Support to Building the Inter-American Biodiversity Information Network

Trust Fund #TF-030388

International initiatives in biodiversity vocabularies and thesauri

(Document 8)

July 2004

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

Review of Thesauri

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 project Building IABIN (Inter-American Biodiversity Information Network) and for assistance in supervision of the project. 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

Report Summary iv

Chapter 1 Introduction 1

1.1 Introduction 1

1.2 Definitions 1

1.2.1 Controlled vocabulary 1

1.2.2 Thesaurus 1

1.3 Thesaurus structure 1

1.4 Controlled vocabulary structure 2

1.5 Multilingual issues 2

1.6 Usage 2

1.6.1 Information cataloguing 3

1.6.2 Information discovery 3

1.7 On-line interoperable access 3

Chapter 2 Overview of main thesauri and controlled vocabularies in the field of environment 4

2.1.1 CBD Controlled Vocabulary 4

2.1.2 UNEP EnVoc 4

2.1.3 FAO AgroVoc 5

2.1.4 GEMET 2001 5

2.1.5 CIESIN (Center for International Earth Science Information Network) Indexing Vocabulary 6

2.1.6 NBII/CSA Biocomplexity Thesaurus 7

2.1.7 Other environmental thesauri and vocabularies 7

Chapter 3 Issues and Concerns 9

3.1 Multi-language 9

3.2 Linking to standards and ecosystem classification systems 9

3.3 Linking to taxonomic references 9

3.4 Usability and focus for the region 10

3.5 Comparative Review 10

Chapter 4 Recommendations 11

Annexes

ANNEX 1 - Acronyms and Abbreviations 122

Table of Tables

Table of Figures

Report Summary

Thesauri and controlled vocabularies are extremely useful in the standardisation of descriptions of entities. This report looks at a number of major thesauri and controlled vocabularies and highlights some of their strengths and weaknesses. The key aspects of thesauri and controlled vocabularies are also defined in this report.

Given the diversity of approaches available, it is difficult to give specific guidance on which to adopt for IABIN. However, it is clear that there are many ongoing attempts to bring together the various descriptive mechanisms. Not least of these was the meeting convened by UNEP in Geneva in April 2004.

The adoption and necessary adaptation of any thesauri will infer overhead costs in terms of management and support, which should not be underestimated.

CHAPTER 1 INTRODUCTION

1.1 Introduction

For many years people have struggled with the concept of how to describe things in a consistent way, such that others can recognise the description. For example, an automobile is the same as a car, but not the same as a vehicle. With the additional difficulties involved when people do not share a common language, it is clear that a structured approach is required. Thesauri and controlled vocabularies have been developed to meet these requirements.

1.2 Definitions

1.2.1 Controlled vocabulary

At its simplest, a controlled vocabulary is a restricted list of words (or short phrases) that can be used when trying to describe or search for an item. This can be extended to include relationships between terms that are synonyms, or even to include narrower and broader terms.

1.2.2 Thesaurus

A list of words (or short phrases) within an extensive hierarchy that that can be used when trying to describe or search for an item.

1.3 Thesaurus structure

A thesaurus is a collection of terms and relationships. Most are built around the set of relationships shown in Table 1.

Table 1: Thesaurus relationships

Broader term (BT)	Term A is a more general description than term B	Vehicle to car
Narrower term (NT)	Term A is a more specific description than term B (*inverse of "broader term"*)	Car to vehicle
Related term (RT)	Term A describes a related item to Term B	Car to motorcycle
Used for (UF)	Term A is the preferred term to use in describing term B	Car instead of horseless carriage
Use (USE)	Term A should not be used and instead Term B should be (*inverse of "used for"*)	Horseless carriage to car

With this set of relationships, a thesaurus (or controlled vocabulary) can be built to describe the interactions between different descriptors in a controlled context.

Example

Within the context of motorised vehicles, the following shows how these simple relationships can be used to build the thesaurus.

