk/qa-focus/documents/ briefings/briefing-43/briefing-43-A5.doc

[JRD02] Jostein Ryssevik, "*The Data Documentation Initiative (DDI) metadata specification*", Nesstar Ltd, 2002 - http://www.icpsr.umich.edu/DDI/papers/ryssevik.pdf

[JJG02] Joe Futrelle and Jeff Gaynor, "*The NEESgrid Metadata Service API: Overview",* Whitepaper, 2002 -www.neesgrid.org/documents/MetadataService_v1_0.pdf

[KAD01] Kuula, Arja. *"The DDI and qualitative data"*, Amsterdam, Netherlands, May 2001 - http://datalib.library.ualberta.ca/conferences/2001/presentations/Kuula.ppt

[KSL03] Kimberly S. Lightle, *"Using Metadata Standards to Support Interoperability"*, 2003, Associate Director, Instructional Resources Eisenhower National Clearinghouse (ENC) - http://telr-research.osu.edu/learning_objects/documents/Lightle.pdf

[KGJ98] Keith G Jeffery*, "What's Wrong With Dublin Core?"*, (A Discussion Paper) CLRC-RAL, November 1998 – http://www.fou.uib.no/fou/grey_lit/dublincore98112.doc

[LDR97] Lorcan Dempsey and Rachel Heery, UKOLN, *"A review of metadata: a survey of current resource description formats"*, March 1997 - http://www.ukoln.ac.uk/metadata/desire/overview/

[LSMNA] Lloyd Sokvitne*, "An Evaluation of the Effectiveness of Current Dublin Core Metadata for Retrieval"* – http://www.vala.org.au/vala2000/2000pdf/Sokvitne.PDF

[MVD03] Mario Valle "*Scientific Data Management", 2003* - http://www.cscs.ch/~mvalle/sdm/scientific-data-management.html

[MDB03] "*MetaStar Digital Library Solution"*, Blue angel Technologies, April 2003 – http://www.blueangeltech.com/Solutions/MetaStar%20DLS.pdf

[MCMNA] "*Metabase - Cross-platform Metadata Management Solutionn"*, MetaMatrix - http://www.metamatrix.com/datasheets/metabase.pdf

[MMP03] "*Metadata Management at* PIK", PIK, 2003 –http://www.pik-potsdam.de/~cera/pikcera/welcome.html

[MIM04] "*Manage and integrate data across the extended enterprise with MetaMatrix enterprise data integration and data management technology"*, MetaMatrix, 2004 - http://www.metamatrix.com/l3_metabase.html

[MEM04] "*MetaMatrix System -- Enterprise Information Integration"*, MetaMatrix, 2004 – http://www.knowledgestorm.com/

[MZE96] Matthias Zingler, *"Architectural Components for Metadata Management in Earth Observation",* 1996 - http://www.computer.org/conferences/meta96/zingler/zingler.html

[MRS99] "*Request For Quotation, Provision of Metadata Repository and Search System, For The National Library of Australia"*, NLA, June 1999 - http://www.nla.gov.au/dsp/rfq/rfq.html

[NRC02] "*The NERC Data Grid"*, Project Proposal, 2002 - http://ndg.badc.rl.ac.uk/public_docs/NDG01_Proposal_Public.pdf

[NES04] "*NEESgrid:: Virtual Collaboratory for Earthquake Engineering"*, 2004 - http://www.neesgrid.org/index.php

[NDO00] "*Applying the OAIS Reference Model to the Deposit System for Electronic Publications (DSEP)"*, NEDLIB, June 2000 – http://www.kb.nl/coop/nedlib/results/OAISreviewbyNEDLIB.html

[NDI03] "*Welcome to NDIIP"*, Library of Congress, 2003 – http://www.digitalpreservation.gov/index.php

[NIS03] "*Data Dictionary - Technical Metadata for Digital Still Images - National Information Standards Organization (NISO)*", NISO, 2003 – http://www.niso.org/committees/committee_au.html#charge

[NED01] "*Nedlib homepage*", 2001 - http://www.kb.nl/coop/nedlib/

[OCS02] "*Reference Model for an Open Archival Information System (OAIS)*", Consultative Committee for Space Data Systems, Blue Book, January 2002 – http://www.ccsds.org/documents/650x0b1.pdf

[OPM02] "*Preservation Metadata and the OAIS Information Model A Metadata Framework to Support the Preservation of Digital Objects*", The OCLC/RLG Working Group on Preservation Metadata, June 2002 - http://www.oclc.org/research/pmwg/

[OMG04] *Catalog of OMG Modeling and Metadata Specifications*, 2004 - http://www.omg.org/technology/documents/formal/uml.htm

[OAI04] "*Open Archives Initiative*", 2004 - http://www.openarchives.org/

[OCL04] "*About OCLC*", 2004 - http://www.oclc.org/about/default.htm

[PMD96] Paul Miller, "*Metadata for the masses*", Dublin Core 1996 - http://www.ariadne.ac.uk/issue5/metadata-masses/#refs

[PAJ03] Philip Lord and Alison Macdonald, "*Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*", Consultation Draft, The JISC Committee for the Support of Research, (JCSR), 2003 - http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

[PSC97] Peter Churchill and Stuart Mills *"CIP - Catalogue Interoperability Protocol"*, 1997 - http://www.loc.gov/z3950/agency/profiles/cip.html

[PARNA] Peter Allan, *"Lessons from Earth Observation"*, Rutherford Appleton Laboratory – http://main.cs.qub.ac.uk/~fmurtagh/astro-grid-papers/peter-allan.pdf.

[PRM03] "*PREMIS (PREservation Metadata: Implementation Strategies)*", OCLC, 2003 - http://www.oclc.org/research/projects/pmwg/default.htm

[PWG99] "*Preservation working group*", NLA, 1999 – http://www.nla.gov.au/preserve/pmeta.html#ref8

[RDF04] "*Resource Description Framework (RDF) Model and Syntax Specification*", 2004- http://www.w3.org/TR/rdf-primer/

[RCM98] Robert Craig, *" Metadata Management",* ENT News Archive Article, October 1998 - http://www.entmag.com/archives/article.asp?EditorialsID=3588

[RSL94] Robyne M. Sumpter Lawrence, *"Whitepaper on Data Management",* , Livermore National Laboratory February 10, 1994 Version 1.0 - http://www.llnl.gov/liv_comp/metadata/papers/whitepaper-draft.html

[RSM01] Ronald Snijder, *"Metadata Standards and Information Analysis, A Survey of Current Metadata Standards and the Underlying Models",* 2001- http://www.geocities.com/ronaldsnijder

[RCD97] Roger Clarke, "*Beyond the Dublin Core: Rich Meta-Data and Convenience-of-Use Are Compatible After All"*, 11 July 1997 - http://www.anu.edu.au/people/Roger.Clarke/II/DublinCore.html

[RED01] Addendum to Final Report "*Electronic Preservation of Data Documentation: Complementary SGML and Image Capture"* SBR-9617813 Results of the Evaluation of the Data Documentation Initiative (DDI), 2001, - http://www.icpsr.umich.edu/DDI/PAPERS/evalsummary.pdf

[RMX02] Ruixin Yang, Menas Kafatos, and X. Sean Wang, "*Managing Scientific Metadata Using XML", George Mason University, 2002- www.computer.org/internet/ic2002/w4052abs.htm

[RLG04] "*RLG Home"*, 2004 - http://www.rlg.org/

[SPMNA] Suresh Purusothaman & Asimkumar Munshi - *"Leveraging Business Objects As A Metadata Managment Solution"*, technical paper- http://www.dmreview.com/whitepaper/WID1101.pdf

