

Digital repository at Nanyang Technological University : implementing a subject metadata scheme

Sam, Lena Choon Foong.; Muralidharan, Padmaja.; Tan, Han Yong.; Goh, Su Nee.

2009

Sam, L. C. F., Muralidharan, P., Tan, H. Y. & Goh, S. N. (2009). Digital repository at Nanyang Technological University : implementing a subject metadata scheme. In Conference proceedings : towards dynamic libraries and information services in Southeast Asian countries/Congress of Southeast Asian Librarians (CONSAL XIV) (pp. 258-272). Hanoi, Vietnam.

<https://hdl.handle.net/10356/91266>

Digital Repository at Nanyang Technological University – Implementing a Subject Metadata Scheme

Sam, L. C. F., Muralidharan, P., Tan, H. Y., Goh, S. N.

Abstract

One information service that has emerged in libraries with the advancement of digital technologies is the online institutional repository. DR-NTU is the digital institutional repository jointly developed by NTU Library and the Wee Kim Wee School of Communication and Information, using open source software DSpace from MIT, to facilitate the capture, storage and preservation of the intellectual output of the NTU community.

Aside from using keywords and metadata elements such as title, author and date to search and browse the collections, the DR-NTU user interface is designed to also allow subject search and browse. This feature would provide value-added library services to the user for navigating information within the digital environment, allowing more precise subject search and browse by the use of controlled vocabulary. NTU Library's Bibliographic Services Division assisted in developing a metadata scheme which would support these information searching and retrieval requirements. This paper provides an overview of the policies, procedures and relevant issues in creating, testing and implementing the DR-NTU subject taxonomy.

Scope and purpose

The collection

The Digital Repository at Nanyang Technological University (DR-NTU) aims to capture, store and preserve the intellectual output of Nanyang Technological University (NTU) and, where possible, make it available to the global research community. The repository contains open access collections as well as restricted collections which are accessible only to the NTU Community and require user authentication to view the full text.

DR-NTU accepts the following types of electronic publications into the repository:

To the Restricted access collection – theses, dissertations, research reports, student reports (written for academic degrees, internships, professional attachments). The authors are mostly students although some research reports are by staff or graduate students); and

To the Open access collection - journal articles, conference papers, working papers and commentaries; the authors are mostly staff of NTU.

The number of items in the whole repository at this point in time is more than 15,000. The earliest publication dates back to the 1970s and the collection is expected to increase at a rate of about 1,000 items per year.

Stakeholders and system

Staff and students may contribute publications to DR-NTU from three main sub-communities – schools, research centers and administrative departments in NTU. The NTU digital repository was jointly developed by NTU Library and the Wee Kim Wee School of Communication and Information using the open source software DSpace from MIT. NTU's Centre for IT Services provides technical support for the repository system and technology.

Purpose

User-centred objectives

Recent developments in information organization within the international library community have shifted to return focus on user requirements; in NTU Library's strategic plan one of the 4 main areas of concern has been to focus all activities on the user. Guided by this user-oriented philosophy, the Library has looked into ways to apply its core capability in the management and organization of information in the digital repository collection.

NTU Library's role in the DR-NTU project is to develop, design and manage the information architecture and system technology for the digital repository, bringing together the fundamental requirements of the organisation context, users and content. The project is led by the Library Technology Division with metadata support from Bibliographic Services Division and user liaison provided by the subject librarians from each school in NTU.