Motor Car

BT	Vehicle
UF	Horseless carriage
UF	Automobile
NT	Sports car
NT	Racing car
RT	motorcycle

Sports car

BT	Motor car
BT	Vehicle

Horseless carriage

USE	Motor car
BT	Vehicle

BT	**Broader term**
NT	**Narrower term**
RT	**Related term**
UF	**Used for**
USE	**Use**

1.4 Controlled vocabulary structure

Different controlled vocabularies implement different levels of relationships; some may not even implement any relationship between terms, only allowing a single term for a given topic.

Many of the large multi-lingual controlled environment vocabularies would be better described as thesauri, given the extensive implementation of relationships within them.

1.5 Multilingual issues

When developing a multilingual thesaurus, it is common to introduce a new qualifier to the information describing the language of the term. In this way, semantic differences between languages can also be accommodated.

Care needs to be taken when developing multilingual thesauri, as the precise meanings of words and phrases may differ between languages and countries. Even within a language, different countries will use different words or associate different meanings with the same word.

1.6 Usage

Thesauri and restricted vocabulary lists can primarily be used to assist IABIN in two key areas of knowledge management - information cataloguing and information discovery.

1.6.1 Information cataloguing

When cataloguing information and information sources, it is best to use a standard set of terms, and by implication meanings, to reference the content in the metadata. In this way the information will have increased compatibility.

1.6.2 Information discovery

As this paper shows, there is not a universally accepted single source of reference terminology. It is therefore necessary to assist users who are trying to discover information from the system by allowing them to enter their keyword, and then matching it up with any other appropriate keywords, prior to searching the knowledge store.

1.7 On-line interoperable access

Many of the initiatives involved in the production and maintenance of controlled vocabularies and thesauri are exploring the establishment of interoperable services. These include web services using XML and SOAP, Z39-50 services and other bespoke solutions. The technology is capable of supporting this style of access, but the sustainable financing of such solutions and large-scale implementations remains an issue. IABIN could make use of one or more of these services in the short-term with a modular approach, which would enable the retention of flexibility to adapt to new or emerging services in the future.

Given the location of IABIN, it may be appropriate to implement such interoperable services themselves for use throughout the network.

CHAPTER 2 OVERVIEW OF MAIN THESAURI AND CONTROLLED VOCABULARIES IN THE FIELD OF ENVIRONMENT

2.1.1 CBD Controlled Vocabulary

The CBD Controlled Vocabulary was developed with the intention of providing the CBD Secretariat with a list of terms that could be used as descriptors, i.e., metadata, for web pages on the Convention's web site including the Clearing-House Mechanism (CHM). The list is also recommended for use by CHM National Focal Points to describe the contents of their national CHM web sites.

The development and adoption of the CBD Controlled Vocabulary is expected to bring many benefits. Firstly, it will facilitate the searching, locating and retrieval of information by linking similar documents and resources with a unique term. Secondly, it would standardize descriptions of web sites, and so assist in efforts to make information interoperable within the CHM network, and with other websites related to the CBD that adopt the same vocabulary. Also, because terms are taken from official Convention documents, UNEP's EnVoc and the FAO's AgroVoc, the terms can be easily translated.

It is worth noting that the CBD Controlled Vocabulary is a work in progress, and is regularly updated with new terms as they are needed. Users can submit a term for inclusion in the CBD Controlled Vocabulary. However, terms submitted for inclusion should be from official CBD documents and/or other published controlled vocabularies or thesauri. The relationship of CBD Controlled Vocabulary terms to terms in other thesauri is recorded. There is not a numerical equivalence between the languages (not all descriptors are in every language).

The CBD Controlled Vocabulary is available from the CBD website in either PDF or XML format, and is available as an online query. The XML format provides an easy way to include the functionality of the Controlled Vocabulary within other systems. At present this will be limited by the lack of documentation available on the XML extraction tool operation.

Finally, it is worth noting that the CBD Controlled Vocabulary has evolved into a thesaurus through the inclusion of a hierarchy.

http://www.biodiv.org/doc/cbd-voc.aspx

2.1.2 UNEP EnVoc

EnVoc is a multilingual thesaurus with a controlled and structured vocabulary for use in indexing, storing and retrieval of environmental information. The latest edition contains categorised and alphabetical lists of subjects, together with a

KWIC (KeyWords in Context) list. This thesaurus is available in the six official United Nations languages. Previously, this publication was known as the INFOTERRA Thesaurus of Environmental Terms. Available for purchase as a printed document, this thesaurus has also been accessible for on-line querying at http://p5uni.ii.pw.edu.pl/infoterra/ although at the time of writing, the service was frequently unavailable.

