

CANRI Thesaurus - Initial Review

Prepared for

**Community Access to Natural Resource Information
(CANRI)**

Version	0.1
Status	Draft
Date	4 July, 2002
Job No.	17044

TABLE OF CONTENTS

1	Executive Summary	i
2	Scope	1
3	Background.....	1
4	Requirements and Current Classifications of Natural Resource Information Stakeholders.....	2
4.1	Nature.net	2
4.2	NSW Metadata Working Group	2
4.3	Department of Land and Water Conservation	3
4.4	Office of Information Technology	3
4.5	AGLS.....	3
4.6	NSW Local Government	4
4.7	National Land and Water Resources Audit (NLWRA).....	4
4.8	Northern Territory Office of Supervising Scientist	4
4.9	Manly Environmental Centre.....	4
4.10	Sydney Olympic Park Authority	4
4.11	National Parks and Wildlife Service (NPWS)	5
4.12	Australia New Zealand Environment and Conservation Council (ANZECC)	5
4.13	Australia New Zealand Land Information Council (ANZLIC)	5
4.14	Community Access to natural resource Information (CANRI)	5
4.15	California Environmental Resources Evaluation System (CERES)	6
4.16	Environment Australia (EA)	6
4.17	NSW Environment Protection Authority (EPA)	6
4.18	Global Spatial Data Infrastructure (GSDI)	6
4.19	National Agricultural Library (NAL, USA).....	7
4.20	Natural Resource Management Ministerial Committee (NRM)	7
4.21	National Land and Water Resources Audit (NLWRA)	7
4.22	Natural Resources Information Management System (NRIMS)	7
5	Significant Issues and Trends	7
6	Impact of initial review on project plan	8

DOCUMENT VERSION CONTROL

Title	CANRI Thesaurus - Initial Review
Document Status	Draft
Version	0.1
Date	4 July, 2002
Filename	C:_All_Projects_AUSDEC_Projects\17044_NSW_NRE_Thesaurus\Reports\CANRI_Data_Framework_Thesaurus_Initial_Review_v01.doc
Author(s)	D.R. Miller
Distribution	CANRI
Job No	17044

DOCUMENT HISTORY

Version	Comments	Date Issued
0.1	Initial draft	5 July 2002

DOCUMENT APPROVAL

Title	CANRI Thesaurus - Initial Review
Version	0.1
Approval	D.R. Miller
Position	Director
Date	5 July 2002

1 EXECUTIVE SUMMARY

This initial review sets out the available information to be reviewed in the development of the CANRI thesaurus. It specifies the sources and some of the significant issues raised by those sources that will need to be addressed in the review.

Australasian Spatial Data Exchange Centre

D.R. Miller
Director

2 SCOPE

The Consultancy brief for the development of the Data Framework Thesaurus required the preparation of an initial review report. That report is to cover:

- Requirements of CANRI stakeholders including alignment with legislation, government programs and classification systems,
- Scope of information and custodians to be considered in formulating the master thesaurus,
- Available classification systems for NR&E information,
- Significant issues and trends in classification and retrieval technologies and other relevant areas, and
- Any changes to the project plan as a result of the initial review.

These issues are dealt with in this report.

3 BACKGROUND

The Community Access to Natural Resources Information (CANRI) provides information products tailored for community-based local and regional environmental management in New South Wales. It is the first program of its kind to offer integrated access to maps and other data held at various Internet sites, by a diverse group of natural resource agencies and other stakeholders. CANRI is a collaborative initiative involving all NSW natural resource agencies.

The Data Framework project is funded by CANRI during 2001-2002 and overseen by the NRIMS Steering Group. The project will provide for the first time an overall structure for natural resources management datasets, endorsed by the key NSW natural resource management agencies. This high-level Data Framework will be published and reviewed by stakeholders. It will be used to create:

- a classification system for data served via CANRI;
- a taxonomy of natural resources terms for use in the CANRI catalogue, used by applications such as the Natural Resources Data Directory (NRDD) and Natural Resources Atlas (NRA);
- an enhanced CANRI homepage and map data selection tool allowing discovery of information within the branches of the Data Framework; and
- an XML schema to which more detailed schemas for individual data themes can be fitted as these become available.

The purpose of the consultancy is to develop a thematic structure for natural resources and environmental (NR&E) information across the NSW government. This structure will be submitted for formal endorsement by the natural resources management agencies under the framework of the Natural Resources Information Management Strategy (NRIMS).