[SKR03] Shien-Chiang Yu, Kun-Yung Lu and Ruey-Shun Chen *"Metadata management system: design and implementation"*, Emerald, 2003 - http://www.emeraldinsight.com/0264-0473.htm

[SMT96] Susan Stitt and Maurice Nyquist & Anne Frondorf, *"Development of a Metadata Content Standard for Biological Resource Data National Biological Information Infrastructure Draft Metadata Standard",* 1996 - http://www.computer.org/conferences/meta96/frondorf/ieeemeta.html

[SJR03] Sarah Currier, Jane Barton, Rónán O'Beirne and Ben Ryan, "*Quality Assurance for Digital Learning Object Repositories: Issues for the Metadata Creation Process",* 2003 - http://metadata.cetis.ac.uk/files/currbartobeiryan_altj_6.doc

[STN02] Joe Futrelle, *"Storage Technologies for the NEESgrid Curated Data Repository"*, Technical Rport, NeesGrid, 2002- http://www.neesgrid.org/documents/StorageTechnologies_v1_0.pdf

[TRE99] Thomas Stohr, Robert Muller and Erhard Rahm, *"An Integrative an Uniform Model for Metadata Management in Data Warehousing Environment",* 1999 - http://dol.uni-leipzig.de/pub/1999-22

[UDA04] "*The UK Data Archive"*, September 2004 - http://www.data-archive.ac.uk/home/

[ULP04] "*Metadata Management"*, University of Leipzig, 2004 – http://dbs.uni-leipzig.de/en/Research/meta.html

[WEC98] William E. Moen, Erin L. Stewart and Charles R. McClure, "*Assessing Metadata Quality: Findings and Methodological Considerations from an Evaluation of the U.S. Government Information Locator Service (GILS)",* 1998 - http://csdl.computer.org/comp/proceedings/adl/1998/8464/00/84640246abs.htm

[XML04] "*Extensible Markup Language (XML) 1.0 (Third Edition)"*, W3C Recommendation 04 February 2004 - http://www.w3.org/TR/REC-xml/

Appendix A: Data (Digital) Curation

Within different information domains, the phrase "Data Curation" has different interpretations. From the museum perspective data curation covers three core concepts – data conservation, data preservation and data access. Data access in this sense may imply preserving data and making sure that the people to whom the data is relevant can find it - that access is possible and useful. Another interpretation of the phrase "Data Curation" may be an active management of information, involving planning, where re-use of the data is the core issue [AMP02].

Therefore, in essence, Data or Digital curation is the continuous activity of managing, improving and enhancing the use of data or other digital materials over their life-cycle and over time for current and future generations of users, in order to ensure that its suitability sustains for its intended purpose or a range of purposes and it is available for "discovery" and re-use. One of the curation activities is Archiving, which ensures that data is properly selected, stored and remains accessible over time by maintaining its logical and physical integrity as well as providing security and authenticity as required [PAJ03].

Studies and researches have indicated that the curation of data assists in maximizing the potential of data by facilitating research, increasing its quality and extending the knowledge base through annotation, links and visibility. However, without the perception of benefit, digital curation could stay grounded, yet benefit can only be demonstrated by actually performing digital curation over a sustained period of time. The figure below illustrates typical activities within digital curation environment for data produced from scientific research.

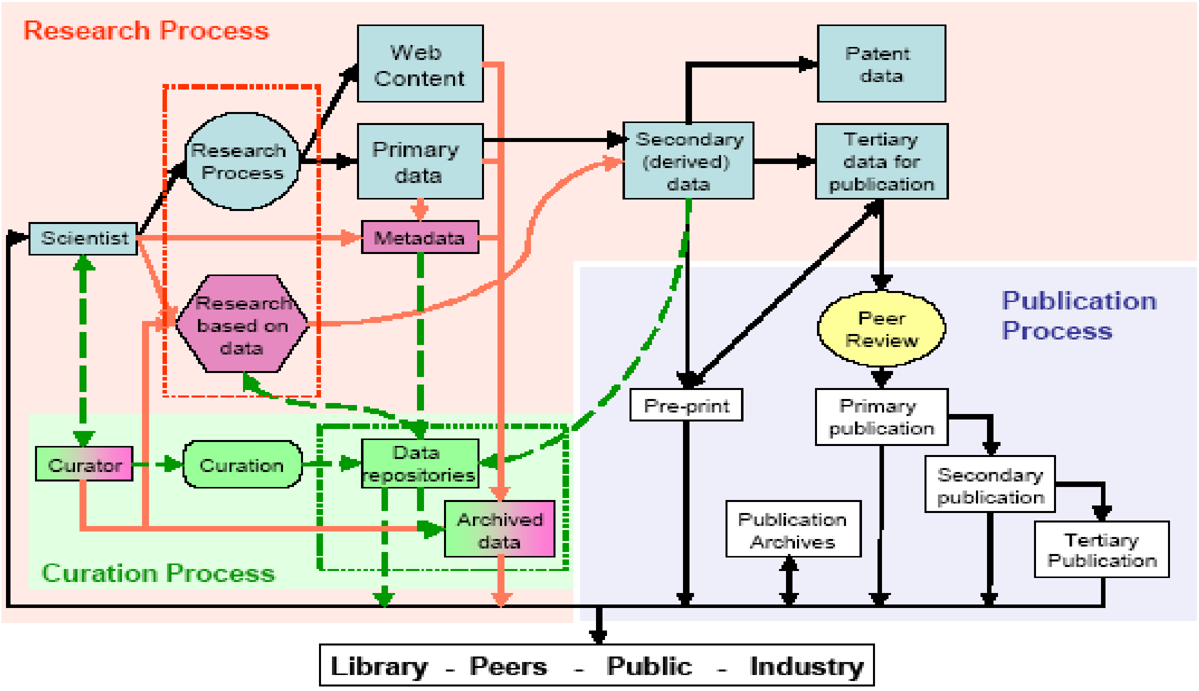


Figure A.1: An example of Digital Curation environment [PAJ03]

Effectively, long-term metadata management is an integral part of long-term data curation, which is therefore the main context of this project.In light of the above construal of digital preservation, Metadata curation may be defined as an inherent part of a digital curation process for the continuous management of metadata (which involves its creation and/or capturing as well as assuring its overall integrity) over the life-cycle of the digital materials that it describes in order to ascertain its suitability for facilitating the intelligent, efficient and enhanced discovery, retrieval, use and preservation of those digital materials over time.

Appendix B: Ancillary Information about Different Metadata Standards

B1: Elements of Dublin Core Metadata Standard [PDU97]

Element Name	Element Descriptions
Title	Typically, a Title will be a name, given to the resource by the CREATOR or PUBLISHER and by which the resource is formally known.
Identifier	An unambiguous reference, typically a string or a number to the resource within a given context, such as accession number, ISBN number, or URL.
Publisher	An entity responsible for making the resource available in the present form. Examples of a Publisher include a person, an organisation, or a service.
Creator	An entity primarily responsible for making the content of the resource. Examples of a Creator include a person, an organisation, or a service.
Contributor	A person or organization in addition to those specified in the CREATOR element that contributed to the creation of the object in a secondary role, such as an editor, illustrator, translator, fabricator, or sponsoring organization.
Date	The date the resource was made available in its present form. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format.
Coverage	Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity).
Subject	Keywords or key phrases to describe the subject or historical association of the resource.

Relation	Relationship to other resources, such as ‘is part of, is a version of, is a reproduction of, is a format of."
Type	Type includes terms describing general categories, functions, genres, or aggregation levels for content, such as text, sound recording, physical object, image, or collection.
Format	The physical or digital manifestation of the resource, such as text/html, ASCII, Postscript file etc.
Description	A textual description of the content of the resource and may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
Rights	Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information.
Language	A language of the intellectual content of the resource.
Source	A Reference to a resource from which the present resource is derived.

