Using taxonomies for knowledge exploration in subject gateways

Marcin Roszkowski
National Library of Poland
Institute of Information Science and Book Studies, University of Warsaw, Poland
m.roszkowski@bn.org.pl

INFORUM 2011: 17th Conference on Professional Information Resources
Prague, May 24-26, 2011

Abstract

The paper presents results of the study conducted on the use of taxonomies in subject gateways. The goal of research was the analysis of methods and tools for information organization. The paper presents static and dynamic models of taxonomies as means for knowledge exploration and access to information. This includes qualitative and quantitative analysis of structure and usage of traditional indexing languages as a source for vocabularies and taxonomies framework.

Introduction

At the beginning of the new millennium F.W. Lancaster, commenting the future of indexing and abstracting services, said: "it seems clear that the continued growth of network-accessible information resources will make subject analysis activities of greater importance than ever before". (Lancaster, 2003, p. 143) Subject gateways are the best example of the potential of manual approach to cataloging the networked resources. The aim of subject gateways is the selection of high quality web resources from heterogeneous and distributed web environment and to make available this information to the users. Subject gateways are services that provide access to Internet resources that have been reviewed, selected and described by subject specialists. The exact selection criteria largely depend on the perceived usage of the gateway, but typically include factors relating to the content and presentation of the resource, the integrity of the information and site provider. Subject gateways are almost always based on the manual creation of descriptive metadata and usually provide end users both search and subject-browse facilities. The existence of rich metadata means that gateways can offer more sophisticated search options than other Web indexes. The application of subject classification schemes means that gateway services often provide hierarchical browse structures for browsing (Koch, 2000). As the Internet itself is constantly evolving, subject gateways also need robust collection development policies that include the regular checking and updating of resources included in the database. (Day, Koch, & Neuroth, 2004)

Subject gateways serve as reliable references in web environment for over decade. But not all of them have stood the test of time. Most of subject gateways were set up within the projects funded by institutions, often associated with the higher education sector. (see Dempsey & Law, 2000) Ending the project was also termination of funding and closure for many subject gateways. Some of the well established subject gateways are:

- Intute (http://intute.ac.uk) – established in 2006 r. as a merge of eight different subject gateways. (Joyce, Kerr, Machin, & Williams, 2010). The scope of Intute covers many fields of knowledge (social sciences, arts and humanities, life sciences, etc.).

- Infomine: Scholarly Internet Resource Collections (http://infomine.ucr.edu) – developed and supported by the Library of the University of California. Infomine presents wide range of multi domain collection.
- The Gateway to 21st Century Skills (http:// www.thegateway.org*)* – educational resources for USA users.
- Special Subject Guides / SSG-Fachinformation (SSG-FI) (http://www.sub.uni-goettingen.de/ssgfi/) – German network for subject gateways:
    - MathGuide (http://www.mathguide.de/) - mathematics
    - GeoGuide (http://www.geo-guide.de) – geology
    - Anglo-American Culture (http://www.sub.uni-goettingen.de/ssgfi/anglo-americana.html) - Anglo-American Literature and history.
    - ForestryGuide (http://www.forestryguide.de) – forestry.
- BUBL LINK Catalogue of Internet Resources (http://bubl.ac.uk) – one of the oldest multi domain systems. Since April 2011 the BUBL is no longer being updated.
- And many more see (MacLeod & Bond, 2011).

The important fact about subject gateways is that on the system output user gets the metainformation. This means that user is informed about relevant, high quality web resources from specific domain.

Concept of information organization (Svenonius, 2000; Taylor, 1999) in subject gateways covers the methods and tools for metadata creation and expression (metadata schemes, coding schemes, controlled vocabularies, etc.) and tools for structural access to collections. Sacco & Tzitzikas (Sacco & Tzitzikas, 2009) identified two different information access modes: focalized search and exploratory search. In focalized search, the user attempts to quickly locate relevant information items on the basis of their contents. In exploratory search (also called browsing) the user explores relationships among items in a database. Exploratory search in subject gateways is conducted by taxonomies.