4 REQUIREMENTS AND CURRENT CLASSIFICATIONS OF NATURAL RESOURCE INFORMATION STAKEHOLDERS

A review of the some 24 stakeholders was undertaken in June by phone and face to face meetings to determine:

- Requirements of CANRI stakeholders including alignment with legislation, government programs and classification systems,
- Scope of information and custodians to be considered in formulating the master thesaurus,
- Available classification systems for NR&E information

The following personnel were contacted:

Contact	Representing
Alan House	National Parks and Wildlife Service Water Information System (WISE)
Chris Williams	NSW State Library
Colin Creighton	National Land and Water Resource Audit (NLWRA)
John Busby	Australian Fisheries and Forests Agency (AFFA)
Kay Winkler	Environment Australia (EA)
Ken Bullock	NSW Office of Information Technology (OIT)
Marcia	National Land and Water Resource Audit (NLWRA)
Mary Gorman	National Office of Information Economics
Mike Thompson	Nature.net
Paul Elton	Natural Resource Management Ministerial Council (NRM) working groups on
Paul Kelly	Australia New Zealand Land Information Council (ANZLIC)
Peter Maganov	NSW Environment Protection Agency (EPA)
Pual Hartley	NSW Planning
Roger Jayasundra	Local Government and Shires Association
Roiss Hohanson	Community Access to Natural resource Information (CANRI)
Stewart Noble	Environment Australia (EA)
Sue Keyes	NSW Metadata Working Group

Listed below are the information sources collected during the review.

4.1 Nature.net

This group reported that there are many common strategic targets which are currently listed inconsistently across Catchment Blueprints in NSW. These Catchment Blueprints are submitted for accreditation and funding and should be reporting to a consistent set of targets. The group suggests that the classification below developed should be adopted within the proposed thesaurus.

The table of terms is given in Appendix 1.

4.2 NSW Metadata Working Group

The NSW Metadata Working Group supports the ANZLIC metadata Working Group. This has published the ANZLIC Metadata Guidelines which includes a controlled vocabulary of

terms for “subject keywords”. All natural resource information to be accessed by CANRI should have ANZLIC metadata to these guidelines. Within the guidelines are a set of subject keywords. These keywords represent a classification as well as a controlled vocabulary. There are some 31 groups and a total of 177 terms.

The content and classification is given in Appendix 9.

4.3 Department of Land and Water Conservation

The DLWC is currently undertaking the development of a thesaurus for the classification of paper based information held within the Department.

This thesaurus is being developed under the KeywordsAAA guidelines developed by NSW State Archives. This provides a “function and activity” thesaurus not a “subject” thesaurus. This is a fundamental difference from that proposed for the CANRI thesaurus which by nature is a subject thesaurus. This development is of little relevance to the CANRI thesaurus development.

The DLWC is also developing a web site. It is understood that this contains a classification of information available from the DLWC. At the time of writing of this report this data has not been made available to the consultant for review.

The DLWC has made available a glossary of terms being used to classify DLWC information sources. This comprises some 747 terms and the terms are either subject or activity descriptors. The listing appears in Appendix 12.

4.4 Office of Information Technology

The NSW OIT has provided a number of developments which may provide requirements for the proposed CANRI thesaurus.

Framework for development of a green pages directory. – This report, produced in April 2002, provides some options for providing access to collection level information resources. The report is a comprehensive review of the issue of discovering information. Of particular relevance to the proposed CANRI thesaurus is :

Search engines are constrained by their ability to correctly interpret a user’s request
Users want to search particular domains of information to see what is there, they do not have a specific information source in mind,
Searching should include the ability to access synonyms and broader and narrower terms.
It is highly desirable to establish a domain thesaurus

4.5 AGLS

The Online Council (www.noie.gov.au/oc) has endorsed the Australian Government Locator System (AGFLS) as the online resource discovery metadata standard, for use by all Australian governments. AGLS is based on the Dublin Core metadata standard. The AGLS specifies a field called “subject” that allows the description of the information by a controlled vocabulary. It is strongly recommended that thesaurus terms be used. The AGLS guide does

not give a thesaurus leaving this to subject experts to create and publish. In this context the proposed CABNRI thesaurus may become one such published thesaurus.

4.6 NSW Local Government

The State of the Environment Reporting undertaken by Local Government in NSW is governed by published environmental guidelines. These guidelines recognise 9 major classification headings and 54 topics. This classification of environmental information should be mappable into the proposed CANRI thesaurus. The classification shown below:

The list of terms and structure is given in Appendix 2.

4.7 National Land and Water Resources Audit (NLWRA)

The NLWRA has published a classification of natural resource information by way of the Australian National Resource Atlas (ANRA). This has the following classification:

The list of terms and structure is given in Appendix 3.

4.8 Northern Territory Office of Supervising Scientist

The Northern Territory Office of the Supervising Scientist has created a comprehensive thesaurus. A listing of their terms appears in Appendix 4. The thesaurus includes term relationships (broader/narrower terms) and is used as a controlled vocabulary for the classification of all reports and information produced by the Office. The thesaurus comprises some 185 terms, many of which are locality and company/agency names. There is no classification system in the thesaurus.