Table B.1: 15 elements of Dublin Core Metadata Standard

B2: Elements of Content Standards for Digital Geospatial Metadata standards (Metadata-Lite) [HSS97]

Elements		
Identity of this entry Originator Publication date Title of data set Citation information Presentation form Online linkage Abstract Purpose Supplemental information	Beginning date Ending date Currentness reference Progress Data set maintenance/update frequency West bounding coordinate East bounding coordinate North bounding coordinate South bounding coordinate Theme keywords	(Theme) reference Place keywords (Place) reference Limits on data accessibility Limits on use of data Browse graphic URL Browse graphic caption Browse graphic file type Spatial data type Distribution organization

Table B.2. CSGDM Metadata "Lite": These attributes indicate minimal standards for CSGDM.

B3: Sections of Data Documentation Initiative (DDI) Metadata Specification [JRD02]

Sections	Sub Sections	Description
The Document Description	**Citation** **Guide to the documentation** **Documentation status** **Documentation source**	This section contains elements that consist of bibliographic information describing the metadata document and the sources that have been used to create it.
The study description	**Citation** **Study scope** **Methodology and processing** **Data access** **Other study description materials**	The elements in this section contain information about the data collection
The Data Files Description	**File description** **Notes**	This section contains elements to describe each single file of a data collection (formats, dimensions, processing information, missing data information etc.)
The variable description	**Variable group** **Variable** **NCube (added in version 1.02)** **Notes**	Describe each single variable in a data file (format, variable and value labels, definitions, question texts, imputations etc.).
Other Study-Related Materials		Includes references to reports and publications, other machine-readable documentation that are relevant to the users of the study (referenced by URI's) etc.

Table B.3: Different Sections of DDI metadata specification

B4: Elements of the Global Information Locator Service (GILS) metadata standards [GLD00]

Elements	Sub Elements	Description
Abstract		Presents a narrative description of the information resource.
Access Constraints		It describes any constraints or legal prerequisites for accessing the information resource or its component products or services.
Agency Program		This element identifies the major agency program or mission supported by the system and should include a citation for any specific legislative authorities associated with this information resource.
Availability	**Distributor, Resource Description, Order Process, Technical, Prerequisites, Available Time Period, Available Linkage, Available Linkage Type**	This element is a grouping of sub elements that together describe how the information resource is made available.
Control Identifier		This element is defined by the information provider and is used to distinguish this locator record from all other GILS Core entries.
Controlled Vocabulary	**Index Terms-Controlled Thesaurus**	This element is a grouping of sub elements that together provide any controlled vocabulary used to describe the resource and the source of that controlled vocabulary.
Cross Reference	**Cross Reference Title, Cross Reference, Linkage, Cross Reference Type**	This element is a grouping of sub elements that together identify another locator record likely to be of interest.
Date of Last Modification		This element identifies the latest date on which this locator record was created or modified.
Local Subject Index		This element is a grouping of descriptive terms to aid users in locating resources of potential interest, but the terms are not drawn from a formally registered controlled vocabulary source. Each term is provided in the repeating sub element: Local Subject Term.

Methodology		This element identifies any specialized tools, techniques, or methodology used to produce this information resource. The validity, degree of reliability, and any known possibility of errors should also be described.
Original Control Identifier		This element is used by the record source to refer to another GILS locator record from which this locator record was derived.
Originator		This element occurs once per locator record. It identifies the information resource originator, named as in the U.S. Government Manual where applicable.
Point of Contact for Further Information	**Contact Name** **Contact Organization** **Contact Street Address** **Contact City** **Contact State** **Contact Zip Code** **Contact Country** **Contact Network Address** **Contact Hours of Service** **Contact Telephone** **Contact Fax**	This element identifies an organization, and a person where appropriate, serving as the point of contact plus methods that may be used to make contact
Purpose		This element describes why the information resource is offered and identifies other programs, projects, and legislative actions wholly or partially responsible for the establishment or continued delivery of this information resource. This description may include the origin and lineage of the information resource, and related information resources.
Record Source		This element identifies the organization, as named in the U.S. Government Manual that created or last modified this locator record.
Sources of Data		This element identifies the primary sources or providers of data to the system, whether within or outside the agency.

Spatial Reference	**Bounding Coordinates** **Geographic Name**	This element is a grouping of sub elements that together provide the geographic reference for the information resource.
Schedule Number		This element is used to record the identifier associated with the information resource for records management purposes. *Mandatory when the GILS Core entry is intended to meet the obligation of Federal agencies to inventory automated information systems or other records series for records management purposes
Supplemental Information		Through this element, the record source may associate other descriptive information with the GILS Core entry
Time Period of Content		This element provides time frames associated with the information resource
Title		This element conveys the most significant aspects of the referenced resource and is intended for initial presentation to users independently of other elements. It should provide sufficient information to allow users to make an initial decision on likely relevance. It should convey the most significant information available, including the general topic area, as well as a specific reference to the subject
Use Constraints		This element in some cases may contain the value "None." It describes any constraints or legal prerequisites for using the information resource or its component products or services. This includes any use constraints applied to assure the protection of privacy or intellectual property and any other special restrictions or limitations on using the information resource.

Table B.4: Elements of GILS metadata Standard

B5: Elements of The Global Change Master Directory (GCMD)'s Directory Interchange Format (DIF) [GCM04]

Elements		
Entry ID *	Data Set Citation	Keyword
Entry Title*	Instrument	Quality
Science Keywords*	Platform	Access Constraints
ISO Topic Category*	Temporal Coverage	Use Constraints
Data Center *	Paleo-Temporal Coverage	Data Set Language
Summary *	Data Set Progress	Originating Center
Personnel	Spatial Coverage	Distribution
Related URL	Location	Multimedia Sample
Parent DIF	Data Resolution	Reference
Metadata_Name	Project	Discipline
Metadata_Version	DIF Creation Date	IDN Node
	Last DIF Revision Date	DIF Revision History
		Future DIF Review Date

Table B.5: GCMD DIF Attributes. Required fields are marked with '*'

B6: Metadata Categories of CLRC Scientific Metadata Model [BSK01]

Metadata Category	Description
Topic	A set of keywords relevant to the particular study, describing the subject domain with which it is concerned.
Study	The type of entry, which this metadata description is capturing. Description of the study within which the dataset has been generated. Includes investigator, experimental conditions, and purpose.
Access Conditions	Access rights and conditions on the data referred to within this entry. Includes ownership and access control information.
Data Description	The data description maintains the description of the data itself.
Data Location	Gives details on the location of the data sets together with any copies or mirrors etc.
Related Material	Contextual information associated with the resource being described.

Table B.6: Different Categories of CLRC Scientific Metadata Model and their descriptions

Appendix C: Other Reviewed Metadata Standards & Formats

C1: ISO 19115, Geographic information – Metadata

The ISO 19100 series is a multi-part International Standard for Geographic Information that is being developed by Technical Committee 211 Geographic information/Geomatics of the International Organisation for Standardisation (ISO). ISO 19115, Geographic information – Metadata is part of the ISO 19100 series. The objective of this International Standard is to provide a clear procedure for the description of digital geographic datasets so that users will be able to determine whether the data in a holding will be of use to them and how to access the data [ISO02].

