Harmonizing Taxonomies for e-Government: a briefing note

1. Why harmonize taxonomies?

The aim is to be able to find information on a given subject, no matter what terminology or language it was expressed in at the time of publication or input.

Typically the form of words used to describe an information need differs from the expressions used by authors and publishers of relevant resources. A taxonomy can act as an interface between the author and the user, mediating the form of language. If everyone used the same taxonomy, then a given concept would always be represented the same way in the interface. In practice however many different taxonomies are used. Harmonization should enable easy translation from one taxonomy to another.

2. Technical Background

2.1 What do we mean by taxonomy?

For the purposes of this note, 'taxonomy' means any controlled vocabulary used to support the discovery of resources by subject, including browsing as well as searching. Classification schemes, thesauri, subject heading schemes and ontologies may all be used as controlled vocabularies. These different types of taxonomy represent concepts in different ways: some as terms, some as codes or categories. In this note, the term 'categories' will be used for all of these types of concept representation.

Taxonomies are very often presented as tree structures, allowing the user to browse or 'drill down' from a limited number of high-level categories to progressively more specific categories at lower levels of the hierarchical tree.

Taxonomies may also support the selection and conversion of keywords entered without hierarchical browsing.

Taxonomies are often used as encoding schemes for the subject element of metadata. In other words, when a resource is published on a website or enters into some other system, relevant categories from the taxonomy are selected and entered into the metadata as values for the subject element. The presence of these values in the metadata makes it easy to retrieve the resource when the taxonomy is used for searching or browsing.

2.2 Harmonization and mappings

The simplest method of harmonization is for everyone to use the same taxonomy from the start. If this is not possible, harmonization may be achieved by establishing mappings between pairs of categories in different taxonomies.

A mapping is a relationship between categories in different taxonomies. The commonest type of mapping is 'equivalence', that is to say the mapped categories represent one identical concept, even though it may be expressed in different words or codes. Other possible mapping types include hierarchical, partial, associative, etc.

Often equivalence is accepted pragmatically between categories that are not strictly synonymous. For example, "devolution" might be accepted as equivalent to "devolved government" even though these do not quite represent the same concept. Similarly, "seagulls" might be mapped to "birds", or "rubbish bins" to "household waste". Such inexact mappings may give good results in one context but fail in another. Also, some are irreversible – all seagulls can be accepted as birds, but conversion of "birds" to "seagulls" could give some

very unacceptable results. For these reasons, it is common to restrict a set of mappings to certain applications, and/or to use them as one-way conversions only.

When two taxonomies are to be harmonized, the systematic approach is to identify and establish mappings between all the equivalence pairs. Other types of mapping are optional. (For example, a hierarchical mapping could be established between "seagulls" and "birds" and this would be helpful in determining when it is or is not safe to switch the categories). The mappings data may be stored in the maintenance databases of either one of the taxonomies, or both, or in a separate database.

2.3 Where and how are mappings applied?

The usual purpose of preparing a database of mappings is to use the data for switching or augmenting categories in storage and retrieval applications, either at the point where resources are entered into a system, or at the point of retrieval, or both. The process may be entirely automatic or subject to human review.

Mapping at the point of input is cost-effective when there is an established need for the resources to be retrievable using the target taxonomy as well as the source taxonomy. If the original indexing (tagging) is done by a human operator, then the validity of the mapping should preferably be checked at the same time, when the context of the resource is known. But if automatic methods are used for the original indexing, then rather than mapping to the target taxonomy, it may be equally cost-effective to use automatic categorization for this too.

There are similar options when mapping is applied at the point of retrieval. If the user expresses his query using categories from a source vocabulary, the mappings database may convert them to the corresponding ones in the target taxonomy. But mappings may not be needed if the user expresses his query in free text, and this is converted directly to the target taxonomy. The best option to choose depends on the capabilities of the user as well as the technology available and the quality of the mappings.

These choices do not arise if the user's approach is by browsing rather than searching. The browse interface compels the user to select one or more taxonomy categories. The latter are then mapped to the corresponding categories in the taxonomy/ies applicable to the resources to be searched.