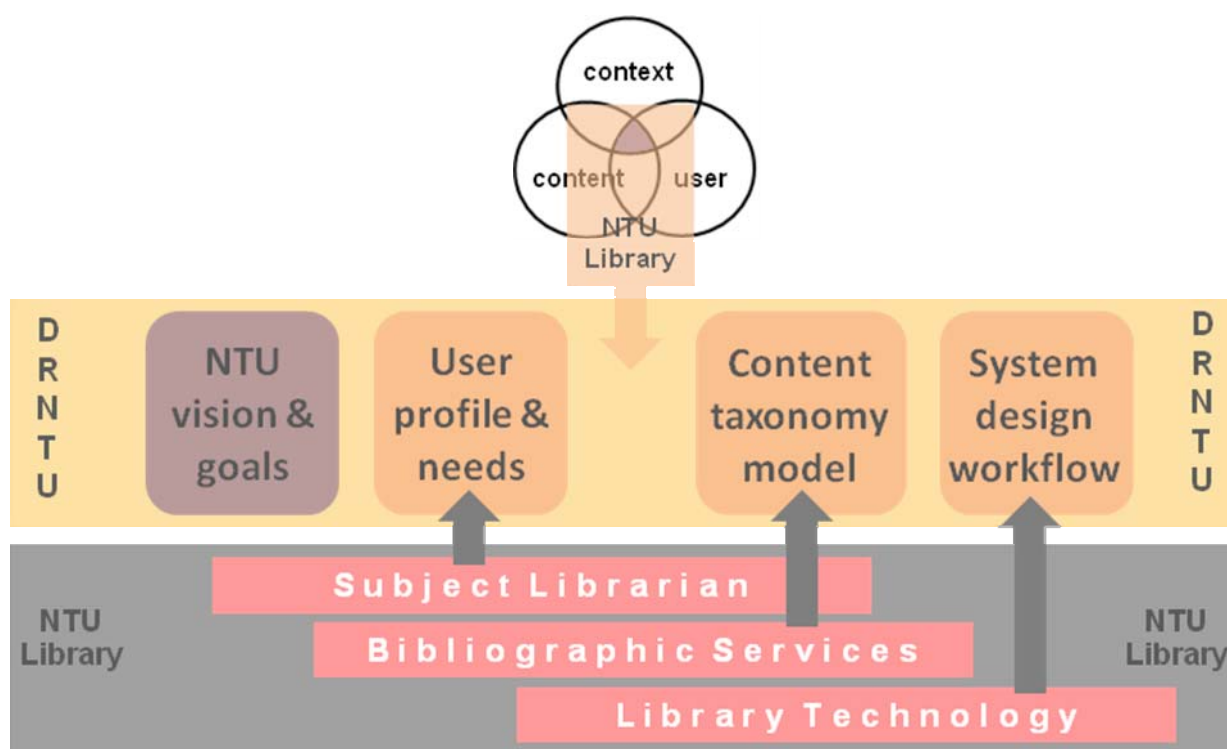


Figure 1. Library capability in DR-NTU information architecture

In coming up with a strategy for the digital information infrastructure to be used in the repository, the Library looked at a collaborative solution where the Library and community potential would be equally realised. It is expected that the initial higher resource deployment at the development stage while the collection size is modest would facilitate:

- opportunity for integrated participatory workflows involving the user and the Library,
- improved user awareness of subject categorization and other knowledge representation tools
- continually organised information growth in a systematic and consistent manner,
- efficient change management processes and extensibility.

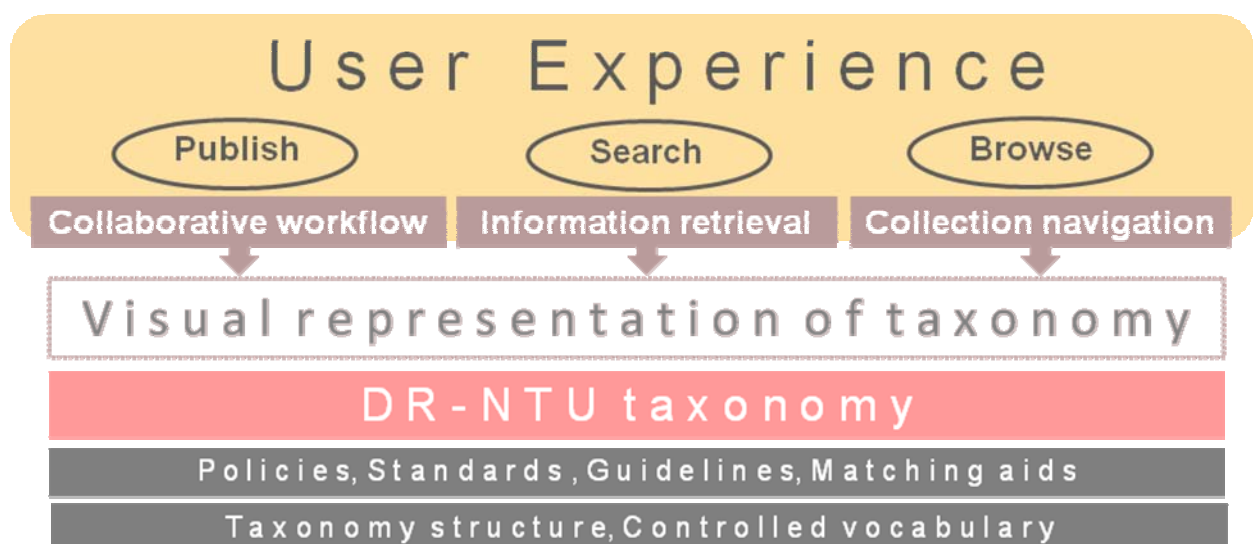
Navigation capability

The DR-NTU information organisation model is intended to add value to the user experience beyond the basic functional requirements of an information retrieval system; FRBR describes in these, the context of a bibliographic system, as finding, identifying, selecting and obtaining an information object. It has been suggested that a full-featured and effective information

retrieval system should also include navigation features. The intention is to allow users the opportunity to search and browse the DR-NTU collection by subject area in addition to keyword searching.

In the context of information organisation, taxonomies refer to structures that provide a way of classifying knowledge resources into a series of hierarchical or other systematic groups for easy identification or location. Within the business environments, taxonomies are used to provide increase knowledge sharing by effective content management (identification, location and organisation) and improve information access and business (navigation and visualization).

The Library conceptualised the subject metadata scheme as a subject taxonomy allowing users navigation within information structures developed from systematic analysis of context, user and content. Depending on the their information requirements when interacting with the DR-NTU interface, the taxonomy allows users to select a subject tag to their submission, to search for a specific item, as well as browse the collection by subject categories in an undirected search. Within the digital environment, the taxonomy structure and its visual representation intuitively aid users in knowledge discovery in the same way classification achieves this in the physical collection.



Picture 2. DR-NTU taxonomy model

NTU Library's Bibliographic Services Division assisted in developing the subject taxonomy to which would support the information searching and retrieval requirements of the digital repository. These two broad objectives - focusing on adding value to the user experience by providing navigational capability to the DR-NTU information organisation system - have shaped the approach when designing and building the subject taxonomy. The next section will describe the subject taxonomy development process and share some issues from its implementation.