Supplementary benefits of this standard for metadata are to facilitate the organization and management of geographic data and to provide information about an organization's database to others. In general, this standard furnishes those unfamiliar with geographic data the appropriate information to characterize their geographic data and it makes possible dataset cataloguing enabling data discovery, retrieval and reuse [ISO02]. The ISO 19115 appears to use the same approach for metadata versioning as the CSDGM – with the help of "Metadata_Data_Stamp" element (non-repeatable) for the date that the metadata was created. However, the standard does not appear to contain any metadata version or update element.

C2: Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

The OAI protocol for metadata harvesting provides a standard mechanism for sharing metadata over the Internet. The underlying concept of harvesting is that the participants agree to take small efforts that enable some basic shared services, without being required to adopt a complete set of agreements. The protocol is an open, freely available standard created by the Open Archives Initiative. It supports any metadata schema, with the base schema being simple unqualified Dublin Core. It is relatively easy to implement and supports both data providers and service providers in the creation of federated discovery services or portals based on aggregated and shared metadata [OAI04].

C3: Learning Objects Metadata (LOM)

The Learning Object Metadata (LOM) metadata specification was developed by the Learning Technology Standards Committee (LSTC) of the Institute of Electrical and Electronic Engineers (IEEE) and became an approved standard in June 2002. The purpose of the development of the LOM standard is to facilitate search, evaluation, acquisition, and use of learning objects (see) by learners or instructors. It also aims to facilitate the sharing and exchange of learning objects, by enabling the development of digital libraries and catalogues, so that users can create and publish educational material.

The standard has a wide variety of elements (almost 80 elements and sub-elements) that can describe digital objects in details. These elements are extensible and have a status of obligatory (must be present) or optional (maybe absent). It should be noted that similar to Dublin Core, LOM is unable to address or capture syntactical changes in different version of learning objects, making it impossible to provide common version control features (e.g. roll-backs), which consequently restricts the semantic understanding of versioning change to the comparison of metadata records [ILM02].

C4: Resource Description Framework (RDF)

The Resource Description Framework (RDF), developed under the auspices of the World Wide Web Consortium (W3C), is an infrastructure that enables the encoding, exchange, and reuse of structured metadata. It is an application of XML (see below, C5) that imposes needed structural constraints to provide unambiguous methods of expressing semantics. The RDF additionally provides a means for publishing both human-readable and machine-processable vocabularies designed to encourage the reuse and extension of metadata semantics among disparate information communities. The RDF is built upon a simple but robust data model that allows resources to be described in terms of their properties. As a result, the RDF provides a flexible architecture for managing diverse, application-specific and machine generated metadata records [RDF04].

In short, the RDF does not address the issues of long-term metadata management. The techniques for managing the RDF elements, such as performing version control, managing the modifications, recording source information, and so on, are somewhat independent of RDF. However, once determined, these techniques can readily be used with the RDF.

C5: eXtensible Markup Language (XML)

The eXtensible Markup Language (XML) is defined by the World Wide Web Consortium as a successor to Hyper Text Markup Language (HTML). It is used to create information objects consisting of elements. The elements are encoded using tags and attributes. Contrary to HTML tags are only used to define the structure of documents, and not the layout. While it is possible to freely define tags XML also gives the opportunity to define strictly ruled applications such as the RDF. The tags function as 'containers' for digital data, regardless of the format of that data [XML04].

However, similar to all of its other predecessors, XML does have some shortcomings. For example, it does not have inherent capabilities to model complex relationships, such as inheritance, association and aggregation, which play very important roles in representing different relationships between data objects as well as in their subsequent implementations.

C6: Document Type Definition (DTD)

The Document Type definition or DTD is a set of rules primarily to define and regulate the structure of an XML document. With DTD, applications and parsers can verify the validity of XML documents and authoring tools can generate XML documents. The syntax of DTD is hard to learn, not to mention the insufficiency of metadata definitions. For instance, DTD handles only text format data, not including the declarations of other formats; DTD provides only the declaration of the default value for attribute field, not the element field; DTD cannot treat an XML document as an object type for redirection [CGK00].

C7: XML Schema

XML schema is an XML-based schema (or metadata) description language that actually provides two pieces of critical data: a definition of the acceptable structure of the elements that make up a valid type of XML document and a representation of the data type used by the document. XML schema is not only an attempt to simplify existing schemas but also it is an effort to create a language capable of defining the set of constraints of any possible data resource [SKR03].

C8: Metadata Object Facility (MOF)

Metadata Object Facility (MOF) is an industry-endorsed standard for Metadata management approved Object Management Group (OMG) in 1997. MOF is an extensible model driven integration framework for defining (using Unified Modelling Language, UML - see C10), manipulating and integrating Metadata and data in a platform independent manner. MOF-based standards, such JMI etc. are in use for integrating tools, applications and data [CMS02].

In general, the MOF standard does not address the issues of long-term metadata management in direct terms as it is more focused on business oriented data management & interoperability rather than successful perpetuation of data. However, the layered architecture consisting of data, models metamodels and a single meta-metamodel that the standard provides may well lend valuable guidance for developing a specialised standard for long-term metadata management.

C9: Common Warehouse Metamodel

The Common Warehouse Metamodel is a standard for describing technical and business metadata in the domain data warehousing and business intelligence. The CWM is hosted by industry consortium Object Management Group (OMG) with an aim to interchange metadata between different tools and repositories. It can also be used for building active object models for storing and maintaining metadata. The CWM is founded on the UML (see C10) metamodel and extends it with specific meta-classes and meta-relationships for modelling data lineages found in the data-warehousing domain. Thus, it provides a complete specification of syntax and semantics necessary for interchanging shared metadata [GEM02]. As the MOF standard (see C8) the CWM also does not tackle issues like those associated with long-term metadata management.

C10: Unified Modelling Language (UML)

The Unified Modelling Language (UML) is a specification defining a graphical language for visualising, specifying, constructing, and documenting the artefacts of distributed object systems. The latest version of UML, 1.5 incorporates Action Semantics, which adds to UML, the syntax and semantics of executable actions and procedures, including their run-time semantics [OMG04].

Appendix D: Data Warehouse & Repository

D1: Data Warehouse

A typical data warehouse environment consists of file systems and DBMSs managing the operational sources, the data warehouse and data marts. Furthermore, a variety of tools from different vendors are usually involved for data modeling, ETL tasks (extraction, transformation, loading), and data access (OLAP, querying, etc.). All of these components create and maintain metadata, e.g. within database catalogues, dictionaries or tool-specific repositories. Typically metadata is maintained independently in a largely isolated way and in specific representation formats [HEW00]. The figure below reflects this description of a typical data warehouse environment.

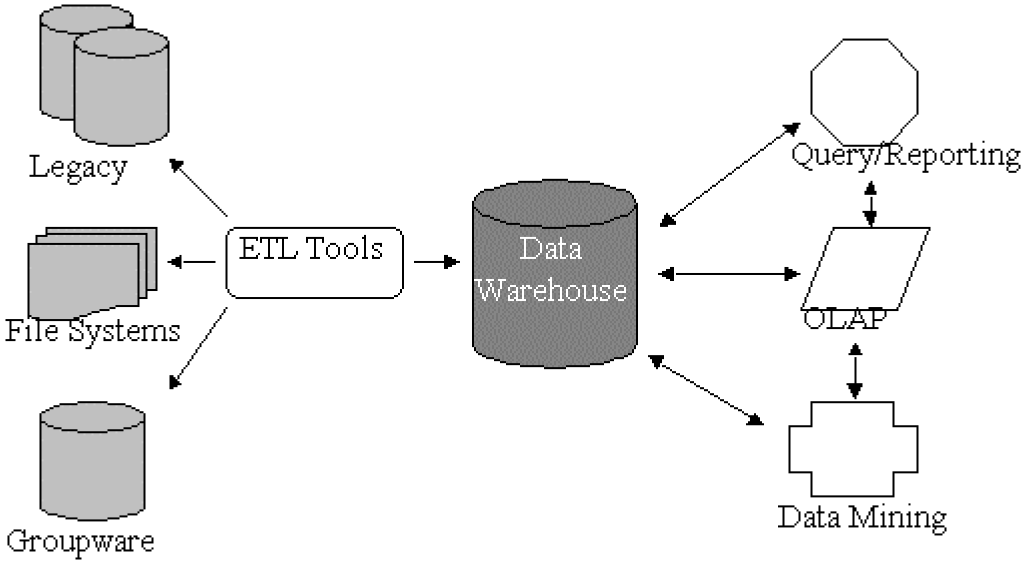


Figure D.1: A typical Data Warehouse Environment [CMS02]