2.4 Multiple taxonomies

When there are more than two taxonomies in play, mappings may be established between any or all of the possible pairs in the set. To reduce the amount of work involved, one can select one of the taxonomies as the 'backbone', and treat all the others as 'ribs' which link to the backbone but not to each other. Any attempt to convert categories from one of the rib taxonomies to another will have to be mediated by the backbone, and this may reduce the quality of the conversion. The alternative of direct mappings between all pairs involves more work but is more versatile and capable of higher quality results.

2.5 Performance issues

Switching from one taxonomy to another usually introduces loss of precision and/or recall. There maybe no loss when the equivalence pairs are truly synonymous, but typically there is some difference in scope between the original category and the one to which it is mapped, so losses are inevitable.

Further losses occur when

a) Mapping is indirect, perhaps routed via a backbone taxonomy
b) One-way mappings are applied in the opposite direction

At the stage of developing the mappings, automated methods can lead to errors. For example a category named "depression" referring to a state of mental health in one taxonomy, could be

confused with a category "depression" in another taxonomy, where it refers to economic conditions. However, some automation is essential for cost-effectiveness when mapping large taxonomies. It is important to strike the right balance between human and automatic processes.

2.6 Standardization issues

The construction of taxonomies is a specialist job, in which relatively few people have experience. Standards such as ISO 2788, ISO 5964 and ANSI/NISO Z39.19 provide some guidance, although they all need updating and are not always followed rigorously.

Even fewer people have experience of setting up mappings, and the standards do not (yet) cover this aspect at all. Guidance would be useful at two levels: firstly the intellectual criteria for establishing mappings, and secondly some methodological guidelines for automating the process.

Additionally, the formats and protocols for data exchange urgently need to be standardized. This will facilitate the use of taxonomy data in applications such as search engines, portal interfaces and content management systems; also the switching from one taxonomy to another.

3. Needs in the context of e-government

Public sector needs are still at a very early stage of exploration. In contrast, the more advanced mapping projects under way (See Appendix, Section A) are planned by or for the academic sector, with useful spin-offs for researchers in all sectors. The taxonomies selected include most of the large universal schemes, and the most widely used thesauri. The applications explored typically allow the expert staff of one centralized conversion service to make results available to a very wide community of users. This makes excellent economic sense, since taxonomy development work is expensive, but does not address taxonomies that serve the citizen or business applications.

Additional challenges for the public sector include the huge diversity of styles of taxonomy in use, the need for data manipulation in thousands of different organizations that issue information/services, and the general lack of software and expertise. Mappings between the DDC and MeSH (see Appendix) are not much help to organizations wanting to convert from their in-house taxonomy to, say, a standard business classification.

Specifically, the e-government sector could benefit from:

- **Standards** for taxonomy building and mapping, and for data interchange
- **Guidance** on techniques, including automation.
- **Demonstrators** of how the mapped taxonomies can be used, both for meta-tagging and for retrieval
- **Products** – some mapped taxonomies available free or for purchase, and software to handle taxonomy data.

Useful projects could be considered along the following lines:

A) Support, encourage and publicize the standardization projects described in Appendix Section B. Improve communication to avoid conflicting standards.

B) Support selected authorities in developing mapping schemes from their own taxonomies to standard e-government taxonomies such as the GCL in the UK. Possibly involve the vendors of software for content management systems, EDRMS and other applications that support meta-tagging. Disseminate the results via demonstrator systems, discussion forums, etc.

C) Encourage the vendors of search engines, portal software and EDRMS to support and exploit mappings in user interfaces.

APPENDIX: Current projects and initiatives

A. Harmonization projects

A1 HILT

http://hilt.cdlr.strath.ac.uk/
Project supporting access to the archival/digital collections of universities, museums and other bodies, indexed with a variety of vocabularies including DDC (Dewey Decimal classification), UDC (Universal Decimal Classification), LCSH (Library of Congress Subject Headings), Art & Architecture Thesaurus, UNESCO Thesaurus and others. The approach will use the 'backbone' model. Mappings will be established between DDC and each of the other vocabularies. A pilot project is under way, funded by JISC.