Taxonomy Development

The benefits of a good taxonomy, Weinstein says, are that users can "navigate from need to resource consistently and quickly." A good taxonomy "allows an organization to inventory and monitor knowledge resources based on a structured understanding of user and community needs." (Pack, 2002)

The subject taxonomy development is collaborative work involving Bibliographic Services Division providing taxonomy expertise, subject librarians, who provide in-depth knowledge of user culture and content, and the Library Technology Division who design and develop the subject taxonomy interface. The flowchart in Picture 3 illustrates an overview of the DR-NTU key taxonomy development stages. Although it depicts a sequential process, in reality, some stages may occur simultaneously or be reiterative, especially at review and refinement.

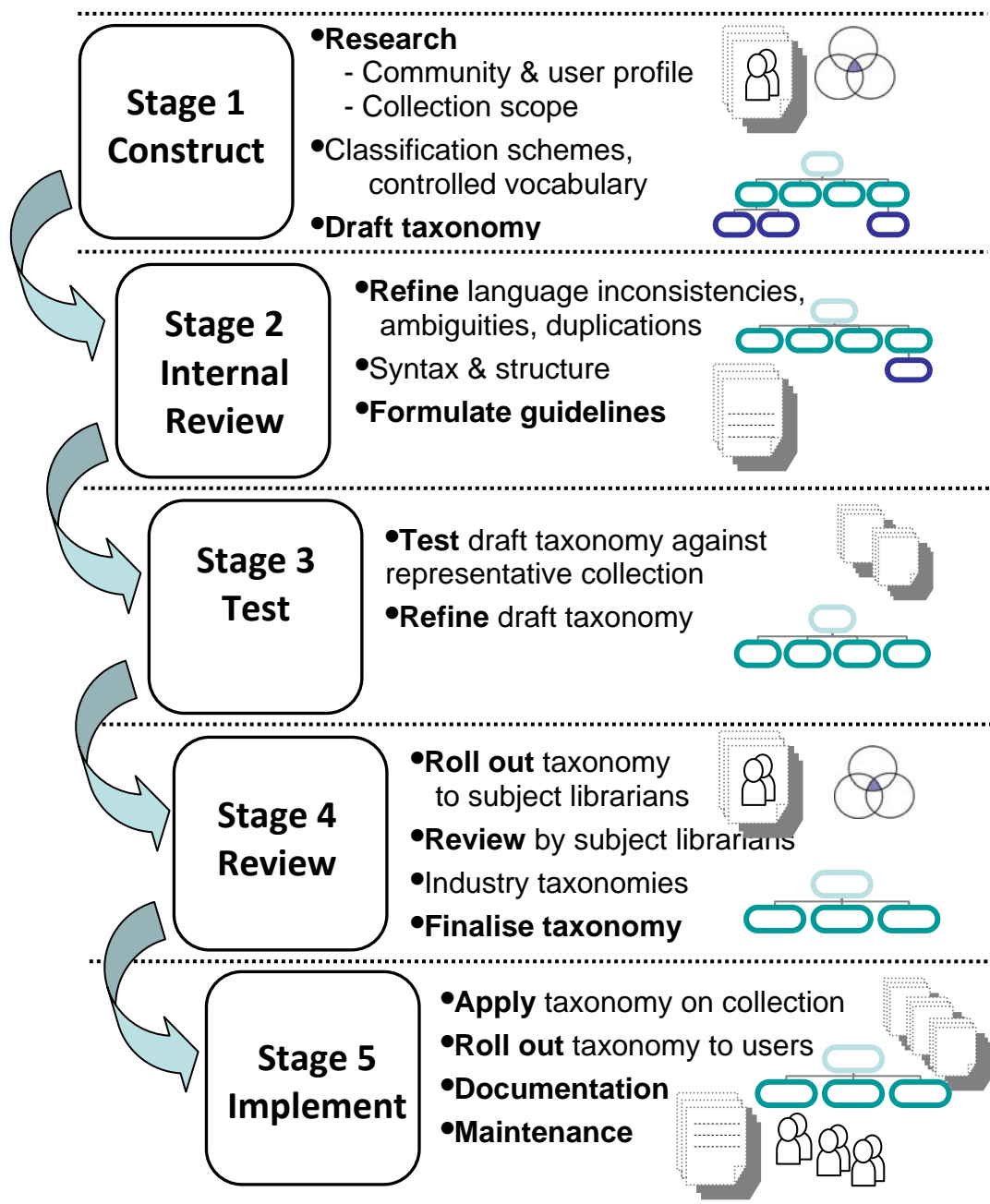


Figure 3. DR-NTU taxonomy development flowchart

Stage 1 Construct draft version of DR-NTU taxonomy

The DR-NTU Taxonomy was constructed using the Library of Congress Classification (LCC) as a guideline for its structure and as a source for the subject language. The advantages are that some users may already be familiar with the structure and language since it is being used in NTU libraries. The LCC system is also integrated into the collection development planning of each NTU school community profile information requirements.

Additionally from the content management perspective, the use of LCC would allow opportunity for cross mapping of subject categories for reference and subject tagging aids.

