iabin
Inter-American Biodiversity
Information Network

Support to Building the Inter-American Biodiversity Information Network

Trust Fund #TF-030388

Review of International Initiatives in Metadata Management

(Document 10(a))

July 2004

NIPPON KOEI

UNEP WCMC

Support to Building IABIN (Inter-American Biodiversity Information Network) Project

Review of International Initiatives in Metadata Management

Project Background

The World Bank has financed this work under a trust fund from the Government of Japan. The objective is to assist the World Bank in the completion of project preparation for the proposed project Building IABIN (Inter-American Biodiversity Information Network) and for assistance in supervision of the project, once and if it is approved. The work undertaken covers three areas: background studies on key aspects of biodiversity informatics; direct assistance to the World Bank in project preparation; and assistance to the World Bank in project supervision. The current document is one of the background studies.

The work has been carried out by Nippon Koei UK, in association with the UNEP World Conservation Monitoring Centre.

Table of Contents

Annexes

Table of Tables

Report Summary

This report reviews existing international biodiversity metadata management initiatives, and makes recommendations as to how this experience and best practice can be utilised by the World Bank in the context of the Inter-American Biodiversity Information Network.

CHAPTER 1 INTRODUCTION

1.1 Introduction

This report reviews existing international biodiversity metadata management initiatives and makes recommendations as to how this experience and best practice can be utilised by the World Bank in the context of the Inter-American Biodiversity Information Network (IABIN).

The IABIN report *IABIN Portal Architecture* (McClarty, 2003) makes recommendations on the system architecture of the IABIN Portal, including a proposal on metadata standards and tools and mechanisms for database interoperability. The current report does not significantly diverge from the recommendations of McClarty, but rather seeks to provide further depth and a broader context for the IABIN metadata activities with reference to other on-going biodiversity information management projects.

Considerable effort and advances in the definition of metadata standards and systems, including the biodiversity subject domain, have been achieved since the mid 1990s. These activities have proceeded in parallel with the more generic developments of the WWW and the emergence of Web Services and their many related technologies. During this period, numerous information management projects, justified by the information sharing aspirations of environmental conventions, have recognised the potential for the electronic medium in providing increased access to biodiversity information and the potential benefits that flow there from. Because generic tools and protocols for information sharing have taken time to develop, there has been an unavoidable period of duplicated effort as consensus emerges. However, sufficient time has now elapsed for the consensus to be reasonably clear, but there is still a process of convergence necessary as legacy systems become fully interoperable.

CHAPTER 2 KEY PRINCIPLES OF METADATA MANAGEMENT AND USE

2.1 A Definition of Metadata

Metadata are frequently defined as being "data about data"[1]. A typical example of metadata is a card catalogue in a library. To efficiently find data of interest, say the contents of a book, the library user first refers to the card catalogue to locate suitable books by subject, title or author. In a large library this is a much more efficient approach to finding a book than randomly selecting books from shelves. The card catalogue therefore provides a means of browsing summary information, in a structured way, so as to discover where to find a particular book on the shelves of the library. The library card catalogue therefore provides a way of discovering information resources - books. Having found a suitable book, the user may then refer to the table of contents or index to rapidly move to a specific section of the book. Both the table of contents and the index contain further ancillary information that can be considered metadata.

This definition - "data about data", is useful but too simplistic to fully define the role which metadata is to provide within IABIN. In defining metadata, we also need to consider the purpose and use of “data about data”, as well as their content and structure (Ahmed *et al* 2001). The context of metadata is important, as one person's metadata may be another person's (or application's) data. There are many borderline examples of where the metadata/data boundary lies. For example, is the content of a telephone directory metadata or data? There are those who will argue the case for each of these positions and both are right, depending on their viewpoint. Furthermore, to a database programmer, metadata means the detailed description of the tables and their fields within a database, rather than a high level description of the database itself.

This report deals with biodiversity information, but illustrates that there are a number of metadata domains within "biological informatics" which may appear as metadata or data depending on the way in which the information is to be used. This is explained in more detail in Deliverable 7.3.

As well as distinguishing between *metadata* and *data*, the distinction between *data* and *information*, and indeed *information* and *knowledge* should also be made. A predictive weather report may be compiled from vast amounts of data, e.g. satellite images and 30-year mean climate statistics, and then processed using advanced climate models on super computers. In isolation, the elements of the raw data may be useless in ascertaining what the weather is going to be like at a

[1] This definition can be extended to “data about data about data...”.

given location on a given day. However, by processing the raw data using predictive models, reports about the daily weather prospects can be produced. Metadata can be used to assist in locating the data needed to perform the "business logic" (weather prediction) and to catalogue the results (daily forecasts). In relation to biodiversity the business logic will be different and may refer to a task such as reporting species conservation status at a national level, or some other such task.

The library example is instructive because it illustrates the generic and multi-level nature of metadata:

- Library - card catalogue – book
- Book – table of contents – data/information

The elements that comprise the card catalogue in fact constitute yet another level of metadata, meta-meta-data.