EnVoc is updated on an irregular basis as funding and other resources within UNEP permit. Similarly, its extension into other languages has been dependent upon the availability of resources. There is not a numerical equivalence between the languages (not all descriptors are in every language).

http://www.unep.org/infoterra/overview.htm

2.1.3 FAO AgroVoc

The AgroVoc thesaurus is designed to cover the terminology of all subject fields of agriculture, forestry, fisheries, food and related domains, in order to describe the documents in a controlled system language. AgroVoc is currently at version 4 and is available for on-line browsing at http://www.fao.org/agrovoc. Currently AgroVoc supports seven languages: Arabic, Chinese, Czech, English, French, Spanish and Portuguese. The thesaurus is also available upon request for download for non-commercial usage. The website also features a link allowing users to submit terms for future inclusion. It is not clear whether there is a numerical equivalence between languages.

There is no clear indication of a mechanism for interoperable browsing of the thesaurus by other systems.

2.1.4 GEMET 2001

The GEneral Multilingual Environmental Thesaurus (GEMET) was developed by the European Environment Agency (EEA) and one of its “Topic Centres” working on a catalogue of datasources (ETC/CDS), together with the co-operation of international experts, to serve the needs of environmental information systems. Analysis and evaluation work led to a core terminology of 5,400 generalised environmental terms and their definitions. This vocabulary ensures validated indexing, cataloguing and retrieval within environmental information services as well as harmonised translations in the multilingual European network. EEA and ETC/CDS are continuously working towards a harmonised environmental terminology. The GEMET 2001 is provided as a polyhierarchically structured thesaurus and is now available in 19 languages: 4 more languages have been recently added (Bulgarian, Czech, Russian and Slovenian) besides English, Danish, Finnish, German, Dutch, Norwegian, Swedish, French, Greek, Italian, Portuguese, Spanish, Hungarian, Slovak, and American English.

GEMET provides a complete numerical equivalence (all descriptors have an equivalent) with the included languages. The semantic equivalence (correct correspondence of meaning between languages) has been separately ensured by the National Focal Points. The translation of GEMET into other languages is an ongoing activity.

http://www.mu.niedersachsen.de/cds/etc-cds_neu/software.html#GEMET

2.1.5 CIESIN (Center for International Earth Science Information Network) Indexing Vocabulary

The CIESIN Indexing Vocabulary was developed to index data resources and datasets related to the human interactions in global change. The "CIESIN Indexing Vocabulary" comprises two elements: CIESIN Indexing Terms and CIESIN Location Indexing Terms. At the time of writing the CIESIN Location Indexing Terms were last revised on 7th June 2002, but the CIESIN Indexing Terms have not been revised since 1997.

CIESIN Indexing Terms consist of a controlled thesaurus of socioeconomic and environmental terms arranged in nine Science Data Domains (commonly referred to as "Topics"). All of the terms are organised in a hierarchical relationship of broader to narrower terms, as follows:

Agriculture and Food Security
Economic Activity
Environmental Protection
Human and Environmental Health
Human Attitudes, Preferences, and Behaviour
Industry and Energy
Land and Freshwater Resources
Policy and Institutions
Population Dynamics

The CIESIN Location Indexing Terms is a controlled vocabulary developed to represent the geographical, geopolitical, and spatial coverage of socioeconomic and environmental data resources.

The location terms are arranged in categories, rather than hierarchically, as follows:

Continent/Region
Country/Island
Water Body

The database is available for on-line interrogation at http://sedac.ciesin.org/metadata/vocab/vocab_intro.html. However, there is no indication of any interoperable querying service.