Graef argues that taxonomies "are structures that provide a way of classifying things--living organisms, products, books--into a series of hierarchical groups to make them easier to identify, study, or locate. Taxonomies consist of two parts--structures and applications. Structures consist of the categories (or terms) themselves and the relationships that link them together. Applications are the navigation tools available to help users find information." (Graef, 2001) In the broad sense, taxonomies are: classification scheme, semantic and knowledge map. (Lambe, 2007) Taxonomy as a classification scheme means use of hierarchical structures as the basis for its structure. The *semantic* element embodies the idea that the terms in the taxonomy mean something to the community for which the taxonomy has been developed. And those terms have some relation to one another, as described in the taxonomy. This leads to the *knowledge map* aspect: the taxonomy can be looked upon as a way to describe how the community thinks about the content. (Vinson, 2007)

The function of taxonomy is to provide:

- Identification--The taxonomy can help control the glut of information and identify where information should be stored by filtering, categorizing, and labeling information.
- Discovery--Additional information on a topic can be inferred by seeing where the entry is placed in context within the taxonomy and provide serendipitous guidance to the person working on the issue.
- Delivery--The taxonomy can improve the retrieval process. The use of the taxonomy's controlled vocabulary enhances searching via browsing. The use of navigation paths or "breadcrumbs" based on the taxonomy's hierarchy provide context and enhance searching via free text. (Bruno & Richmond, 2003)

Two key components of taxonomies are hierarchical structure and labels. These two elements are also core features for classification and thesauri, respectively. In other words, hierarchical structure is the foundation of classification schemes; while labels naming concepts and represented by terms are building blocks of thesauri. Thus, some researchers state that taxonomies can be considered as a combination of features of classification and thesauri. (Zhonghong, Chaudhry, & Khoo, 2006) This statement means that taxonomies have common features of these tools for knowledge organization but they cannot be seen as a combination of classification and thesauri.

(Zhonghong et al., 2006) compiled a table showing the differences between taxonomies and classification schemes and thesauri. (Table 1.)

Features		Classification schemes	Thesauri	Taxonomies
Scope		Library community Academic disciplines	Online environment Academic community	Web environment Organizational environment
Treated objects		Collections	Documents	Digital resources
Roles		Classifying Shelving	Indexing Searching	Categorizing Browsing and navigation
Forms	Hierarchical Structure	One-dimensional Use combination of Notations	Networked term relationships	Dynamic structure
	Terms	Classes	Terms	Categories
Focus		More on content	More on content	More on users

Table 1. Differences between taxonomies and classification schemes and thesauri. Source: (Zhonghong et al., 2006)

Lambe (Lambe, 2007) argues that taxonomies can take many forms. These are:

- Lists,
- Trees,
- Hierarchies,
- Polyhierarchies,
- Matrices,
- Facets,
- System maps.

They can be represented as anything from a flat (nonhierarchical) structures to a mono- and polihierarchical ones including faceted approach. The final structure of a specific taxonomy depends on what best fits for the community and the content being taxonomized. (Vinson, 2007)

The aim of this study is the analysis of taxonomies in subject gateways and attempt to investigate the taxonomy models. Identification of the dependencies and patterns in subject gateway's taxonomies is the clue for ways of information exploration in such environments.

Methodology

The survey was conducted at the beginning of 2009 on a set of 20 subject gateways. Research material covered mainly large systems with English language interface. Taxonomies were extracted from user interface and moved to spreadsheet software. Taxonomies were analyzed quantitatively and qualitatively. Quantitative research tended to answer questions about:

- depth of the hierarchies,
- Categories per level distribution,
- Resources per level distribution,
- Number of categories on the first level of division.

Qualitative research tended to answer questions about ways of taxonomy exploration, principles of division, taxonomy structure.