The figure below shows an outline of the LCC:

A -- GENERAL WORKS
B -- PHILOSOPHY. PSYCHOLOGY. RELIGION
C -- AUXILIARY SCIENCES OF HISTORY
D -- HISTORY: GENERAL AND OLD WORLD
E -- HISTORY: AMERICA
F -- HISTORY: AMERICA
G -- GEOGRAPHY. ANTHROPOLOGY. RECREATION
H -- SOCIAL SCIENCES
J -- POLITICAL SCIENCE
K -- LAW
L -- EDUCATION
M -- MUSIC AND BOOKS ON MUSIC
N -- FINE ARTS
P -- LANGUAGE AND LITERATURE
Q -- SCIENCE
R -- MEDICINE
S -- AGRICULTURE
T -- TECHNOLOGY
U -- MILITARY SCIENCE
V -- NAVAL SCIENCE
Z -- BIBLIOGRAPHY. LIBRARY SCIENCE. INFORMATION RESOURCES
(GENERAL)

Figure 4 Library of Congress Classification outline

Although the DR-NTU Taxonomy uses LCC as a basis for its structure and content, there were considerable modifications made to suit the specific subject area requirements of the community, user and collection. The selected LCC subject captions are the closest match to subject categories derived from research into each school and research centre in NTU and their respective publications. The customizations made were based on the principle of enhancing and adding value to the user experience; some of the main modifications include the following:

- Only subject category captions relevant to NTU curriculum and research areas were selected; these may or may not be arranged according to the hierarchical structure in LCC. Instead the broad subject categories were rearranged to reflect the NTU community representative subject structure and content resulting in a more compact tree structure for the draft taxonomy.
- DR-NTU taxonomy does not adopt the enumeration system of the LCC since DSpace software allows a structured hierarchy display of taxonomy with or without enumeration. There may not be strong advantages of enumerating the taxonomy for the user based on current search and browse features of the DSpace software. Instead the main subject categories were indexed alphabetically for user convenience.

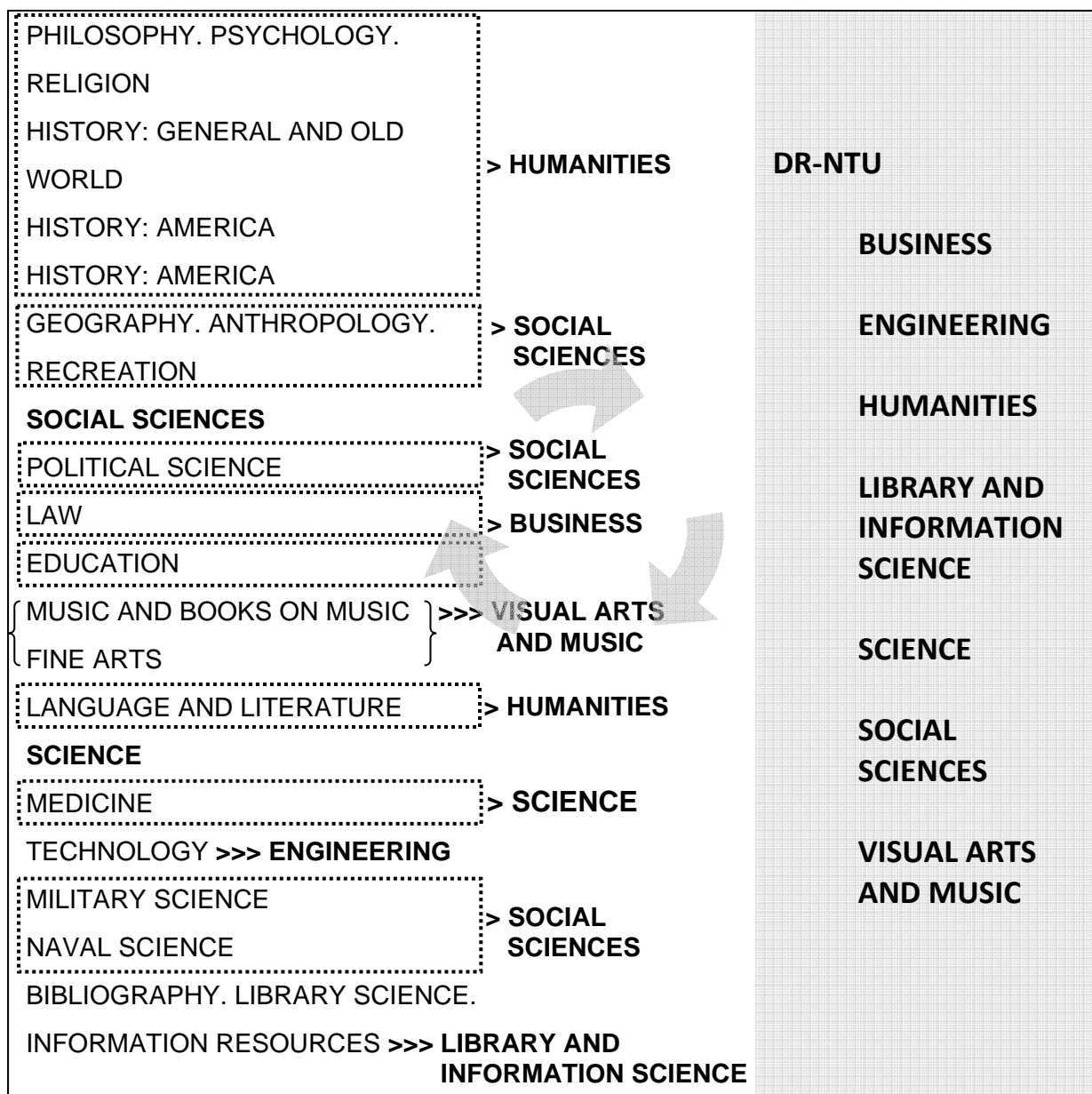


Figure 5 Modifications to Library of Congress Classification outline

Stage 2: Review draft version of DR-NTU Taxonomy

Stage 2 involves refinement of the draft taxonomy after an in-depth review of the broad issues concerned with consistencies across all main subject categories such as duplication of topics in more than one category, selection and arrangement of first level categories and consideration of user navigation.