2.2 Levels of Metadata and the Purpose of Metadata within IABIN

As explained above, many levels of metadata can be identified. For example, as shown below, the Global Biodiversity Information Facility identifies five levels. The discussion in this report is limited primarily to "Discovery Metadata", i.e. metadata that is used to assist finding or discovering information resources. Report 7.3 considers in more detail other levels of metadata that are more concerned with data interoperability.

2.3 Metadata Policy

IABIN should have developed and published a metadata policy that clearly states the purpose of IABIN metadata, what standards it adheres to, how it relates to other biodiversity information sharing and metadata initiatives, and what its metadata contain. This information can be formally set out in an IABIN Collections Policy, i.e. what metadata are collected by IABIN and for what purpose. This is analogous to a library collections policy. The policy need not be lengthy, but should be readily accessible probably as a publication on the IABIN web site.

2.4 Metadata Location

As stated above, metadata can be used at many different levels and can reside in a number of different places. For the purposes of IABIN, we should identify:

1. Metadata that resides within a structured catalogue – analogous to the card catalogue, which refers to information resources that are held elsewhere, perhaps at a given URL or on a bookshelf;

2. Metadata that are carried within an information resource, such as in a <META> tag in an HTML www page, or the Properties dialogue of a MS-Word document;
3. Metadata that are used in a database system and which describe the contents and structure of the database, its fieldnames, data types, etc.

Metadata of the first category will probably be compiled by a person who has the task of indexing information resources in a catalogue.

Metadata of the second category will be compiled and entered into a www document by a person during the course of compiling the page. Some elements may be populated by a www authoring tool. The content of the <META> tags, however, will be harvested by a web robot, or Webbot, and made available through a web search engine or indexing system. An example of this type of mechanism is the normal web robot that crawls the web looking for content. The UK National Biodiversity Network (NBN) uses metadata embedded in the meta tags of HTML, pages of which can in turn be harvested by a Webbot. Where web pages are dynamically generated from a database the metadata tags should be appropriately completed at the same time as the content is assembled.

Metadata of this last category are discussed in Report 7.3 on database interoperability.

CHAPTER 3 METADATA STANDARDS

There are three types of metadata standards:

- Discovery standards – for example, what metadata elements do we need to use to describe a specific information resource?
- Semantic standards – for example, what is the meaning of a keyword, or what is its equivalent in another language?
- Syntactic standards – for example, how do we encode our metadata?

The emergent syntactic standard is the eXtensible Markup Language (XML), whilst numerous standards exist for discovery and semantics. Both semantic and discovery standards can be subject domain specific. Because information management requirements differ, both between subject domains and even between organisations operating in the same domain, it is unreasonable to expect that a single, all encompassing, metadata standard will emerge to cater for all applications. However, there are only so many ways of describing information resources, and many standards include a core set of identical or similar descriptive elements.

A number of metadata standards are of relevance to IABIN, and some of these have been identified in the IABIN Portal Architecture Report in relation to the IABIN Catalogue Services. This section is restricted to what are clearly discovery metadata, rather than data or data transfer standards. Standards that are more closely associated with data transfer and exchange (such as Darwin Core and ABCD) are discussed in more detail in report 7.3.

In relation to the IABIN Catalogue Services, McClarty (2003) identifies the requirement for IABIN Catalogue Services to catalogue the following types of information resources:

- Bibliographic;
- Datasets;
- Websites.

McClarty (2003) recommends the utilisation of a specific metadata system. Here, however, we prefer to discuss agnostic metadata standards rather than specific proprietary implementations.

3.1 The Dublin Core Metadata Initiative

The Dublin Core Metadata Initiative (DCMI or DC) is a set of 15 metadata elements that are designed to describe generic electronic information resources - typically electronic documents. Having only 15 elements, the DC is both simple to implement and to understand. Furthermore, its 15 generic elements are more often than not found in other more complex metadata standards, and it can therefore be used as a common core that can be used across diverse metadata domains, thereby enabling cross-domain searching.

The DC element set is not intended to replace other more comprehensive systems, should they be required. Furthermore, the DC element set is not fixed, and elements can be qualified and new elements defined to extend the set to suit a particular application. It should be emphasised that the DCMI element set is intended to describe information resources as metadata, rather than, necessarily, to contain data. The DCMI element set may therefore be used to describe an electronic document, such as a WWW page, indeed the DCMI element may be embedded into the page's HTML. It may also be used to describe a database, its content description and location, but not its actual contents.

The 15 basic DCMI elements are as follows:

- **Title** - A name given to the resource;
- **Creator –** An entity primarily responsible for making the content of the resource;
- **Subject** - A topic of the content of the resource;
- **Description –** An account of the content of the resource;
- **Publisher** - An entity responsible for making the resource available;
- **Contributor** - An entity responsible for making contributions to the content of the resource;
- **Date -** A date of an event in the lifecycle of the resource;
- **Type** - The nature or genre of the content of the resource;
- **Format** - The physical or digital manifestation of the resource;
- **Identifier** - An unambiguous reference to the resource within a given context;
- **Source** – A Reference to a resource from which the present resource is derived;
- **Language** - A language of the intellectual content of the resource;
- **Relation –** A reference to a related resource;

- **Coverage –** The extent or scope of the content of the resource;
- **Rights** - Information about rights held in and over the resource.