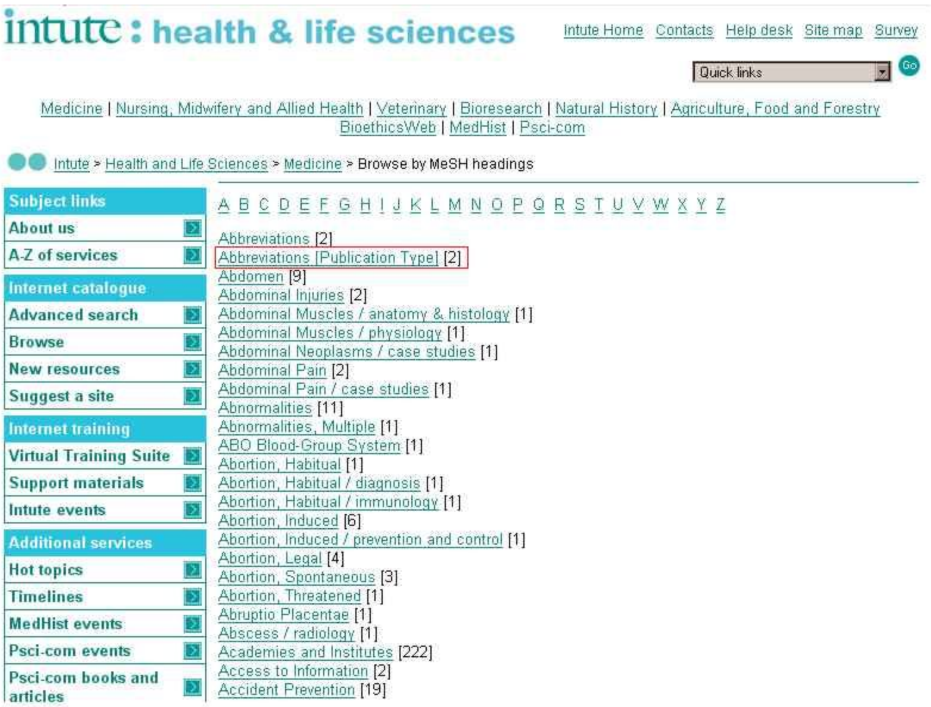
Methodology of this study was adopted and modified from works of Kuyng-Sun et al. (Kuyng-Sun, Sei-Ching, & Soo-Jin, 2006), Vizine-Goetz (Vizine-Goetz, 2002), Wheatley (Wheatley, 2000) and Zins (Zins, 2002).

Results

This study resulted in the development of the three models of models of taxonomies in subject gateways. They were identified on the basis of hierarchical relationships application. These are: flat (non-hierarchical) and hierarchical model. The last one includes the static and dynamic type. The concept of category in this study is used in its narrow sense, as a basic unit of taxonomy, not in a sense of theory of indexing languages.

Flat model

Taxonomies representing this model are built of subject categories in alphabetic order without any semantic relationships between them. This includes hierarchical and associative associations. The content of flat taxonomy is a set of categories representing both subjects and form of resources, often in one listing. Form oriented categories often have some kind of apposition pointing the fact that they represent non-topical point of view. Mostly it takes form of expression like: "resource type" (Fig. 1.)