D2: Repository

A repository is a specific database application for managing, analysing and providing metadata [PCM02]. In contrast to ordinary database management systems the content in a repository can only be accessed through certain repository services. Granting repository access only through specified repository services is a means to secure concurrent access without conflict, maintenance of metadata integrity and consistency, as well as error recovery [MBM00]. Important features of commercial repository products usually include version and configuration management for repository elements. Figure D.2 shows the conceptual architecture for interchanging metadata via central metadata repository.

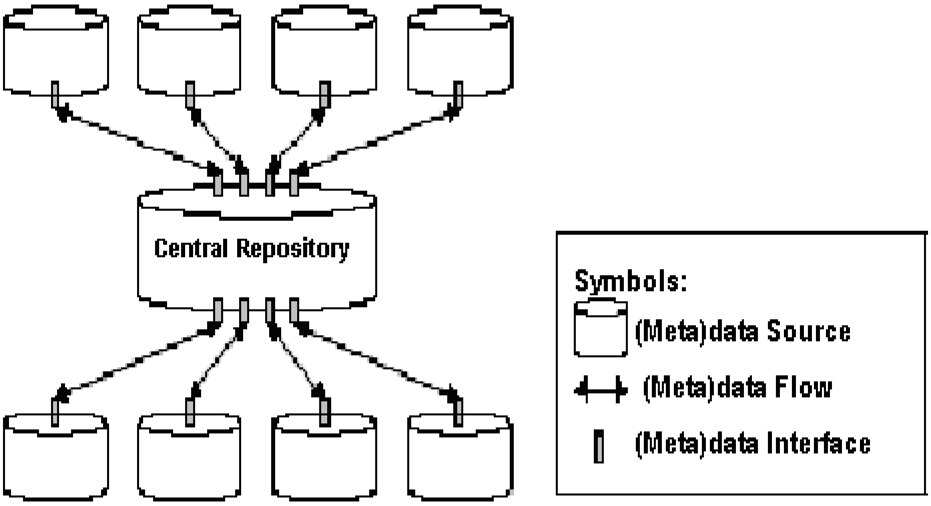


Figure D.2: Interchanging metadata via central metadata repository [GEM02]

Appendix E: Summary of Other Related Published Works

E1: Generic Metadata Management

In addition to those detailed in the main report, the research studied two other research papers that address the issues of generic metadata management. One of them is a project report titled, "**Efficient Metadata Management in Large Distributed Storage Systems**", the result of combined effort by Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long and Lan Xue of Storage Systems Research Center University of California, Santa Cruz of USA. The paper presents a new approach called Lazy Hybrid (LH) metadata management that combines the best aspects of Directory sub-tree partitioning and pure hashing, which are two common techniques used for managing metadata in large distributed storage systems, while eliminating their shortcomings, such as bottlenecks at very high concurrent access rates [JDG85].

Another research-oriented publication is a paper written by Heiner Stuckenschmidt and Frank van Harmelen of Department of Mathematics and Computer Science Vrije University Amsterdam, The Netherlands [HFH03]. The paper, titled**, "Generating and Managing Metadata for Web-Based Information Systems**", published in May, 2003, introduces the Spectacle Approach[66] for the validation of semi-structured information that can be used to check the completeness and consistency of metadata of an existing information system, called BUISY[67], with respect to the information it describes.

In short, this approach is mainly aimed at validating the metadata embedded within web pages and is done on the basis of rules which must hold for the information found in the Web site, both the actual information and the metadata (and possibly their relationship). As described in this research report, this approach enables locating pages with missing metadata, compare information contents and metadata as well as producing hints towards missing keywords. As far as the main objectives of this MSc. project are concerned, this metadata validation approach may not be quite efficient for ensuring metadata quality (during creation or preservation process) over long-term. Nevertheless, this validation technique may certainly be considered as a probable approach for ensuring the quality of metadata embedded within the web pages of the web interface or portal (if any) of the archive or warehouse, preserving information from obsolescence, hence preventing its extinction.

[66] This approach has been implemented in the Spectacle content management tool, developed by the Dutch company AIdministrator (www.aidministrator.nl).

[67] An environmental information system for the city of Bremen that has been developed by the Center for Computing Technologies of the University of Bremen in cooperation with the public authorities.

E2: Data Warehouse Metadata Management

The research report, titled "**An Integrative an Uniform Model for Metadata Management in Data Warehousing Environment",** identifies the increasing complexity of data warehouses and determines the necessity of a centralised and declarative management of Metadata for data warehouse administration, maintenance and usage. Based on this ground, this report, written by Thomas Stohr, Robert Muller and Erhard Rahm of Institute of Informatics department, University of Leipzig in Germany, describes an extensive research scenarios for developing a uniform and integrating model for data warehouse Metadata. This model uses a uniform representation approach based on the Uniform Modelling Language UML [TRE99].

E3: Scientific Metadata Management

In 2002, Chad Berkley, Matthew Jones, Jivka Bojilova & Daniel Higgins Of National Center for Ecological Analysis and Synthesis (NCEAS), University of California, Santa Barbara (XML) undertook a project to design and implement a schema-independent data storage system for XML which is called Metacat. Their project report, titled "**Metacat: a Schema-Independent XML Database System**" describes the Metacat XML data storage system and its relevance to scientific data management in the ecological sciences [CMD02].

In generic terms, the system described in this paper does not address the issues of long-term metadata management. Nevertheless, there are a few good features of the systems, which are worth mentioning and may be seen as useful for generic metadata management purposes. One of these features is the hybrid approach used in Metacat to store XML documents gives the flexibility of a dedicated XML database coupled with the enterprise features of a commercial Relational Database Management System. In addition, Metacat's replication mechanism allows a dispersed community to bring their data together into a central searchable location, yet it allows dataset owners to retain autonomous control. Above all, each features of Metacat has been designed with maximum flexibility in mind, hence, eliminating the restriction on its use only within the ecological community.

Another scientific metadata oriented publication is a research article by Matthias Zingler Department of Remote Sensing Exploitation, European Space Agency - ESRIN (Italy), Casella Postale 64, I-00044 Frascati. The article, titled **"Architectural Components for Metadata Management in Earth Observation**"**,** discusses and proposes several themes and planning criteria, including main requirements for a metadata management system, for Earth Observation metadata systems. The paper also outlines a generic architecture for metadata management system as well as mentioning recognised efforts within the related problem domain [MJE96].

Although, the paper discusses potentially efficient mechanisms for effective and enhanced access, namely *Thematic Mapping* [68] to Earth Observation (EO) metadata, it does not appear to provide any formation as to how the quality of these EO metadata will be

[68] Recognizes that the query for metadata matches a thematic pattern and follows an index mechanism to the leaves of the index tree, which made of two layers, the thematic layer organizing the different application typical queries into an index structure and the spatial and temporal layer reflecting the common reference properties of metagranules (i.e. Metadata describing small set of data - granule).

maintained for long periods of time in order to facilitate long-term management of high quality data.