A good taxonomy, according to Charles Weinstein, director of solution development at the content categorization company Sopheon, is one in which content is distributed evenly across the classification scheme. "The depth of the taxonomy should be relatively uniform," he said. When some categories have too much or too little information, "It usually means that the people didn't understand the nature of the content they were classifying, or they believe that they had more or less than they actually did." (Pack, 2002)

The guidelines for constructing the taxonomy of DR-NTU was formulated at this stage in preparation for roll out of the draft taxonomy to other stakeholders having input to the taxonomy construction process. This document explains the purpose of the taxonomy, the guidelines for punctuation, maximum number of levels for each category, consistent common terminology, etc.

The following is version 1 of the set of guidelines:

Guidelines for constructing the taxonomy of DR-NTU

1. The main purpose of the taxonomy is to create a set of subject terms or categories to provide a way for users to **browse** the DR-NTU collection by major subject areas. The intention is not to use it for cataloging or indexing items for retrieval at a very specific level.
2. The category term in the taxonomy should not be too specific (which will require more time and effort for librarians to assign) nor too broad (which will make browsing ineffective).
3. The category terms are constructed as a pre-coordinated string, i.e.

Discipline - Level 1 – Level 2 – Level 3, etc.,
e.g.
SCIENCE – Physics - Acoustics
SCIENCE – Physics – Atomic physics
SCIENCE – Physics – Classical mechanics
SCIENCE – Physics - Climatology
SCIENCE – Physics – Condensed matter
SCIENCE – Physics – Condensed matter – Electric & magnetic properties
SCIENCE – Physics – Condensed matter – Electronic structure
SCIENCE – Physics – Condensed matter – Mechanical properties

generally from broad to specific, for example, <Chemistry – Crystallography – Liquid crystals>. The advantages are

- a. It is extensible, i.e. more specific levels can be added over time
 - b. It provides a systematic structure for consistency
 - c. It serves as a good mnemonic aid.
4. Whenever possible, there should not be more than 3 levels in the string after the main discipline category. Generally, 2 levels should be sufficient.
 5. Each level should be easy to browse – i.e. each level should not have more than 20 terms so that the user can easily find a relevant term by running through the list
 6. The string need not be strictly hierarchical. Thus some intermediate terms can be dispensed with to make the term more readable, e.g.

Electronics – applications – Biometric identification	TO	Electronics - Applications
Electronics – applications – Computer hardware		Electronics – Biometric identification
Electronics – applications - Detectors & sensors		Electronics - Computer hardware
		Electronics – Detectors &

	sensors
--	---------

7. There 2 ways in which articles in DR-NTU are categorised using the terms in the taxonomy:

- a. Users select the appropriate term for their articles when they submit it to DR-NTU online.
- b. Library staff assigns the categories to existing collection and also to revise or edit entries by users

8. Asterisk means that users could stop at selecting the subject keyword or key phrase before it, e.g.

Communication and mass media	*
Communication and mass media	Children
Communication and mass media	Communication models

The first line of ‘Communication and mass media’ followed by an asterisk refers to general aspects of ‘Communication and mass media’.

9. How the subject taxonomy displays in the DR-NTU document submission page for users to select the controlled vocabulary for the subject keywords that are to be captured in the metadata records:

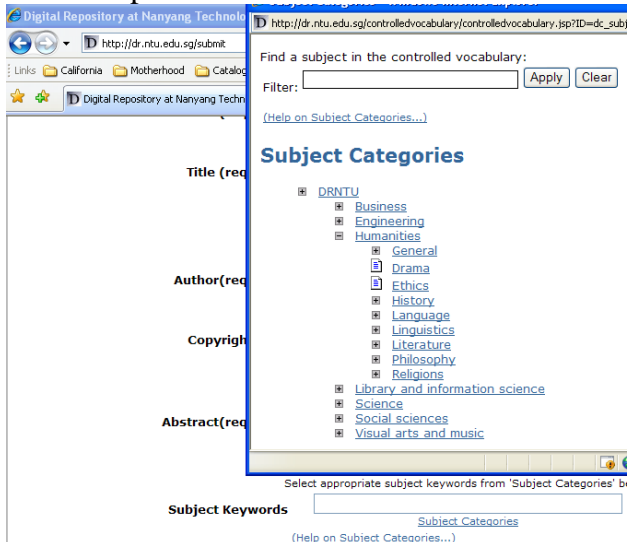


Figure 6. Guidelines for constructing the taxonomy of DR-NTU

Stage 3: Test draft version of DR-NTU taxonomy

Before the next stage of taxonomy review, the draft version of the taxonomy was tested on a representative collection of items. This testing was useful to identify gaps and discrepancies as well as potential usability issues such as the ease for users to find specific information while tagging, searching and navigating. The feedback from this exercise was used to further refine the taxonomy.