Clearly there is nothing about this element set that is specific to biological information resources. However, there is no reason why this element set cannot refer to biological information resources, or why it cannot be made specific to the description of biological resources through the qualification of elements or the addition of new elements.

Because the DCMI presents such a generic element set, it can be mapped directly to equivalent elements in other metadata standards, in a process known as cross-walking.

Full details of the DCMI can be found at http://www.dublincore.org.

3.2 ISO 19139 and ISO 19115 Standards / Federal Geographic Data Committee

The US Federal Geographic Data Committee (FGDC) is responsible for developing the Content Standard for Digital Geospatial Metadata (CSDGM). Federal agencies concerned with compiling and managing geospatial information are required to compile metadata records, according to the standard, as a contribution towards the development of the National Geospatial Data Clearinghouse (FGDC, 1999). Furthermore, the FGDC encourages other public and private organisations to adopt the same standard.

The role of the standard is to enable the documentation of metadata of digital geospatial datasets for the stated purposes of:

- Preserving the meaning and value of datasets;
- Contribute to a catalogue or clearinghouse;
- Aid data transfer.

(Source FGDC, 1999)

Clearly, obliging agencies at the Federal level to adopt and implement a geospatial metadata standard is a powerful way of empowering that community and those that have access to the clearing-house services.

A related Spatial Data Transfer Standard (SDTS) was also developed to enable the transfer of digital geospatial data between systems.

The FGDC metadata standard has a somewhat confusing relationship with the emerging International Organization for Standards (ISO) geospatial metadata standard. The FGDC standard, together with others, was used as the basis of the

ISO geospatial metadata standards - ISO 19139 and ISO 19115. ISO 19139 provides an implementation specification, based on a Unified Modelling Language (UML) abstract model defined by ISO 19115.

The standard is extensive and includes structured information in the following sections:

- Identification;
- Data Quality;
- Spatial Data Organisation;
- Spatial Reference;
- Entity and Attribute;
- Distribution;
- Metadata Reference.

Space does not permit a complete discussion of the CSDGMD, but suffice to say it is comprehensive and has more than 200 elements and sub-elements. Furthermore, it has a number of associated tools for compiling, storing and searching for metadata.

Full details of the FGDC standard can be found at http://www.fgdc.gov

An easy to understand image map of the FGDC standard can be found at http://biology.usgs.gov/fgdc.metadata/version2/ in both English and Spanish.

Full details of the ISO geographic information standards can be found at http://www.isotc211.org

3.3 NBII Biologic Data Profile

FGDC provides a metadata standard for digital geospatial data, whilst this report is concerned with metadata in relation to biodiversity. Clearly, biodiversity information may be either aspatial or spatial and, as such, the FGDC CSDGM may or may not be suitable for cataloguing biodiversity data, both in terms of the elements required to describe biodiversity datasets and whether they contain spatial entities. To address both of these issues, the FGDC Biological Data Working Group has compiled a Biological Data Profile (BDP) which amends the CSDGM to cater for both of these issues, by providing additional elements to fully describe biological data, modifying the conditionality of other elements to suit biological data and, lastly, by making the geospatial elements only "mandatory-if-applicable."

This issue raises a number of questions regarding the nesting of information attributes, and whether spatial location is an attribute of a biological entity, and if so, whether it is necessarily important in all situations, or whether a biological entity is an attribute of a spatial location or feature. The emergence of a geospatial metadata standard before a biological metadata standard may have more to do with the profile of geospatial systems, in both the private and public sectors, than with the wisdom of adding a biological profile to a geospatial standard, rather than the other way around. In the Knowledge Network for Biocomplexity below, this notion is turned on its head.

The BDP contains all of the CSDGM metadata elements unchanged, although the conditionality of some of them differs. Additional sections to those listed above, for the standard FGDC, are provided by the BDP and include:

- Identification;
- Entity;
- Attribute.

Full details of the NBII Biological Data Profile can be found at http://www.nbii.gov

3.4 Knowledge Network for Biocomplexity

The Knowledge Network for Biocomplexity (KNB) is a national network in the United States, which is intended to help facilitate research on “biocomplexity” through information sharing. The network is sponsored by the National Science Foundation, the National Center for Ecological Analysis and Synthesis, Texas Tech University, the Long Term Ecological Research Network and the San Diego Supercomputer Center. In the context of the KNB, biocomplexity refers to the relationships between attributes of biodiversity and ecosystem function. In relation to considering KNB in the context of metadata use within IABIN, we can substitute biocomplexity with biodiversity, as the issues are effectively the same.

These goals are approached through information sharing, by providing open tools to the community to – “discover, access, interpret, integrate and analyse complex ecological data from a distributed set of field stations, laboratories, research sites and individual researchers”.