4.9 Manly Environmental Centre

The Manly Environmental Centre provides a catalog of environmental reference materials. It has been developed over some time and has a focus on providing a classification of the information that is publicly accessible and comprehensible. The classification is tabled below. It has a high level classification of some 12 terms under which terms are grouped. These terms are pitched at a community level.

The terms and structure are given in Appendix 5.

4.10 Sydney Olympic Park Authority

The Sydney Olympic Park Authority (SOPA) has developed a Thesaurus for the classification of the large amounts of information they have collected, particularly relating to the remediation of the Olympic Park site. They have adopted GEMET in total and added some 600 additional words that were not in the GEMET terms. These relate to localities and company/agency names (some 200 terms) and remediation terms (400 terms) that SOPA found were not well represented in the original GEMET terms.

SOPA have added the terms to the terms list only. They have not provided relationships to these terms (broader/narrower terms) as they use the list as a controlled vocabulary rather than as a search tool where the provision of broader and narrower terms can assist the user to search

for information. SOPA experience suggests that when classifying information that they have not included all the broader terms for a selected term. That is if “solar heating” is used to index information then that information is not tagged with the broader terms energy source” and “energy”. SOPA cite the reason for this being that the classification in GEMET includes terms that are not in current usage. In addition this was done to keep the memory overhead low for the index field.

SOPA has the ability to collect how users are using the thesaurus and to retire information (terms) not in use. This has not been done to date but is possible.

The list of additional terms is given in Appendix 6.

4.11 National Parks and Wildlife Service (NPWS)

The NPWS has developed a Water Information System (WISE) on a CD-ROM. This product is moving towards distribution using the Web. It contains a glossary of terms and classification. Which comprises some 30 top terms each with 3 to 5 sub-topics and then a total of about 300 terms.

The terms and structure are given in Appendix 7.

4.12 Australia New Zealand Environment and Conservation Council (ANZECC)

The ANZECC has developed a set of core environmental indicators for the reporting of the State of the Environment. These indicators are presented in a classification of theme/Issue and core indicator. This classification relates to much of the natural resource information to be accessed by CANRI and the classification is relevant to the proposed CANRI classification.

The classification is given in Appendix 8.

4.13 Australia New Zealand Land Information Council (ANZLIC)

ANZLIC have developed the ANZLIC metadata guidelines. All natural resource information to be accessed by CANRI should have ANZLIC metadata to these guidelines. Within the guidelines are a set of subject keywords. These keywords represent a classification as well as a controlled vocabulary. There are some 31 groups and a total of 177 terms.

The content and classification is given in Appendix 9.

4.14 Community Access to natural resource Information (CANRI)

CANRI has developed a web site. On this the home page displays a simple classification of the natural resource data sets already held in CANRI. The classification comprises some 7 groups with no further sub-classes. Some classes are non-specific (OTHER) or overlap (FLORA & FAUNA and VEGETATION).

The classification is given in Appendix 10.

4.15 California Environmental Resources Evaluation System (CERES)

The California Environmental Resources Evaluation System (CERES) is a comprehensive web site that includes a search engine that references a environmental thesaurus. This thesaurus has used the GEMET thesaurus in its development. The CERES web site has a classification of the terms in the thesaurus that is shown in Appendix 11. This gives 3 high level classes and some 39 sub-classes. This has been tailored for use on the web site ensuring that the user is not presented with too many choices at each stage in their search. The terms list, comprising nearly 200 terms) is to extensive to include in the appendix. The terms are fully relational (broader/narrower terms) and each data set registered on the site is linked to a term. It is does not appear that selecting a term for a data set also registers that data set with all the broader terms for that term. The hierarchy goes down at least 6 levels.

4.16 Environment Australia (EA)

EA has produced a thesaurus of environmental terms. The EA Thesaurus is broadly based on a subset of the *Thesaurus of Environmental Protection Terms*, Department of Environmental Protection, WA, 1995. Over the years, current specific Commonwealth terminology has been incorporated into the thesaurus. The thesaurus comprises some 296 terms but has not overlying classification scheme.

A copy of the thesaurus is given in Appendix 13.

In recent conversation EA has indicated that this thesaurus is no longer in use and that a whole of government thesaurus is being developed to replace it. That thesaurus is being developed at a high level and is a functional thesaurus by the National Office of Information Economics (NOIE) and is called TAGS (Thesaurus of Australian Government Subjects). Currently It contains few environmental/natural resource terms and assumes that individual agencies will populate the thesaurus for their particular functional areas.

4.17 NSW Environment Protection Authority (EPA)

The NSW EPA has produced, since 1993, a series of reports on the State of the Environment. Each report has classified environmental data into a series of themes, issues and indicators. These have been successively refined over each report. A copy of the structure and terms used for the 2003 State of the Environment report is given in Appendix 14.