E4: Metadata Model Management Approach

A common and standard way to represent Metadata is to model it. A well-defined model will have very precise definitions of what the features and attributes of particular model instances (the metamodels) mean. These precise definitions will then allow one to define exact and unambiguous mappings of the model features to particular languages and interchange formats. Example of model may be XML Schema, UML (see Appendix C) etc [CMS02].

Model Management is a powerful approach to generic Metadata management not limited to a specific language or application domain. Models and mappings between these models (SQL view definitions, XSLT transformations, XML-to-relational shredding specifications, ER-to-SQL DDL mappings, etc.) are manipulated using high-level algebraic operators, such as Match, Merge, or Compose. These operators are applied to models and mappings as a whole rather than to their individual building blocks. This approach, proposed by Phil Bernstein of Microsoft Corporation (One Microsoft Way Redmond, WA 98052-6399) promises to make the programming of metadata-intensive applications substantially easier [CMS02].

E5: Business-Oriented Metadata Management

One of American leading telecommunication industries, Bell South[69], addresses the issues metadata versioning and quality assurance within its metadata repositories. As described in a technical paper, titled **"Best Practices in Metadata Management**" published by Bellsouth Corporations' Metadata Service Group (MSG), Bellsouth facilitates metadata versioning by providing a special repository of information, called "Bronze" system [BLSNA]. Users that need the versioning utility are provided with access to this repository. Dependence metadata is also utilised for new versions of information. If a new reusable component is released, the dependency of metadata is extracted in order to determine if notification is required to the user base or if the new version can stand-alone. However, it is not quite clear from this description, how the actual metadata versioning process works.

For the purpose of metadata quality assurance, Bellsouth metadata within a repository go through a review and quality assurance check. For example, before an enterprise database is scanned for a new release the logical model, physical model and Oracle tables are reviewed in order to ensure they are in synchronization with each other. It should also be noted that the metadata used or provided by Bellsouth conforms to Dublin Core Metadata Standard (see section 3.2).

[69] BellSouth Corporation, Metadata Services Group, 754 Peachtree Street, Atlanta, GA 30308-1206, URL: http:\\www.bellsouth.com or http:\\www.bellsouth.net

E6: Educational Metadata Management

In an attempt to address Metadata management issues is within a Grid systems, George Samaras, Kyriakos Karenos, and Eleni Christodoulou of Department of Computer Science, University of Cyprus published a paper in 2001, detailing their work that contributes to the effort of enhancing current Grid technologies to support semantic descriptors for resources (termed also the *Semantic Grid).* In essence, the paper, titled **"A Grid Service Framework for Metadata Management in Self e-Learning Networks**" proposes a set of services that are applicable in such a case in alignment to the *Open Grid Services Architecture (OGSA)* for metadata management in Self e-Learning Network (Se-LeNe) that concentrates on providing services for the utilization of Learning Objects' (LO) (see 4.6) metadata.

E7: Metadata Quality Assurance

The research studied an online article titled, **"Improving the quality of Metadata in Eprint Archives",** written by Marieke Guy, Andy Powell & Michael Day of the UKOLN, that outlines a number of quality assurance procedures that may improve the quality of metadata in Eprint[70] archive [IQM04]. Although the overall focus of these quality assurance mechanisms lies on ensuring metadata quality during its creation process, the article does provide some logically effective post-creation metadata quality control techniques, which may be used as guidelines for long-term metadata management. While the pre-creation quality assurance processes include defining suitable metadata elements, controlled vocabularies, metadata encoding schemes, cataloguing guidelines on the basis of core functional requirements of the archive, the post-creation mechanism involves randomly sampling metadata entered into the archive and testing its effectiveness using a commercially available visual graphical analysis tool called Spotfire DecisionSite. In addition, the random sample of metadata is suggested to be subject to further assessment on the basis of:

- How often the document author has to amend automatically generated metadata.
- How often information specialists need to modify the metadata supplied by the document author, with steps being taken to improve the cataloguing guidelines and metadata entry tools being offered to the document authors as a result.

Furthermore, the article also recommends usability tests to be performed to ensure that end-users are able to undertake the activities specified in the initial functional requirements. The results of these post-creation quality control processes are fed back into the system through redesign of the metadata elements, encoding schemes etc. in order to improve overall quality of metadata.

[70] Dedicated to opening access to the refereed research literature online through author/institution self-archiving – http://www.eprint.org

Appendix F: Other Reviewed Metadata Management Systems

F1: ArcCatalog

The ArcCatalog is a CSDGM & ISO 19115 compliant metadata editor developed by Environmental Systems Research Institute (ESRI)[71] Australia. It is an application, included with the ArcGIS[72]; that allows automatic creation and maintenance of metadata. The ArcCatalog automatically attaches metadata to the data set to ensure integrity. When the data changes—for example, when a new attribute has been added, the ArcCatalog automatically updates it with the new information. All metadata created by the system are stored in XML format [ARC01].

The question that arises when assessing its suitability for the purposes of long metadata management is how this automated metadata management functionality ensures the quality of metadata during its creation. A brief answer to this question is that the automatic metadata creation and update functionality of ArcCatalog, though seemingly efficient, may not be able to ensure both syntactic and semantic accuracy of the metadata as it relies only on Document type Definition (DTD) validation of XML documents for such purposes. It should also be noted that this DTD validation is an optional functionality within ArcCatalog, which is only capable of checking the syntactic correctness of the metadata to certain extent. Besides, the ArcCatalog does not provide relevant functionality to address the issues associated with metadata versioning. In addition, the metadata storage facility of this tool is not capable of preserving metadata over the long-term.

However, on the positive side of the ArcCatalog, it offers an open and extensible architecture, which provides a powerful framework for building a custom environment to capture metadata. In addition, within the ArcCatalog, metadata searching can be performed using a simple metadata browser in a distributed environment.

F2: MetaStar Suite

Developed by the Blue Angel Technologies in Australia, this software provides innovative solutions for delivering and managing information on the Internet or Intranet [ASD04]. The MetaStar Suite provides a number of metadata management features, such as, Real-time entry, update, deletion of metadata records from a Web browser, harvesting of metadata from HTML records or files etc. In addition, it is capable of generating and maintaining relational database tables automatically. It provides interoperability between XML schema (metadata) originating from disparate sources. Furthermore, this tool supports metadata of numerous standards, including GILS, CSDGM, DIF (Chapter 3) etc. and locally defined standards. It also overcomes the language barriers by supporting for multiple languages through Unicode. Compatibility between different operating systems, environment

[71] ESRI - The GIS Software Leader - http://www.esri.com/index.html

[72] The ArcGIS is an integrated collection of GIS software products for building a complete GIS for business organization. The ArcGIS framework enables deployment of GIS functionality and business logic wherever it is needed—in desktops, servers (including the Web), or mobile devices.

etc. does not also pose a problem, since the MetaStar suite is a platform and database independent, extensible, interoperable and scalable piece of software.

On the negative side of the MetaStar suite, the software does not appear to have any reliable and efficient quality assurance technique for metadata. However, it supports a list of controlled terms, which presumably may aid in ensuring adequacy of metadata during its creation. Moreover, the lack of security or access control for its associated database, where the metadata is said to be stored as well as metadata versioning facility, may be regarded as an assertion of the MetaStar suite's incapability of handling complex issues associated with long-term metadata management. It should also be noted that absence of security or access control facility might enable the associated database to be vulnerable and subject the stored metadata to potential quality threats. In addition, the MetaStar suite does not facilitate preservation of metadata for long-periods of time, which only adds on to its incompetence for long-term metadata management.