Whilst the KNB research goals focus specifically on the relationship between biodiversity and ecosystem function, the tools developed for information access and management are of direct relevance to IABIN. Indeed, the combination of advanced computing and ecological expertise has produced a formidable set of tools unencumbered by legacy informatics.

Specifically the KNB provides the informatics products shown in Table 1.

Table 1 KNB Informatics Products

Ecological Metadata Language (EML)	A metadata standard implemented as a set of XML Schemas, which can be utilised in a modular and extensible manner.
Morpho	A data management tool including functions for metadata creation, data query and data management.
Metacat	A metadata database, utilising XML as a representational syntax and suited for storage of EML
Monarch	A data exploration and analysis tool, utilising structured metadata and supporting linkages with analytical packages such as SAS, R and MatLab.
Itislib	A Java library providing an API to the ITIS*ca database of taxonomic nomenclature.
Storage Resource Broker (SRB)	A request broker providing links to distributed networks of heterogeneous storage resources. Used in conjunction with Metacat for the management of large data objects.

In addition, Arizona State University is developing the KNB-related informatics products shown in Table 2.

Table 2. KNB Related Informatics Products from University of Arizona

Xylographa	An XML input wizard, not available at the time of writing, although Morpho provides EML data input – see above.
Xanthoria	A SOAP based query tool. Operational at the time of writing, although not available as a download.

These tools are located within an architecture that is designed to provide:

- Data access;
- Information management;
- Knowledge management;

Of particular concern to this report and IABIN, is the metadata component provided by the Ecological Metadata Language (EML). The concept of EML within the KNB is of a single component in a comprehensive information sharing and analysis system. This report therefore focuses on EML; a number of the other components are considered in report 7.3.

3.5 The Ecological Metadata Language (EML)

3.5.1 Purpose of EML

The stated purpose of EML is:

"to provide the ecological community with an extensible, flexible, metadata standard for use in data analysis and archiving that will allow automated machine processing, searching and retrieval"

Key points of note in this definition are: 1) that EML metadata is concerned not only with archiving and retrieval, but also with data analysis and automated machine processing, and 2) that the starting point is the ecological community, not the geospatial community.

3.5.2 Features of EML

The key features of EML, as identified in the EML specification, are:

- Modularity – designed as a set of related modules, implemented in an extensible architecture, EML can be used either selectively or can be extended;
- Detailed structure – in levels of detail caters for both human and machine-readable applications;
- Compatibility – borrows syntax from other existing metadata standards, including DCMI and FGDC, NBII, ISO 19115, ISO 8501, OGC, STMML and XSIL;
- Strong Typing – implemented in XML Schema;
- Differentiation between content model and syntactic implementation – the normative specification defines the content model, whilst XML Schema defines the syntactic implementation.

The fundamental EML unit is the *Data Package*. Unlike other metadata systems, an EML data package may contain not only metadata, but also references to, or the actual data, to which the metadata refers. EML may be extended using the <additionalMetadata> sub-field.

3.5.3 EML Modules

EML uses the DCMI definition of a resource as referring to a "networked digital resource", although non-networked resources can also be catalogued. Some EML modules can stand alone, whilst others have dependencies on other modules. Also, module content can be referenced from multiple locations, meaning that the same content, for example an address, can be repeated without having to repopulate it each time.

EML is comprised of a set of metadata modules, as shown in Table 3.

Table 3. EML Metadata Modules

Root-level	
eml	The metadata container
eml-resource	Base information for all resources
Top-level	**Used to describe separate resources**
eml-dataset	Describes dataset resources
eml-literature	Describes citation (bibliographic) resources
eml-software	Describes software resources
eml-protocol	Describes "abstract, prescriptive procedures for generating or processing data"
Supporting Modules	**Add detail to top level resources**
eml-access	Describes access control
eml-physical	Describes physical characteristics, e.g. filename, size, encoding, etc.
eml-party	Describes responsible party
eml-coverage	Describes geographic, temporal and taxonomic coverage.
eml-project	Describes research context
eml-methods	Describes methods used in creation
Data organisation	**Used to describe logical layout of dataset.**
eml-entity	Describes logical characteristics of each entity in dataset, i.e. tables of data.

eml-attribute	Describes all attributes in a data entity.
eml-constraint	Describes integrity constraints between entities.
Entity Types	**Extends entity module elements with entity specific elements.**
eml-dataTable	Describes logical characteristics of each tabular set.
eml-spatialRaster	Describes rectangular grids of georeferenced data
eml-spatialVector	Describes vector spatial data including points, lines and polygon geometries and topology level.
eml-storedProcedure	Describes procedures that produce data output in the form of a data table.
eml-view	Describes a "view" from a DBMS.
Utility modules	
eml-text	Wrapper container to allow general text descriptions in structured or unstructured text blocks.