2.1.6 NBII/CSA Biocomplexity Thesaurus

The Biocomplexity Thesaurus was developed in 2002-3 through a partnership between the NBII and CSA, a leading bibliographic database provider. The thesaurus was developed through the merger and reconciliation of the following five thesauri:

CSA Aquatic Sciences and Fisheries Thesaurus

CSA Life Sciences Thesaurus

CSA Pollution Thesaurus

CSA Sociological Thesaurus

CERES/NBII Thesaurus

The thesaurus is overseen by the NBII Thesaurus Working Group, which considers its expansion and addition/modification of terms. The thesaurus holds its terms with relationships including Subject Categories (SC).

The thesaurus can be accessed on-line at http://thesaurus.nbii.gov/. At present, this thesaurus appears to only provide results in English on the website.

2.1.7 Other environmental thesauri and vocabularies

The above six examples provide an insight in to the extensive range of available thesauri and vocabularies. There are many more available. The following list is not intended as an exhaustive list, but merely to indicate the extensive range of thesauri available.

EPA Terms of Environment (terms, definitions, abbreviations, acronyms)

EURODICATOM (multilingual EU terminology, with definitions)

Global Legal Information Network (GLIN) Thesaurus

HASSET: Humanities and Social Science Electronic Thesaurus (based on a UNESCO thesaurus)

Legislative Indexing Vocabulary (LIV; LC)

SOSIG Thesaurus, based on HASSET (linked to a search feature for the SOSIG Internet Catalogue)

OECD Macrothesaurus (Social and economic sciences; English, Spanish, French).

Population Multilingual Thesaurus (English, French, Spanish). POPIN Population thesaurus (demography, population studies. CICRED France)

Eurovoc Thesaurus (multilingual thesaurus covering the fields in which the European Communities are active; government, parliaments. 11 EU languages, plus transl. into a further 10 languages)

TRIBLEX Thesaurus, International Labour Organization (English, French)

Government of Canada Core Subject Thesaurus (English, French)

Canadian Government Thesauri and Controlled Vocabularies

CERES (California Environmental Resources Evaluation System)/NBII Thesaurus Partnership Project

Umweltdatenkatalog Online Thesaurus, Oesterreichisches Umweltbundesamt (in German and English)

Florida Environments Online Thesaurus

European Education Thesaurus (to download. Covering the 11 official languages of the European Union plus in Albanian, Croatian, Czech, Hungarian, Polish, Slovene and Turkish)

European Treasury Browser Thesaurus (European Schoolnet. For download: Multilingual thesaurus in twelve languages: Albanian, Danish, Dutch, English, Finnish, French, German, Greek, Hungarian, Italian, Spanish, Swedish aimed to index educational resources)

Cedefop's Multilingual Thesaurus of Vocational Training (VET-Thes 1992)

CHAPTER 3 ISSUES AND CONCERNS

3.1 Multi-language

The number of languages, and the complexity of providing thesauri or controlled vocabularies in multiple languages, should not be underestimated. The terms are often very specific and require a high level of language skill in order for equivalences and the hierarchy to be correctly established. In addition to this, some countries speak the same language but do not use the same terms to describe an entity.

The technical issues of dealing with different character sets are largely addressed through the adoption of the Unicode format for character storage. Although for many end users this is still an issue because of the problems of converting between different code pages that are not Unicode compatible.

A challenge for IABIN will be to access thesauri or controlled vocabularies that are usable by its constituents. These will need to ensure that the content is extensive and comprehensive whilst retaining an appropriate level of access.

3.2 Linking to standards and ecosystem classification systems

Many standards already exist for the exchange of information, and it is important that IABIN should not implement a new format for linking (without a very good reason). The standards supported and recognised by the International Standards Organisation (ISO) should be adopted where appropriate. For example, ISO 639 Codes for the representation of names of languages and the closely linked ISO 3166 Codes for the representation of names of countries and sub-divisions.

With the multiplicity of ecosystem classification systems available, and the very different premises under which they are all established, it is important for IABIN to identify the key ecosystems in the area and the classification schemes that are well-placed to serve those ecosystems. In terms of interoperability, IABIN would be well advised to look at the work undertaken by the European Environment Agency as part of EUNIS (European Nature Information System) at http://eunis.eea.eu.int in terms of cross-referencing classification systems. This system cross-references 20 different classification systems, and in some instances multiple versions, to allow a user to select an appropriate classification.