F3: Eco Companion Document Management Service

The role of the Eco Companion document management service, developed by IndexGeo Pty Ltd, is to manage the collection of metadata documents, validate them, produce presentation versions, and automate their publishing, management, and indexing [ECO04]. In general, it is not intended to address long-term metadata management. However, it does offer some features, which may be used as guidelines for developing a customized metadata management system to serve such purposes. Highlights of these features are:

- Ease of metadata creation and updates using the familiar interface of a WWW browser as well as online editorial assistance to ensure granularity and adequacy of the metadata. However, lack of controlled keywords or vocabularies may be seen as shortcoming in ensuring metadata quality during its creation or updates.

- As far as the quality assurance of metadata is concerned, functionalities of the system are limited to automatic validation of XML metadata, which only ensures that the document structure is correct, that any additional HTML elements are valid, and the content of certain metadata fields is legitimate. However, this validation technique, although enables increased search efficiency, is not sufficient for ensuring semantic accuracy of the metadata, hence may not be fully effective in ensuring metadata quality.

Appendix G: Contact Information of Project Collaborators

G1: Contacts of the CEDARS Working Group

Maggie Jones,
Cedars Project Manager
Edward Boyle Library
University of Leeds
Leeds LS2 9JT
England, UK
Telephone: +44 (0) 113 343 6386
Fax: +44 (0) 113 343 5539
Email: libmjj@leeds.ac.uk

Derek Sergeant,
Cedars Project Officer
Information Systems Services
The University of Leeds
Woodhouse Lane
Leeds LS2 9JT
Telephone: +44 (0) 113 343 5698
Fax: +44 (0) 0113 343 5411
Email: d.m.sergeant@leeds.ac.uk

G2: Contacts of the NEDLIB Research Group

Ms Titia van der Werf,
Koninklijke Bibliotheek P.O.Box 904072509 LK The Hague The Netherlands
Telephone: +31 70 3140467
Fax: +31 70 3140501
E-mail: titia.vanderwerf@kb.nl

G3: Contacts of the OCLC/RLG Working Group

Brian Lavoie, Ph.D.
Research Scientist (PREMIS)
(Preservation Metadata)
Tel: +1-614-764-4399
Email: lavoie@oclc.org

RLG membership for Europe
Nancy Elkington
Email: nancy.elkington@notes.rlg.org
Tel: +1-646-495-5331
Jennifer Hartzell
Email: jlh@notes.rlg.org
Tel: +1-650-691-2207
Nita Dean
Email: nita_dean@oclc.org
Tel: +1-614-761-5002

OCLC Online Computer Library Center, Inc.
6565 Frantz Road
Dublin, OH 43017-3395
USA
Tel: +1-614-764-6000, 1-800-848-5878
Fax: +1-614-764-6096
Email: oclc@oclc.org

Hilde van Wijngaarden
Digital Preservation Officer
PO Box 90407
2509 LK The Hague
The Netherlands
Tel: + 31-70-3140467
e-Mail: hilde.vanwijngaarden_kb.nl

G4: Contacts of the NLA Preservation Working Group

Tony Boston,
Metadata Repository and Search System RFQ & Digital Object Management System RFQ
Tel: +612 6262 1518
E-mail: tboston@nla.gov.au.

Digital Services Project
E-mail: dsp@nla.gov.au.

G5: Contacts of the NERC Data Grid Project

Dr Bryan Lawrence

NERC DataGrid Principal Investigator
Head NCAS/British Atmospheric Data Centre
Space Science and Technology Department
CCLRC - Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, OX11 0QX, England, UK.
Email: b.n.lawrence@rl.ac.uk

G6: Contacts of the UKDA

Ken Miller
Metadata Management and Production System for Surveys in Empirical Socio-economic Research
Email: millk@essex.ac.uk

Mike King
Systems and Preservation Manager
Tel: +44 (0)1206 873560
Email:mking@essex.ac.uk

Hilary Beedham
Cluster Of Systems of Metadata for Official Statistics (COSMOS)
Email:beedh@essex.ac.uk
Phone: +44 1206 872570
Fax: +44 1206 872003

David Hugh-Jones
Data Services Officer
Tel: +44 (0)1206 872250/2251
Email:djhugh@essex.ac.uk

Pam Miller
Metadata Standards and Resource Discovery Manager.
Tel: +44 (0)1206 873395
Email:millp@essex.ac.uk

G7: Contacts of the NISO (USA) Working Group

Meg Bellinger
Preservation Resources, a division of OCLC
9 South Commerce Way
Bethlehem , PA 18017
Phone: 610-758-8700
Fax: 610-758-9700
Email: bellingm@oclc.org

Robin L. Dale
Chairperson
RLG Member Programs & Initiatives
1200 Villa Street
Mountain View , CA 94041
Phone: 650-691-2238
Fax: 650.964.0943
Email: Robin.Dale@notes.rlg.org

Oya Y. Rieger
Chairperson
Cornell University Library
215 Olin Library
Ithaca , NY 14853
Phone: (607) 254-5160
Fax: (607) 254-7493
Email: oyr1@cornell.edu

G8: Contacts of Other Collaborators

The Digital Archiving Consultancy (DAC)

Philip Lord
2 Wayside Court, Arlington Road,
Twickenham, TW1 2BQ, United
Kingdom
Tel: +44 (0)20 8607 9102
Fax: +44 (0)70 5067 5010
E-mail: support@d-archiving.com

NEESgrid Project

Joe Futrelle
Team Leader
Data & Metadata Management Team
Email: futrelle@ncsa.uiuc.edu
Carl Kesselman,
System Configuration and Design
Email: carl@isi.edu
Ian Foster
Email: itf@mcs.anl.gov

DCMI Preservation Working Group

Dr Andrew Wilson, BA (UNE), Dip Archiv Admin (UNSW), MA (ANU), PhD (Syd),
Email: andreww@naa.gov.au

Dr. Heike Neuroth
Research & Development
Göttingen State and University Library (SUB)
Email: mailto:neuroth@ mail.sub.uni-goettingen.de

The European Bioinformatics Institute (EBI)

Array Express Data Curation Team
Email: arraysubs@ebi.ac.uk,
miamexpress@ebi.ac.uk

Database Group, the computer science department of University of Leipzig, Germany

Phil Bernstein
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
Email:
philbe@microsoft.com
Phone**:** (425) 706-2838
Fax: (425) 936-7329

Paolo Atzeni
Database Group
Dipartimento di
Informatica e Automazione
Università Roma Tre
Via della Vasca Navale
7900146 Roma, Italy
Tel: 39-06-55173213
Fax: 39-06-557.30.30
email: atzeni@dia.uniroma3.it

Rachel Pottinger
Dept. of Computer Science & Engineering
University of Washington
Allen Center, CSE 101/ Box
352350, Seattle, WA 98195-2350
Phone: (206) 616-8067
Fax: (206) 543-2969
Email: rap@cs.washington.edu

G9: Other Useful Contacts

Lola Olsen
Project Manager
(Metadata Management)
GCMD, NASA
Email:Lola.M.Olsen@nasa.gov

Marieke Guy
(Metadata Quality assurance)
Subject Portals Project Manager, ePrints UK Project Manager and QA Focus
UKOLN
Email: m.guy@ukoln.ac.uk
Web site:
http://www.ukoln.ac.uk/

Michael Day
(Metadata Quality Assurance)
Research Officer,
Research and Development
UKOLN
Email:
m.day@ukoln.ac.uk
http://www.ukoln.ac.uk/