3.5.4 EML Summary

EML provides a comprehensive open and extensible metadata standard, which has been developed from the point of view of the ecologist and which encompasses the best practice and elements of other standards such as DCMI and FGDC. Furthermore, EML and its supporting tools have been developed and implemented using the latest W3C standards such as XML Schema. The EML Data Package considers, not only discovery metadata, but also metadata describing data and how to extract it, and can indeed even contain the data themselves. Coupled with the Morpho, Metacat, Monarch and related tools, the entire package offers a data sharing and processing system of great breadth and depth, which, coupled with Web Services technologies (such as Xanthoria), will provide a powerful set of information sharing and processing tools.

Full details of the KNB, EML and its related tools can be found at: http://knb.ecoinformatics.org/software/eml/eml-2.0.0/index.html

CHAPTER 4 CONTROLLED VOCABULARIES

Reports 6.1 and 6.2 are concerned specifically with biodiversity related vocabularies and thesauri. However, it is important to note that controlled vocabularies are an important component of metadata systems and can be used effectively to enhance categorising information resources and in searching. Multi-lingual vocabularies are particularly powerful, as they provide a simple means with which to cross language barriers. Multiple controlled vocabularies can be used in metadata systems, but it is important to record, not only the word selected from the vocabulary, but also the vocabulary from which it came. Vocabularies for consideration for use within IABIN metadata systems include:

AGRO-VOC Multilingual Food and Agriculture Organization's agricultural vocabulary

GEMET Multilingual environmental vocabulary of the European Environment Agency

CBD-VOC Multilingual biodiversity vocabulary used by the Convention on Biological Diversity

ENVOC Environmental vocabulary use by the United Nations Environment Programme

AOS Agricultural Ontology Service is an initiative of the Food and Agricultural Organization to produce a service that brings together various thesauri in an ontology, to provide improved semantic searching.

CHAPTER 5 METADATA CROSS-WALKING

Clearly, a number of metadata systems are in use by biodiversity organisations, networks and initiatives. Each initiative has its own particular requirements and capacity to construct metadata, and whilst standardisation is desirable, it is unlikely that it will be achieved by using the same systems or even the same standards. For example, the members of the CBD CHM, the US NBII and members of the Knowledge Network on Biocomplexity may all have metadata catalogues of their information resources, and it would be most useful to be able to search across all of their metadata. However, each of these initiatives uses a different metadata standard, namely the Dublin Core, the FGDC Biological Data Profile and the Ecological Metadata Language.

How can searches be made across each of these diverse systems? A pragmatic and relatively simple solution is to use metadata cross-walking, or what is known as the “dumming down principle” in Dublin Core parlance. This principle takes the common terms that are found in each of the three systems, and uses their equivalence to provide a level of interoperability. Clearly, if the FGDC BDP has potentially hundreds of metadata elements and the Dublin Core has only 15 unqualified elements, there will be a partial overlap. However, as the Dublin Core is designed specifically as a core of essential terms, it is likely that all 15 Dublin Core elements have their equivalents in both FGDC and EML. Indeed, EML was designed with this in mind and with reference to FGDC, and the cross-walk equivalence between FGDC and Dublin Core elements is published.

A comprehensive list of metadata cross-walks, although not including EML, can be found at http://www.ukoln.ac.uk/metadata/interoperability/

CHAPTER 6 BIODIVERSITY METADATA INITIATIVES

6.1 CBD Clearing-House Mechanism

The Clearing-House Mechanism (CHM) of the Convention on Biological Diversity (CBD) is a response to the call for biodiversity information sharing, and the benefits thereof, as stated in Article 17 of the Convention. The mission of the Clearing-House is to:

- Promote and facilitate technical and scientific co-operation, within and between countries;
- Develop a global mechanism for exchanging and integrating information on biodiversity;
- Develop the necessary human and technological network.

(Source www.biodiv.org)

To assist in achieving these aims, the CHM has developed a CHM Toolkit, which includes information on developing National CHM websites and information on common standards and metadata. With respect to metadata, the CHM has adopted the Dublin Core as its description standard, RDF as its semantic schema and XML as its syntactic standard. Currently, the Dublin Core is only utilised by the CHM for description of the content of web pages.

In addition, the CBD has developed the CBD Controlled Vocabulary as a multi-lingual thesaurus for use in metadata descriptions.

The CBD CHM has a stated policy of use of Dublin Core for metadata compilation of its web site, and is currently investigating using RDF, XML and other technologies, such as SOAP, for interoperable data transfer.

It should be noted that the development of the CHM and the establishment of National CHM Focal Points did not necessarily develop using these objectives, and that different National Focal Points currently use different metadata systems. For example, the US NBII uses the BDP of FGDC, whilst the UK National Biodiversity Network uses its own metadata standard based on the UK National Geospatial Data Framework standard for geospatial data (NGDF). Again, the Dublin Core represents a lowest common denominator between these systems.

6.2 Global Biodiversity Information Facility

The goal of the Global Biodiversity Information Facility (GBIF) is to “make the world's **primary data on biodiversity** freely and universally available via the Internet” through an “interoperable network of biodiversity databases and information technology tools using web services and Grid technologies" (www.gbif.net).