Stage 4: Review by subject librarians

Stage 4 is an important stage in the taxonomy development process as it is the first time that the draft version of the taxonomy is made available and accessible to the subject librarians. In the case of DR-NTU, the subject librarians represent a key group of stakeholders; as proxy users they provide valuable input on the information requirements from their respective community subject areas. At this stage of the taxonomy review process, collaborative work is an essential method of working to gain more insightful and user-directed results. Discussions between the taxonomy development team and subject librarians resulted in addition, withdrawal, modification and combination of subject terms; further refinements were made leading to the final structure and content of the taxonomy.

Classification schemes and thesauri were found to be helpful in creating the structure and categories related to the subject facet of the taxonomy, but organizational community sources had to be consulted and several methods had to be employed. Main categories were determined by identifying the stakeholders' interests and consulting organizational community sources and domain taxonomies. (Wang, Chaudhry & Khoo, 2008)

In some subject areas, for example in information technology and computer engineering, the use of industry taxonomy scheme was preferred instead of the structure proposed in the draft taxonomy. In traditional information organization this may raise issues of language inconsistencies across the subject taxonomy, however in the DR-NTU collaborative information model, the users' convenience remains the foremost consideration.

A good taxonomy also is one in which "everything has a place and only one place." Weinstein says. "The sum total of the taxonomy is mutually exclusive of all of the content, and it's collectively exhaustive as well." Also, "the terms used in the taxonomy should be native terms to the user community. They have to be terms that the users will understand instantly, intuitively, and clearly." (Pack, 2002)2)

Stage 5: Implement Version 1 of DR-NTU Taxonomy

The taxonomy 'View and search' feature on the interface is built incorporating the final version of the taxonomy structure and content. Users of DR-NTU are able to view the DR-NTU taxonomy and select subject terms from the DR-NTU taxonomy to be included as subject keywords for their documents which they were going to submit. A work plan has also been created for the subject tagging of the retrospective documents in the DR-NTU to populate the taxonomy. In the final section, some of the issues from this last stage of the taxonomy development will be discussed.

Implementing the DR-NTU taxonomy

The subject taxonomy of DR-NTU serves the purposes of categorising and organising the publications of the NTU community according to subject disciplines. The taxonomy also aids in establishing relationships among the subject categories because of its hierarchical structure and helps in eliminating or reducing ambiguity of the content.

As described in earlier sections of this paper, there are 2 main ways items in DR-NTU are subject tagged:

- a. Users select the appropriate term for their articles when they submit it to DR-NTU online.

- b. Library staff assigns the categories to existing collection and also to revise or edit entries by users

Presently there are not many instances of user assigned subject categories as the DR-NTU was recently launched so this section will be mostly describing issues from the second scenario where categories are assigned by catalogers from the Bibliographic Services Division.

The subject taxonomy was first implemented on NTU's past years' digital collection consisting of theses, dissertations, research reports and student reports. The collection size is more than 10,000 items and comprises submissions from the different schools and research centers in NTU. The collection size is not evenly distributed; for established communities such as the College of Engineering the collection size is, as expected, larger than schools which have been recently set up.

One concern when the implementation started was to create an efficient workflow for the assignment of subject categories. There were two main considerations when prioritizing the starting point for the assignment of subject categories. Agee (Agee, 2008) has pointed out that skillful taxonomy development will require a clear understanding of the terminology of the field. The process of assigning the appropriate level of subject categories from the taxonomy to the collections may be expedited when the subject designator has knowledge of the discipline and terminology used. Another consideration was to tackle the larger sized collections first as this would allow the opportunity to fully populate the taxonomy.

As the subject tagging implementation team comprised two catalogers who were also engineering subject librarians, the first subject areas of assignment were in Civil Engineering and Electrical and Electronic Engineering. Another area of focus was the Business collection which held more than 4,000 items. In both these collection areas, the work started on the more specific subject items i.e. post-graduate theses and dissertations.

The final DR-NTU subject taxonomy was highly customised from the Library of Congress Classification framework so in addition to traditional cataloging tools for subject analysis, it was useful to also create a customised reference guide for Library staff to use in assigning subject categories for DR-NTU. On a broad level, this guide notes the source of the subject category captions for that subject area. For each specific subject category, the reference guide functions similar to scope notes, listing relevant keywords that would aid to classify a specific topic under a broader subject category. It may also list where applicable possible cross-references and LCC classification numbers for cross-mapping.

There are some positive operational aspects of creating a comprehensive reference guide:

- It reduces misinterpretation of the scope of a subject category since subject designators may not always be familiar with the subject domain
- It improves consistency of subject category assignment and downsize the variance where there are multiple subject designators assigning subject categories from a single school
- It facilitates taxonomy maintenance, for example by annotating placeholders for specific topics categorised under a general subject category for future review
- It allows the possibility of testing the DR-NTU collaborative tagging model by using NTU students or non-Library staff to assign subject categories in a controlled environment and hopefully increase the operational efficiency

After catalogers had populated the taxonomy for the initial three subject areas and established their respective reference guides, the use of NTU students or non-Library staff was integrated into the subject tagging operational workflow. The students were assigned to work on the student report collection where subject topics may not be as specific in comparison to the theses collection.