4.18 Global Spatial Data Infrastructure (GSDI)

The Global Spatial Data Infrastructure supports ready global access to geographic information. This is achieved through the coordinated actions of nations and organisations that promote awareness and implementation of complimentary policies, common standards and effective mechanisms for the development and availability of interoperable digital geographic data and technologies to support decision making at all scales for multiple purposes. These actions encompass the policies, organisational remits, data, technologies, standards, delivery mechanisms, and financial and human resources necessary to ensure that those working at the global and regional scale are not impeded in meeting their objectives. Its web site presents spatial data in a classification of some 21 categories. These are given in Appendix 15.

4.19 National Agricultural Library (NAL, USA)

The NAL has developed and published a Agricultural Thesaurus. This was developed for anyone describing, organising, and classifying agricultural resources.

The thesaurus includes a classification of some 17 “subject categories”, one of which is “Natural resources, Earth and Environmental Sciences”. This subject category alone has some 2117 terms. This sub-set is included in Appendix 17.

4.20 Natural Resource Management Ministerial Committee (NRM)

The NRM has published a framework for reporting national level environmental indicators (standards & targets and monitoring & evaluation). A review of these yields some 4 themes and 17 indicators. The terms do contain multiple terms. A copy of the classification and terms is given in Appendix 18.

4.21 National Land and Water Resources Audit (NLWRA)

The NLWRA has developed a web site – the Australian Natural Resources Atlas. The web site gives a classification of natural resource information with 11 themes and some 156 terms. The classification is given in Appendix 19.

4.22 Natural Resources Information Management System (NRIMS)

NRIMS has published a classification of natural resource information as part of its strategy. The strategy lists some 18 themes as listed in Appendix 20.

5 SIGNIFICANT ISSUES AND TRENDS

The review of information sources detailed above has yielded a number of significant issues that relate to the proposed CANRI Thesaurus. These are discussed below.

- The requirements of the Online Council (www.noie.gov.au/oc) has endorsed the AGLS as the online resource discovery metadata standard for use by all Australian governments. This standard, based on the Dublin Core, does not define a thesaurus to be used but requires the reference to a thesaurus or controlled vocabulary list for subject keywords. The proposed CANRI thesaurus will need to be published so that this criteria can be met.
- AGLS recommends that subject keyword lists are best developed by subject experts and the current TAGS (Thesaurus of Australian Government Subjects) is adopting this principle. Currently a subject keyword list does not exist under TAGS for the natural resource theme.
- Overseas-developed thesauri (GEMET, CERES) are very comprehensive, with many containing many thousands of terms. These have been developed over years. The creation of a new thesauri which differs greatly from them would mean that future attempts at international cataloguing of natural resource data sets would be impeded. However, most of the international thesauri also include a classification system which

- allows the thesauri terms to be grouped as distinct from their inter-relationships (as in broader and narrower terms). These classifications reflect strong political and regional bias.
- Natural resources is a very broad theme and this is reflected in the many thesauri and glossaries, both international and Australian. There is no clear cut boundary for subject terms as to whether they fall clearly in or clearly out of such a theme.
 - Polyhierarchical thesauri allow multiple groupings of terms to be developed based on the one glossary of terms. As such multiple classifications or “views” of the terms can be easily accommodated. On this basis the proposed CANRI thesaurus should have this capability.
 - Modern search technology in web search engines allows the use of the term relationships (broader and narrower terms) to be enabled rather than the set classification presentation of terms in a hierarchy. The proposed CANRI thesaurus should enable this capability to be exploited.
 - The issue of using the term relationships raises an issue as to whether an information source is tagged with a term and its broader terms. If this is done then the information source can be found at every level above the term. So an information source describing “solar energy” will also appear under the term “energy source” and under “energy”. The issue is to how many broader terms are tagged as this has an impact on the size of the subject keyword index field. Current experience varies. SOPA, in their implementation have not adopted tagging with broader terms but others (CERES) may have.
 - Modern web search technology can now accommodate fuzzy searches. These engines should be able to exploit the term relationships (related term, broader/narrower term). This means that the proposed CANRI thesaurus must have this structure and not be a simple glossary of terms.

6 IMPACT OF INITIAL REVIEW ON PROJECT PLAN

The most significant impact on the project plan from the initial review is the need to consider multiple classifications for natural resource information. There are competing needs in determining the most appropriate structure to adopt. Currently the EPA and ANZECC have well developed and published classifications but these do not cover the whole subject field. Other high level classifications, such as NRIMS, represent a good starting point but have a combination of high and low level terms in the top level classification (for example, biodiversity- an all encompassing theme and restricted sites – a specific term). The project plan will have to reflect a greater degree of consideration of these competing classifications at the expense of the consideration of the terms that should make up the thesaurus.