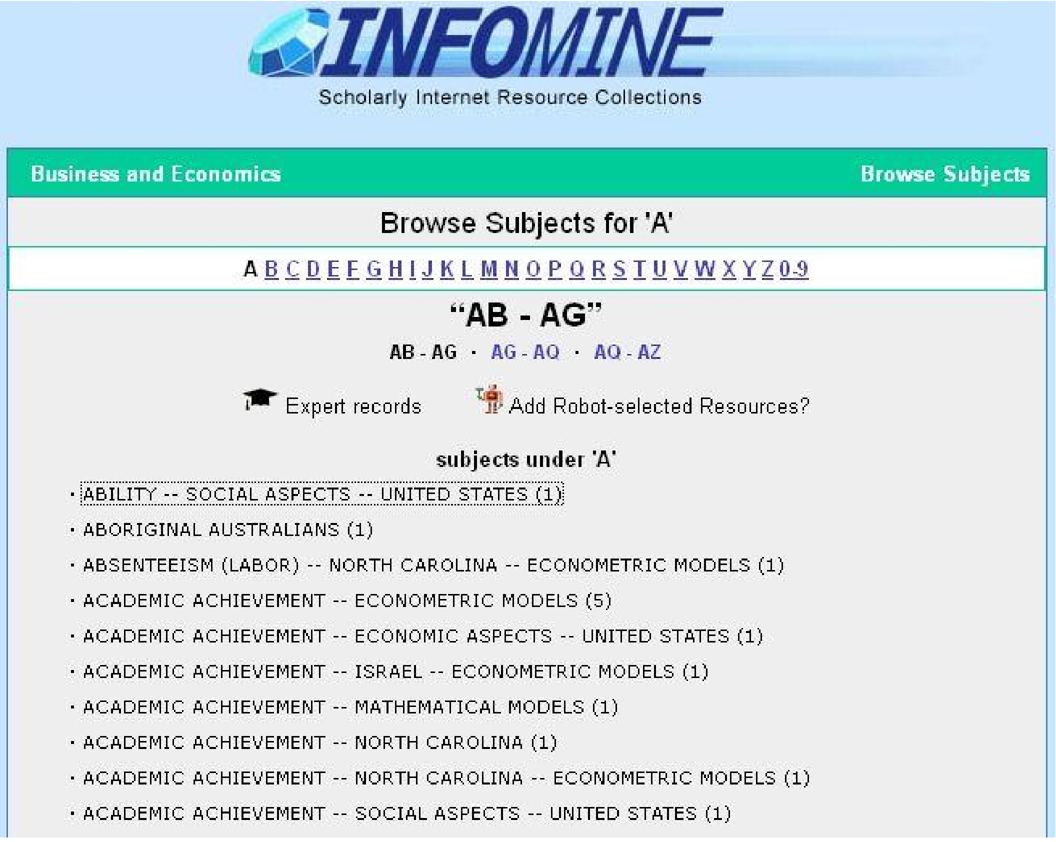

intute : health & life sciences
Intute Home
Contacts
Help desk
Site map
Survey
Quick links
Go
Medicine | Nursing, Midwifery and Allied Health | Veterinary | Bioresearch | Natural History | Agriculture, Food and Forestry
BioethicsWeb | MedHist | Psci-com
Intute > Health and Life Sciences > Medicine > Browse by MeSH headings
Subject links
About us
A-Z of services
Internet catalogue
Advanced search
Browse
New resources
Suggest a site
Internet training
Virtual Training Suite
Support materials
Intute events
Additional services
Hot topics
Timelines
MedHist events
Psci-com events
Psci-com books and articles
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Abbreviations [2]
Abbreviations [Publication Type] [2]
Abdomen [9]
Abdominal Injuries [2]
Abdominal Muscles / anatomy & histology [1]
Abdominal Muscles / physiology [1]
Abdominal Neoplasms / case studies [1]
Abdominal Pain [2]
Abdominal Pain / case studies [1]
Abnormalities [11]
Abnormalities, Multiple [1]
ABO Blood-Group System [1]
Abortion, Habitual [1]
Abortion, Habitual / diagnosis [1]
Abortion, Habitual / immunology [1]
Abortion, Induced [6]
Abortion, Induced / prevention and control [1]
Abortion, Legal [4]
Abortion, Spontaneous [3]
Abortion, Threatened [1]
Abruptio Placentae [1]
Abscess / radiology [1]
Academies and Institutes [222]
Access to Information [2]
Accident Prevention [19]

Fig. 1. Categories and appositions in the Intute: health & life sciences flat taxonomy

Flat taxonomies are representation of topics distribution, expressed during the indexing stage. These taxonomies can be easily generated from specific field of metadata scheme. Information exploration through this taxonomy is based only on identification of relevant category in alphabetic order and analysis of assigned resources.

An interesting example of flat taxonomy is navigation tool in Infomine (http://infomine.ucr.edu) gateway. This flat browsing structure is built on compounded Library of Congress Subject Headings. (Fig. 2.) So the name of category consists of heading and proper number of subheadings designating the scope of the category.



INFOMINE

Scholarly Internet Resource Collections

Business and Economics

Browse Subjects

Browse Subjects for 'A'

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z 0-9

“AB - AG”

AB - AG · AG - AQ · AQ - AZ

Expert records

Add Robot-selected Resources?

subjects under 'A'

- ABILITY -- SOCIAL ASPECTS -- UNITED STATES (1)
- ABORIGINAL AUSTRALIANS (1)
- ABSENTEEISM (LABOR) -- NORTH CAROLINA -- ECONOMETRIC MODELS (1)
- ACADEMIC ACHIEVEMENT -- ECONOMETRIC MODELS (5)
- ACADEMIC ACHIEVEMENT -- ECONOMIC ASPECTS -- UNITED STATES (1)
- ACADEMIC ACHIEVEMENT -- ISRAEL -- ECONOMETRIC MODELS (1)
- ACADEMIC ACHIEVEMENT -- MATHEMATICAL MODELS (1)
- ACADEMIC ACHIEVEMENT -- NORTH CAROLINA (1)
- ACADEMIC ACHIEVEMENT -- NORTH CAROLINA -- ECONOMETRIC MODELS (1)
- ACADEMIC ACHIEVEMENT -- SOCIAL ASPECTS -- UNITED STATES (1)

Fig. 2. Compound subject headings as names for categories in Infomine gateway.

In flat taxonomies there are no means for any relationships establishment and therefore no other ways of information searching than alphabetical listing exploration.

Hierarchical models

This type of taxonomy was distinguished on the basis of the scope of the hierarchical relationships as a basis of taxonomy structure. During the analysis of research material static and dynamic types have been discovered.