GBIF is a relatively new initiative and has the advantage that its vision is based on the new standard technologies available through Web Services, without the need to carry the baggage of legacy systems. In this respect, it is starting with a clean technology slate.

Central to GBIF are Web Services. Web Services, as explained in Report 7.3, are biodiversity "business logic" that is available through standard-based Internet protocols such as HTTP and SMTP" (Chappell and Jewell, 2002). The important thing about Web Services is that they are based on standards and are platform independent. Their technologies are based on XML interfaces between the Simple Object Access Protocol (SOAP), the Web Services Description Language (WSDL) and Universal Description, Discovery, and Integration (UDDI).

Indeed, GBIF identifies five levels at which metadata are utilised, namely:

- Interface;
- Provider;
- Service;
- Reply;
- Record.

(Source Hobern, 2003)

Metadata are a key component of Web Services, which use metadata for a number of purposes. The technologies of Web Services are explored further in Report 7.3.

6.3 Harmonizing Metadata Initiatives Throughout IABIN

In 1999 IABIN under the leadership of Vincent Abreu undertook a review of existing metadata initiatives in the IABIN region. This review had four key deliverables:

1. A web page providing metadata guidelines
2. Development of a prototype biodiversity catalogue for Central America
3. Survey of existing metadata activities in the region
4. Final report providing guidelines and recommendations.

The final report can be accessed at the following URL (http://www.iabin-us.org/documents/proj_reports/metadata_fnl.pdf) and provides a comprehensive snapshot of the metadata initiatives under way in the region at the end of the 20th Century. It should be noted that at present this is an extremely active area with new initiatives frequently being undertaken. With this warning in place it is interesting to note the extensive range of activities already underway in the region. It would be of particular interest to see how many of the recommendations from the report have been followed. In particular, the recommendation of the establishment of a "clearinghouse" and the associated quality control on the content is a very strong recommendation. The list of metadata elements to be used by IABIN on pages 10 and 11 of the report are very sensible and provide an appropriate compromise between usability and excessive levels of interoperability.

It is not clear at the time of writing which of these recommendations have been adopted throughout IABIN and which are still seen as future goals. A recommendation from this review is that the survey is repeated to ascertain what progress has been made within the individual organisations to adopt the recommendations already made. In particular this second review should focus on reasons for non-adoption. It would be anticipated that other initiatives would have taken precedence over the recommendations.

The recommendation of the co-operation with other initiatives on the development of a multi-lingual biodiversity focused thesaurus and its implementation in systems throughout IABIN is one which appears to have had some success with the development of BIOBot and its ability to tap into many disparate and distributed datasources.

6.4 Interoperability between Initiatives

In that IABIN needs to operate among a number of existing metadata initiatives and domains, it is important that it can interoperate between different systems. We have seen that a common theme among the different discovery metadata standards is the Dublin Core, as a lowest common denominator. Clearly, IABIN should strive to meet the requirements of this element set as a prerequisite. This requirement is not unique to IABIN, and clearly other initiatives have similar goals.

CHAPTER 7 THE IABIN METADATA PROPOSAL IN THE CONTEXT OF THE IABIN PORTAL ARCHITECTURE

At the time of writing, Version 0.5 of the IABIN Portal Architecture report by Darrell McClarty was available, which describes the IABIN portal, its requirements and the underlying system architecture. This includes in sections 3.3 and 4.2 details of the IABIN metadata Catalogue System. This report proposes that the IABIN catalogue services should include metadata references to the following types of content:

- Bibliographic Data;
- Datasets;
- Websites.

Furthermore, the catalogue system should be able to perform the following functions:

- Management of metadata for bibliographic resources and datasets;
- Online submission of records;
- Search and display of metadata records according to specified search criteria, based on:
  - Free text search;
  - Search on an individual field;
  - Search on a combination of fields;
  - Geographic search;
- Download of query results;
- Multilingual presentation.

With regard to metadata standards the McClarty report proposes that:

- The DCMI element set is used for bibliographic resources;
- The FGDC standard, together with the Biologic Data Profile, is used for “Datasets”; and
- An indexing system is used for cataloguing web resources, i.e. HTML pages of appropriate sites.

Whilst not disagreeing with the broad outline presented by the McClarty report, the above discussion suggests that the following qualifications should be made:

1. The DCMI element set is a suitable metadata framework for the bibliographic information in IABIN. It is likely, however, that qualifications to the basic elements

will be required to tailor the element set to the specific purposes of IABIN. The DCMI is extensible and can cater for qualified elements;

2. Whilst the DCMI has its roots in the bibliographic arena it is actually more generic and is designed to cater for “electronic resources” defined as “anything that has identity”, namely a URI. Furthermore, DCMI state that there “... are no fundamental restrictions to the types of resources to which Dublin Core metadata can be assigned.” The DCMI can, with qualification, be used to describe non-bibliographic resources, such as datasets, at a resource discovery level.
3. The FGDC is a national standard, whereas IABIN is an international initiative. It may be more politic to refer to the broadly equivalent ISO standard instead of, or in association with, the FGDC.
4. Whilst many biodiversity datasets contain a spatial component, not all do. The NBII Biologic Data Profile makes the spatial component of FGDC non-mandatory if appropriate. This non-mandatory spatial option should be honoured where appropriate.
5. It should not be assumed that all data providers or users are GIS users.
6. It is not clear what a user will do with discovered resources once they have located them. The use to which they are put, whether by human or machine processing, may dictate the granularity of the metadata required, i.e. whether DCMI or a more comprehensive metadata profile such as ISO 19139 / FGDC or EML.
7. The argument that comprehensive metadata should be compiled for all datasets is only sustainable if people have the time (resources) and tools to do so.