Appendix H: Project Milestones

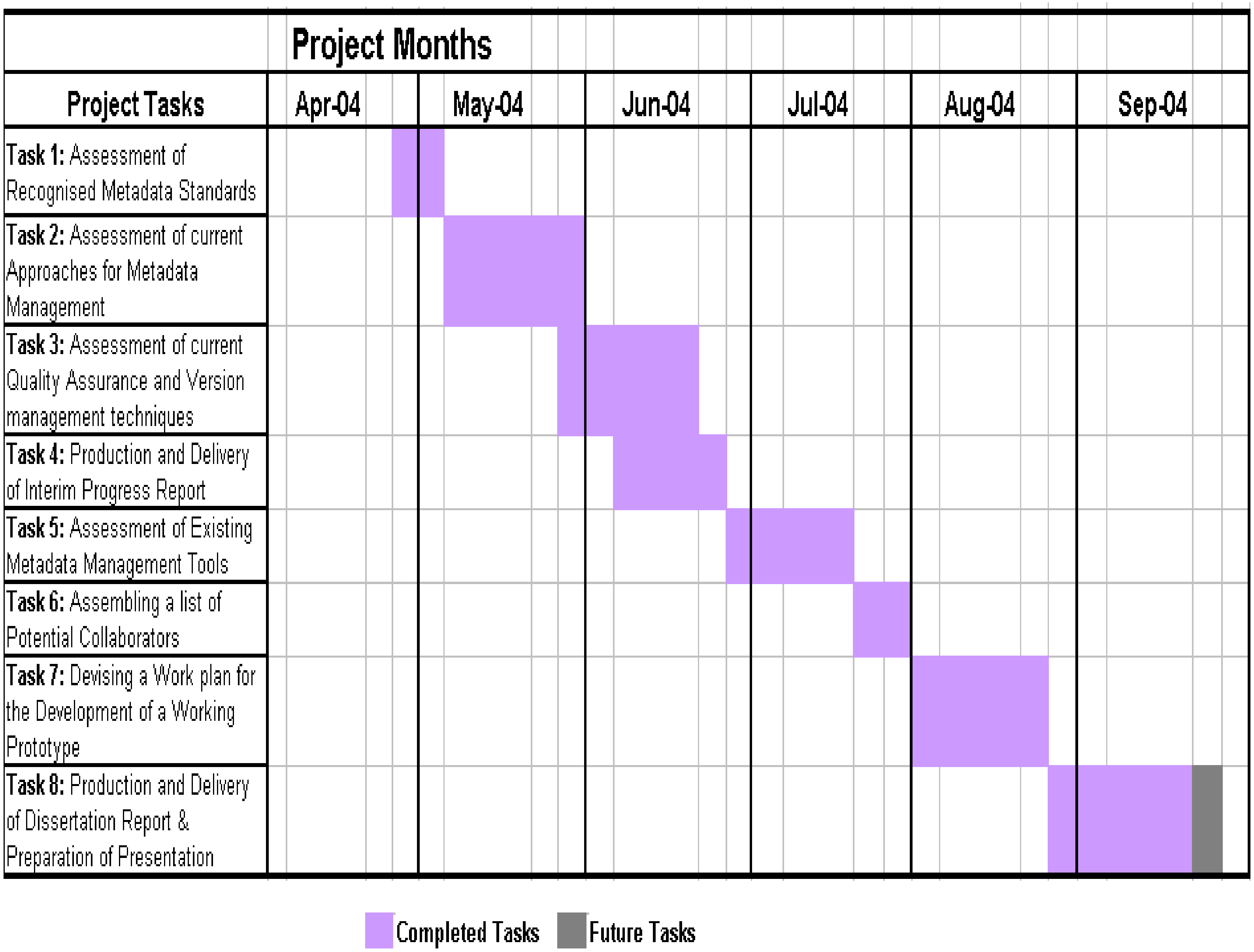


Figure H.1: Project Milestones

Appendix I: Estimated Time Scales for Future Work

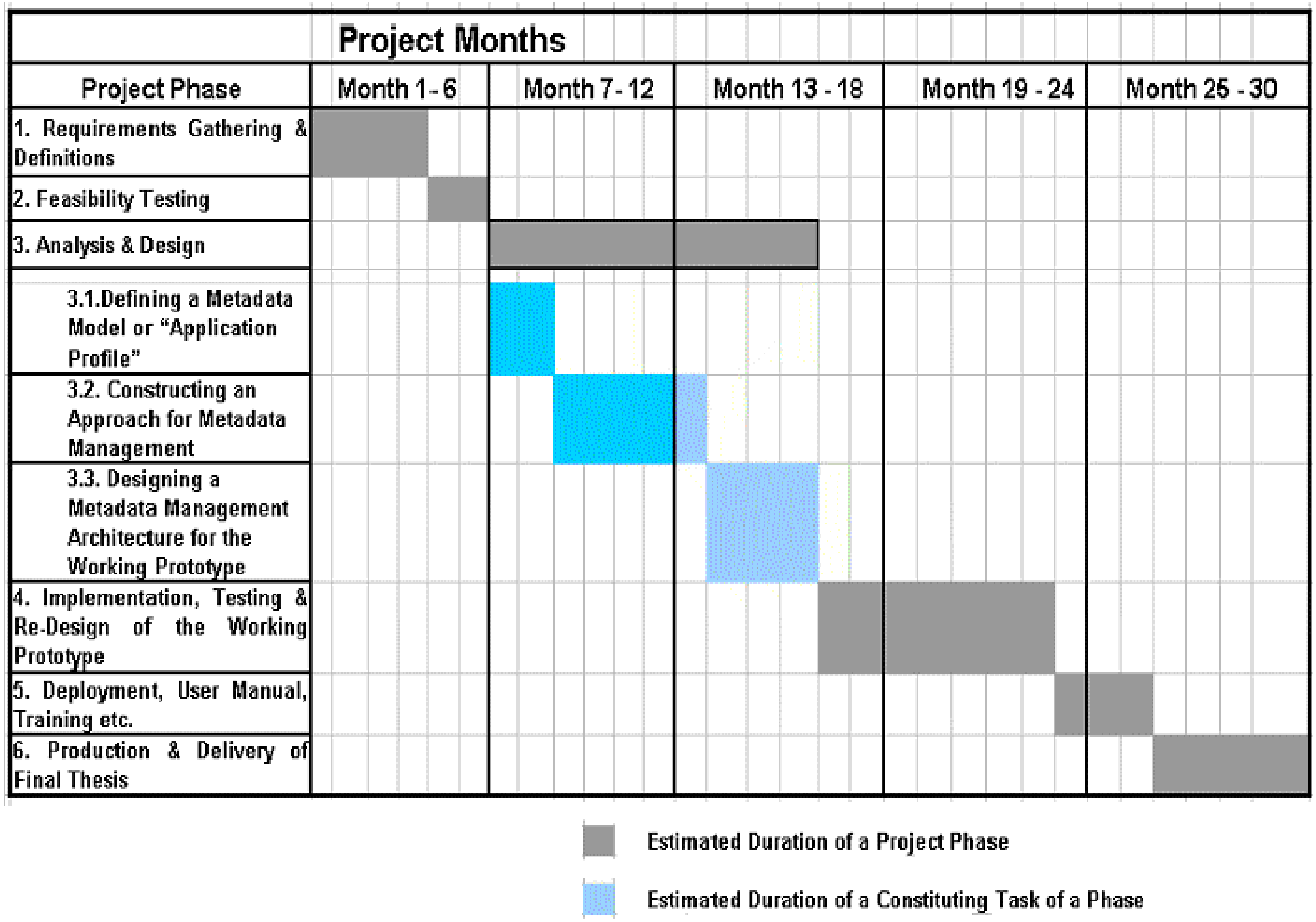


Figure I.1: Estimated Time Scales for the Future Project