The table below shows the correlation between the catalogers' subject assignment and the students' selection. The students were able to assign more appropriate categories within their area of study; however they are still able to assign suitable subject categories more than sixty percent of the time outside of their knowledge domain. Allowing for minimal on the job training, these statistics indicate positive user response to the DR-NTU taxonomy.

Collection	Business+*		Electrical & Electronic Engineering*		Civil Engineering*		Mechanical Engineering+	
	No.	Percentage	No.	Percentage	No.	Percentage	No.	Percentage
Subject Tag Match	596	63.74%	310	62.00%	327	83.85%	172	77.13%
Subject Tag Not Matched	339	36.26%	190	38.00%	63	16.15%	51	22.87%
Total	935	100.00%	500	100.00%	390	100.00%	223	100.00%

* Subject designator is 2nd year Civil Engineering Student

+ Subject designator is 1st year Mechanical Engineering Student

Figure 7 Correlation between catalogers and non-catalogers subject category assignment using reference guide

Multidisciplinary content

Richardson, S, et al (Richardson, Childs, & Dempster, 2004) mentions that in multidimensional situations a resource might locate itself within more than one subject category or be defined in terms of different purposes or perspectives. The collections in DR-NTU from today's multidisciplinary curriculum have titles which will fit inside more than one subject category in the taxonomy. For such cases, two subject categories are assigned to illustrate the multidisciplinary nature of the content. The title "Coffee over the Internet" can be analysed from the perspective of control engineering. The second subject category assigned, DRNTU::Engineering::Electrical and electronic engineering::Control and instrumentation::Control engineering highlights the contents of the topic from that perspective.

Another such example is an Electrical and Electronic Engineering thesis entitled "SECS/GEM message discovery for rapid equipment integration" which aims to analyze and suggest a solution to alleviate the problem of slow and tedious equipment integration process in the semiconductor manufacturing industries. This can be analyzed from Electronic engineering perspective as well as manufacturing engineering perspective. Hence the following two subject categories are assigned:

DRNTU::Engineering::Electrical and electronic engineering::Semiconductors and
DRNTU::Engineering::Manufacturing.

Subject tag distribution

Once a subject taxonomy area has been populated, the distribution of subject categories used is helpful in analyzing areas for taxonomy maintenance. Subject tags which are in high use may need to be monitored for future sub-categorization to ensure optimum navigational usability. Conversely a long range of low use categories may indicate these areas are sufficiently represented at higher level categories without further sub-categorization. These content analytics may also provide useful data for subject librarians to assess their user profile requirements in future.

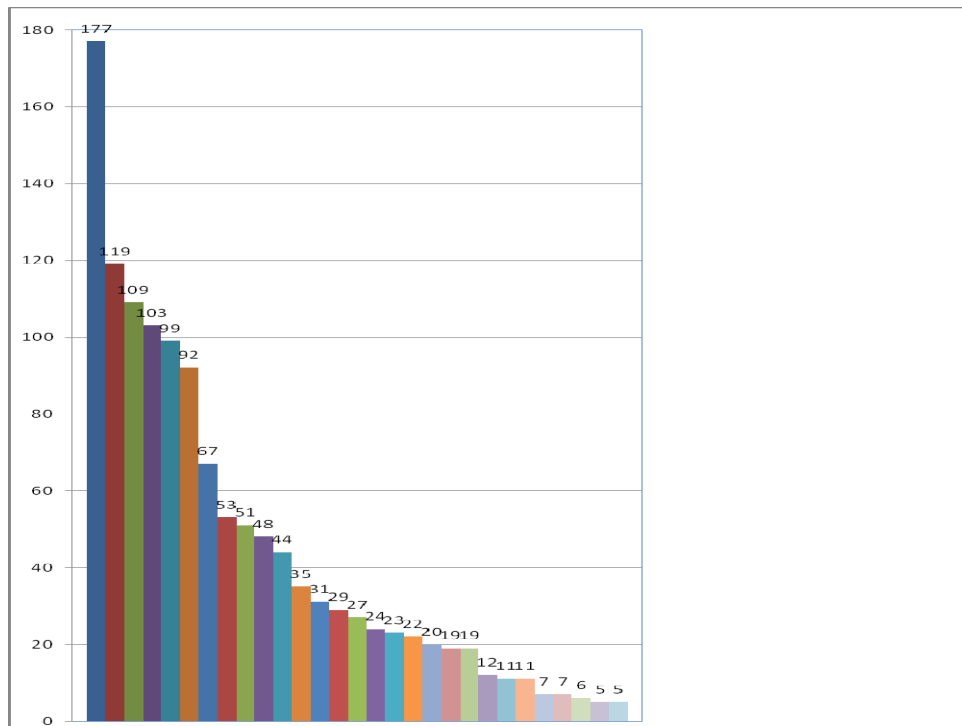


Figure 8. Subject category distribution for Electrical and Electronic Engineering collection