CHAPTER 8 RELATED TECHNOLOGIES

A number of discovery metadata standards are discussed above and the distinction has been made between discovery metadata semantics and syntax. Each of these technologies deserve text books to describe them sufficiently but, for completeness, they are considered very briefly below.

8.1 Resource Description Framework (RDF)

The Resource Description Framework is a simple but powerful data model for describing resources, using triplets of name/value pairs and identifiers. RDF can be used as a model for metadata by identifying the location of an information resource in a document, the name of the descriptor, and the value of the descriptor. This simple triplet can be used with great power in defining metadata.

8.2 Topic Maps

Topic Maps enable the navigation of large quantities of information through a navigation layer consisting of topics (representing subjects), associations (representing relationships between subjects) and occurrences (instances of objects that are relevant to a given subject) (Ahmed, 2001). Topic maps are potentially a powerful way of indexing information resources. However, they are also laborious to construct, and probably therefore outside the scope of the IABIN initiative at present.

8.3 eXtensible Markup Language (XML)

The eXtensible Markup Language is a meta-language that can be used to define other markup languages, such as EML. As well as being extensible, XML is also operating system and platform neutral, easy to understand, and well suited to hierarchical information. XML, unlike HTML, is concerned with information structure and content, rather than with presentation. XML can be used to both define a document’s format and to contain a document’s data content. There are numerous XML related tools, such as XML Schema, XPath, XLinks, Xpointers, which make it a powerful tool for storing and specifying metadata containers.

8.4 ANSI Z39.50

ANSI Z39.50 is a network protocol that enables the searching of diverse and heterogeneous metadata catalogues to retrieve information using a single user interface. It has also been adopted by ISO as ISO 23950. The protocol is mainly in use by the library community for searching library catalogues. The protocol is, however, difficult to configure and somewhat esoteric. Whilst it has been adopted by some metadata initiatives, such as FGDC, its prominence appears to be waning somewhat in the face of new developments such as DiGIR and Web Services, which can be configured to achieve the same ends. However, that is not to say that Z39.50 will not remain an important and valuable protocol for metadata searching and information exchange in the future.

CHAPTER 9 METADATA IN THE INSTITUTIONAL INFORMATION CULTURE

Metadata can occupy an uneasy position in an organisation's framework. Who in an organisation is responsible for the maintenance of metadata, and where should it reside? These roles have traditionally been the domain of the librarian, and indeed they still are, but the now ubiquitous WWW is making more of us our own librarians, in the same way that GIS has made us our own cartographers, and word processors our own secretaries. Compiling metadata is not a trivial task, either in terms of understanding it role, its utility, the tools used to manage it, or in its compilation.

It is valuable to take the view that a dataset is not complete without its metadata, and that metadata should be compiled now rather than being left till later. Systems that have a close binding between their data and metadata are those that will probably be the most successful in the long term. It is no coincidence that the KNB Morpho tool's basic entity is the Data Package, which is comprised of both data and its associated metadata.

To make metadata a successful data sharing and information discovery tool, its compilation and management must be part of the institutional data management culture and its costs should be accounted for, rather than being assumed to be an additional task that perhaps an already over worked staff can necessarily absorb. Institutions employ librarians to manage libraries, and the role of the metadata compiler should not be neglected, as metadata, when properly managed, retains and adds value to what is often an institution's most valuable asset – its data repository.

CHAPTER 10 CONCLUSIONS AND RECOMMENDATIONS

10.1 Conclusions

Great efforts have been made over the past ten years to establish and standardise electronic metadata collection, recording and searching as an aid to information sharing. This is true across a broad range of subject domains, including biodiversity and related disciplines. A number of standards are available as a basis for biodiversity metadata compilation in IABIN, but it is unlikely that a single standard and related tools are yet available that can fully satisfy all of the requirements of the IABIN network. They would also have to strike the right balance of complexity whilst, being easy to use, not requiring an onerous level of data input, being available in the language of choice, and providing interoperability across language boundaries.

It should also be remembered that ten years is a short period of time, and that the development of electronic information systems is still evolving rapidly. However, a clear convergence can be observed in terms of consensus on requirements, systems and technologies in the realms of both metadata and data. This rapid evolution of technology and standards means that some systems that were at the cutting edge only a few short years ago, are now legacy systems in need of updating. For example emerging systems that use HTTP as their transport mechanism is now eclipsing the Z39.50 standard for interoperable catalogue searching. In short, things are still in a state of flux.