3.3 Linking to taxonomic references

With the continued expansion of Global Biodiversity Information Facility (GBIF) and the participation of IABIN as a node, the linkages to taxonomic references will continue to grow. The availability of taxonomic interoperable tools will provide both opportunities and issues in the future.

3.4 Usability and focus for the region

Key aspects of the usage of these thesauri and controlled vocabularies are the requirement to ensure that the constituency of IABIN supports their adoption and deployment. This will require major effort in terms of marketing and adaptation of the systems and content to meet the specific requirements within the region.

Given the issues with regard to Internet access and access speed in the region, it is likely (at least in the short-term) that the adoption of some of these standards will require the development of off-line as well as on-line tools. Those on-line tools which are developed, will need to be developed with the lowest possible bandwidth and hardware/software requirements possible.

3.5 Comparative Review

It is important to realise that these tools do vary from one to another. In order to demonstrate this, 50 words were chosen at random from the CBD controlled vocabulary and searched for in each of the other seven thesauri. The results are shown in the table below. This highlights the different coverage of all eight systems, and certainly does not mean that one is better than any other.

	AgroVoc	**GEMET 2001**	**NBII**	**UNESCO**	**CAB**	**INFO-TERRA**	**USGS**
Number of the 50 terms present	42	35	30	23	23	20	14
Number of times this was the preferred term	40	33	29	21	20	19	13

CHAPTER 4 RECOMMENDATIONS

IABIN should examine their current thesauri in the light of the developments being undertaken with the broader biodiversity/environment community, to ensure that they are collaborating in such a way that all parties benefit from the shared experiences.

With the work undertaken in Europe, particularly within the European Commission and its associated agencies, there are many initiatives that are already working in English, Spanish and Portuguese, which may be able to contribute to the future developments in IABIN.

The continued developments in the field on Information Communication Technology must continue to be monitored to ensure that appropriate technologies and protocols are being used.

UNEP convened a meeting of major participants in the field of multi-lingual thesauri in Geneva in April 2004. For the first time, this brought together the major providers of environmental terminologies with the aim of discussing the status of their terminologies, how they are applying new technologies, and how these resources can be "integrated" using new technologies. The meeting examined many of the duplicated thesaurus initiatives underway, and discussed opportunities to bring them together using the available web-based collaborative technologies as a coherent mechanism for developing a global multi-lingual system. Representatives from many of the major thesaurus initiatives participated.

The outcomes of the meeting can broadly summarised as follows:

1. There is an identified need for consolidation and collaboration on the development of thesauri and controlled vocabularies. The meeting proposed the over-arching organisation of EcoTerm to take this forward.

2. To hold an annual meeting to allow all parties to be appraised of recent developments and to foster collaborative efforts.

3. Implement a high-level web service for GEMET.

4. Technical issues relating to the outcome of the meeting will be included at a later date on a website.

It is strongly recommended that IABIN keeps a careful eye on the outcomes of the meeting, to look at the synergies between different thesauri and also the opportunities for them to draw upon the subsequent activities and participate in the multi-lingual developments.

ANNEX 1 - Acronyms and Abbreviations

CBD	Convention on Biological Diversity
CHM	Clearing House Mechanism
CSA	Cambridge Scientific Abstract
CERES	California Environmental Resources Evaluation System
EAA	European Environmental Agency
ENVOC	INFOTERRA Multilingual Thesaurus of Terms
ETC/CDS	European Topic Centre / Catalogue of Data Sources
GEMET	GEneral Multilingual Environment Thesaurus
IABIN	Inter-American Biodiversity Information Network
INFOTERRA	The global environmental information exchange network of the United Nations Environment Programme
ITIS	Integrated Taxonomic Information System
NBII	National Biological Information Infrastructure
TRS	Terminology Reference System
UNEP	United Nations Environment Programme
UNESCO	United Nations Educational, Scientific and Cultural Organisation
USDA	United States Department of Agriculture
U.S. EPA	United States Environment Protection Agency
USGS	United States Geological Survey