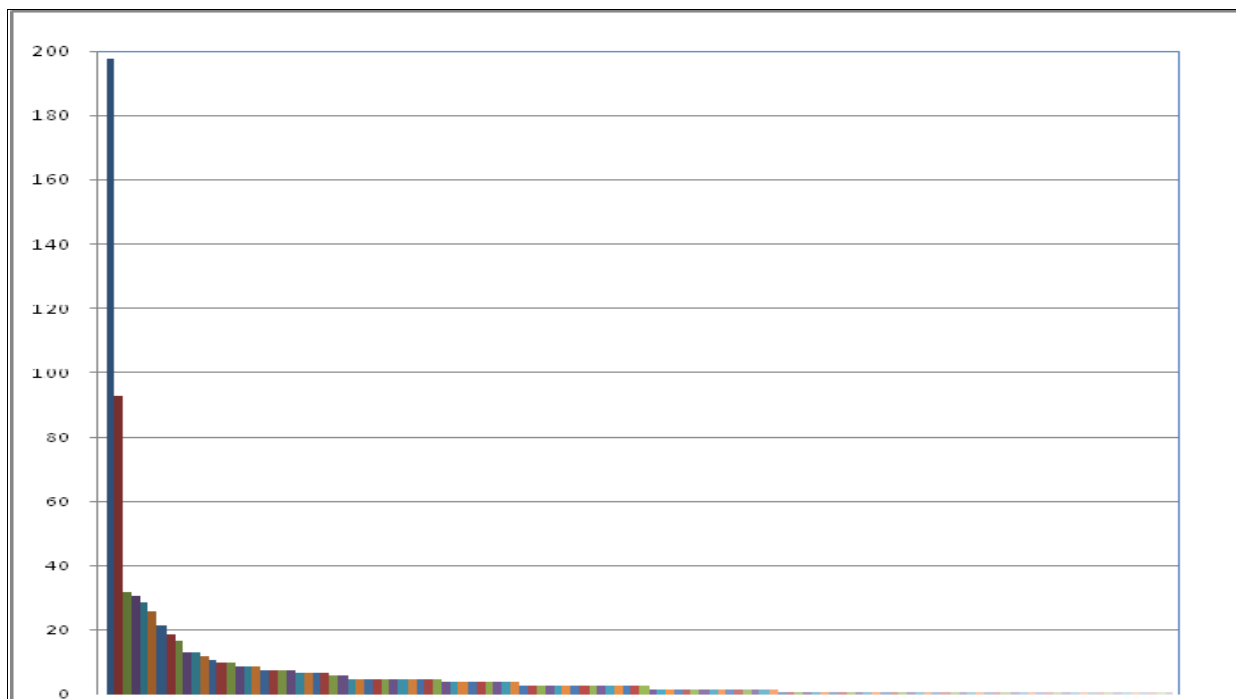


Figure 9. Subject category distribution for Business collection

For the subject taxonomy of DR-NTU, one of the aims is to organise the categories into subsets which are likely to appear more logical and intuitive from the user's viewpoint, as opposed to a librarian's viewpoint. The DR-NTU subject taxonomy is positioned to be a value added feature to web search methods allowing the use of both uncontrolled keywords and controlled vocabulary to facilitate knowledge discovery in NTU's digital collection.

References

- Agee, V. (2008). Controlling Our Own Vocabulary: A Primer for Indexers Working in the World of Taxonomy. *Keywords*, Vol. 16(No.1 January/March), pp. 30-31
- Broughton, V., & Slavic, A. (2007). Building a faceted classification for the humanities: principles and procedures. *Journal of Documentation*, 63(5), 727 - 754.
- Dill, E., & Palmer, K. L. (2005). What's the Big IDeA? Considerations for Implementing an Institutional Repository. *Library Hi Tech News*, 22(6).
- IFLA. (2007, Feb 2009). Functional Requirements of Bibliographic Records : final report. from <http://www.ifla.org/VII/s13/frbr/frbr1.htm#2>
- Lubas, R. L., Wolfe, R. H. W., & Fleischman, M. (2004). Creating metadata practices for MIT's OpenCourseWare project. *Library Hi Tech*, 22(2), 138 - 143.
- Melzer, J. (2003). Enterprise information architecture in context, version 2.0. Retrieved Jan 2009, from http://jamesmelzer.com/EIAinContextv2_final.pdf
- Morrison, J. H. (2003). Implementing taxonomy during development. Understanding information taxonomy helps build better apps. from <http://www.builderau.com.au/program/development/soa/Implementing-taxonomy-during-development/0,339024626,320276137,00.htm>
- Pack, T. (2002). Taxonomy's role in content management. *EContent*, 25 (no.3), pp. 26-31.
- Richardson, S., Childs, M., & Dempster, J. A. (2004). Developing interoperable taxonomy systems for sharing resources within multidisciplinary communities of practice. *Interactive Technology & Smart Education*, Vol.1(No. 2, May), pp. 91-100.
- Rosenfeld, L., & Morville, P. (2002). *Information architecture for the world wide web* (2nd ed. ed.).
- Svenonius, E. (2001). *The Intellectual foundation of information organization*. Cambridge, MA: The MIT Press.
- Tillet, B. (2004). What is FRBR? : a conceptual model of the bibliographic universe [Electronic Version]. Retrieved Feb 2009, from <http://www.loc.gov/cds/downloads/FRBR.PDF>
- Wang, Z., Chaudhry, A. S., & Khoo, C. S. G. (2008). Using classification schemes and thesauri to build an organizational taxonomy for organizing content and aiding navigation. *Journal of Documentation*, 64(6), pp. 842-876.