A2 OCLC Terminology Services

http://www.oclc.org/research/projects/termservices
The project will establish mappings between some very widely used taxonomies (including DDC, LCC(Library of Congress Classification), LCSH, MeSH (Medical Subject Headings) and Eric Thesaurus) in order to offer accessible, web-based terminology services.
Taxonomies are first encoded using MARC format, then matching categories are detected automatically. Human review of the mappings has been found necessary for some types of match. Access is likely to be through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

A3 RENARDUS

http:www.renardus.org/
A broker service that integrates access to European subject gateways, using Z39.50 protocol. One-way mappings from DDC to local taxonomies enable cross-browsing.

A4 UMLS (Unified Medical Language System)

http://www.nlm.nih.gov/pubs/factsheets/umls.html
This is one of the longest established harmonization projects, bringing together over 100 biomedical taxonomies for multiple medical applications, mostly research-oriented. Coordinated by the US National library of Medicine, it has collaborators from many different countries and language groups.

A5 Some UK public sector projects

The UK's e-Government Metadata Standard (eGMS) requires public sector bodies to use a taxonomy called the GCL, bringing the need for harmonization with the many taxonomies used by the different organizations. Several of these are now developing their own methodologies in-house.
The SeamlessUK project, which offers portal access to the resources of several different local authorities and voluntary bodies, is studying the options for harmonizing its own taxonomy with the GCL and also the taxonomies of all the collaborating bodies. Mapping tools are under development, and will be integrated with the interface for inputting resources as well as the search system.
The LAWs project (developing standards for Local Authority Websites) is also preparing mappings from at least two local authority taxonomies to the GCL. The mappings are not yet integrated into downstream applications.

B Standardization and discussion forums

B1 NKOS

http://nkos.slis.kent.edu/
NKOS (Networked Knowledge Organization Systems/Services) is an active forum for the discussion of functional and data models enabling taxonomies as networked interactive information services to support the description and retrieval of diverse information resources through the Internet. It organizes regular meetings n the USA and Europe.

An overview of "Activities to advance the powerful use of vocabularies in the digital environment" is available at http://www.lub.lu.se/%7Etraugott/drafts/seattlespec-vocab.html

Discussions at a workshop on "Evolving standards" held in 2003 (See report at http://dlib.ejournal.ascc.net/dlib/september03/09inbrief.html#TUDHOPE) included:

- maintaining an inventory ('namespace') with standard definitions of core taxonomy relationships
- the need for standards on distributed taxonomy protocols and data exchange
- relationships and frameworks for mapping between taxonomies.

Upcoming NKOS Workshops include one in Tucson, USA in June 2004 and another in Bath, UK in September 2004 in conjunction with the ECDL conference (http://www.ecdl2004.org/).

B2 British and International Standards for thesauri

BSI is currently revising and extending the British standards for monolingual and multilingual thesauri, BS5723 and BS 6723 respectively, which are identical to ISO 2788 and ISO5964. The two will be brought together in one new 5-part standard. Part 4 of the new standard will cover interoperability between taxonomies, including multilingual thesauri and mapping. Part 5 will cover syntactic aspects of interoperability, including formats and protocols for data exchange.

B3 ANSI/NISO thesaurus standard

NISO is currently planning revision of ANSI/NISO Z39.19, the American standard for monolingual thesauri. One aim is to provide for a broader group of controlled vocabularies including ontologies, classifications, taxonomies, and subject headings, in addition to standard thesauri. Interoperability is another area of concern, but there has been no mention yet of including mapping or other forms of taxonomy harmonization.

B4 CEN/ISSS Activities

CEN/ISSS has already issued a Workshop Agreement document on "Controlled Vocabularies for Learning Object Metadata: typology, impact analysis, guidelines and a web-based Vocabularies Registry". A new project is just starting on harmonization of vocabularies.

B5 SWAD

http://www.w3c.rl.ac.uk/SWAD/thesaurus.html
SWAD-Europe conducts research, creates demonstrations, outreach materials, software and services to support W3C's Semantic Web Activity. One workpackage concerns taxonomies, including for example an RDF schema for inter-thesaurus mapping.