The IABIN Portal Architecture report identifies a multi-tiered Web Services model as the most logical choice for the IABIN Portal. Metadata, service discovery and data description play an important role in this model at a number of levels. This topic is discussed more fully in Report 7.3.

Most biodiversity practitioners are not interested in how metadata systems function, they are instead interested in how a metadata catalogues can help them to locate relevant information.

10.2 Recommendations

Tanenbaum’s adage that "the nice thing about standards is that there are so many of them to choose from" is starting to have resonance in the realm of biodiversity informatics – FGDC, BDP, ISO 19139, EML, DCMI, NBN Metadata Initiative, etc. Clearly IABIN needs to move forward in providing tools for discovery level metadata compilation and searching. The IABIN Portal Architecture (McClarty, 2003) is explicit in its proposal for the utilisation of the Mercury system, developed by the Oak Ridge National Laboratory, which supports the use of DCMI, FGDC and the Z39.50 protocol, for the IABIN Catalogue Service. This system would be provided as a "turn-key" service to IABIN. Whilst this proposal has merit, not least because Mercury is in use in the LBA Project in Latin America, the adoption of Mercury has not been sufficiently justified, as it is understood that Mercury is based on open standards but is a proprietary system.

1. Open standards and open systems have much to recommend them and IABIN should use them as an aid to interoperability. It is understood that the Mercury system, whilst using open standards, is a proprietary system, so careful consideration and justification should be given to its adoption.

2. Consideration should be given to the use of EML as a metadata framework and its associated tools. EML offers a rich and extensible metadata paradigm unencumbered by legacy technologies. Furthermore, EML offers a tight linkage between data and metadata and offers a range of support tools.

3. IABIN should more clearly state what the role of its catalogue services are. The catalogue services will assist in locating information resources, but what happens next and how does this metadata layer interface with the Web Services architecture? In particular IABIN should look at the on-line support that can be provided to the partner organisations in terms of how to best deploy many of these tools and methods. The establishment of many interoperability standards allows each site to implement in their own preferred way but to still contribute to the system as a whole and it is this concentrator activity which is crucial to the success/failure of a distributed metadata tool.

4. The IABIN catalogue service should be based on a discovery metadata profile that is appropriate, and has good supporting tools and training materials.

5. The IABIN network should encourage the incorporation of metadata compilation into the institutional culture of its participating members.

CHAPTER 11 REFERENCES

Ahmed, K., Ayers, D., Birbek, M., Cousins, J., Dodds, D., Lubell, J., Nic, Miloslav., Rivers-Moore, D., Watt., A., Worden, R., Wrightson, A, 2001, Professional XML Meta Data, Wrox Press, Birmingham, UK.

Biodiversity Conservation Information System, 2000, *Framework for Information Sharing*, Busby, J.R, (Series Editor).

Chappell, D., A., Jewell, T., 2002, Java web services, O'Reilly, Sebastopol, CA.

FGDC, 1999, Content Standard for Digital Geospatial Metadata, Part 1: Biological Data Profile, Biological Data Working Group Federal Geographic Data Committee and USGS Biological Resources Division

Hobern, D., 2003, GBIF Metadata Standards, GBIF Secretariat.

McClarty, D., 2003, IABIN Portal Architecture, IABIN GEF PDF Project Report, Version 0.5 July 2003. http://www.iabin.net.

ANNEX 1 - Key Contacts

Dublin Core Metadata Initiative http://www.dublincore.org

Knowledge Network for Biocomplexity http://knb.ecoinformatics.org/index.jsp

Federal Geographic Data Committee http://www.fgdc.gov

International Organization for Standardization http://www.isotc211.org

National Biological Information Infrastructure http://www.nbii.gov

UK Office for Library Networking http://www.ukoln.ac.uk

ANNEX 2 - Acronyms and Abbreviations

BCIS	Biodiversity Conservation Information System
BDP	Biological Data Profile
CBD	Convention on Biological Diversity
CHM	Clearing House Mechanism
DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
DiGIR	Distributed Generic Information Retrieval
FGDC	Federal Geographic Data Committee (US)
IABIN	Inter-American Biodiversity Information Network
ISO	International Organization for Standards
NBN	National Biodiversity Network (UK)
KNB	Knowledge Network for Biocomplexity
RDF	Resource Description Framework
SOAP	Simple Object Access Protocol
UDDI	Universal Description, Discovery, and Integration
UML	Unified Modelling Language
URL	Uniform Resource Locator (a type of URI)
URI	Universal Resource Identifier
WDSL	Web Services Description Language
WS	Web Services
WWW	World Wide Web
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

ANNEX 3 - Glossary of Terms Used

Term	*Definition*
Metadata	"Data about data"
Web Services	"A web service is a piece of business logic, located somewhere on the Internet, that is accessible through standard-based Internet protocols such as HTTP or SMTP." (Chappell & Tyler, 2002).