Static model

Static model of hierarchical taxonomies in subject gateways involves the use of broader/narrower semantic relationship for organization of categories. In these subject trees no logical division takes place.

The generalized conclusions from quantitative research are:

- The depth of taxonomies varies from two to seven levels. Most of analyzed taxonomies presented 3-5 levels depth. (Fig. 3.)
- Categories per level distribution showed that great number of categories (almost 70 %) was placed on second and third level. (Fig. 4.)
- Resource per level distribution showed that great number of postings was assigned to categories on second and third level. (Fig. 5.)
- There is a strong correlation between resource and category per level distribution. (Fig. 6.)
- The average number of categories on first level is 14 but it varies from 3 to 40. (Fig. 7.)
- One resource is assigned to average two categories (access point factor = 2,07).

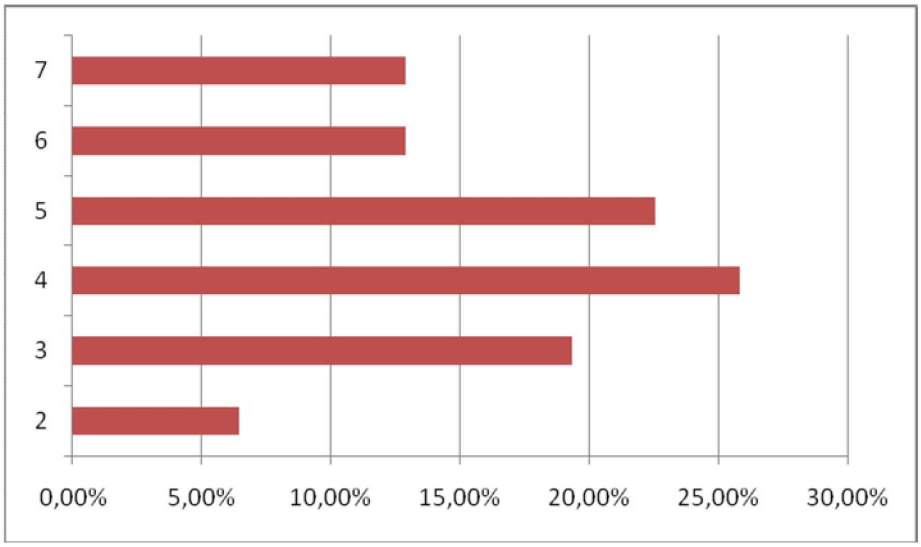


Fig. 3. Depth of the hierarchies.

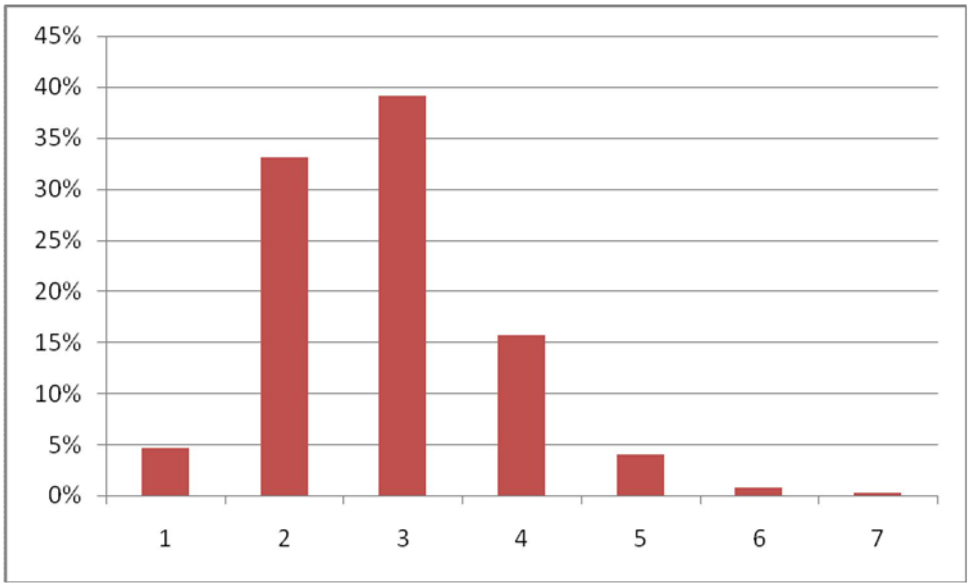


Fig. 4. Categories per level distribution.

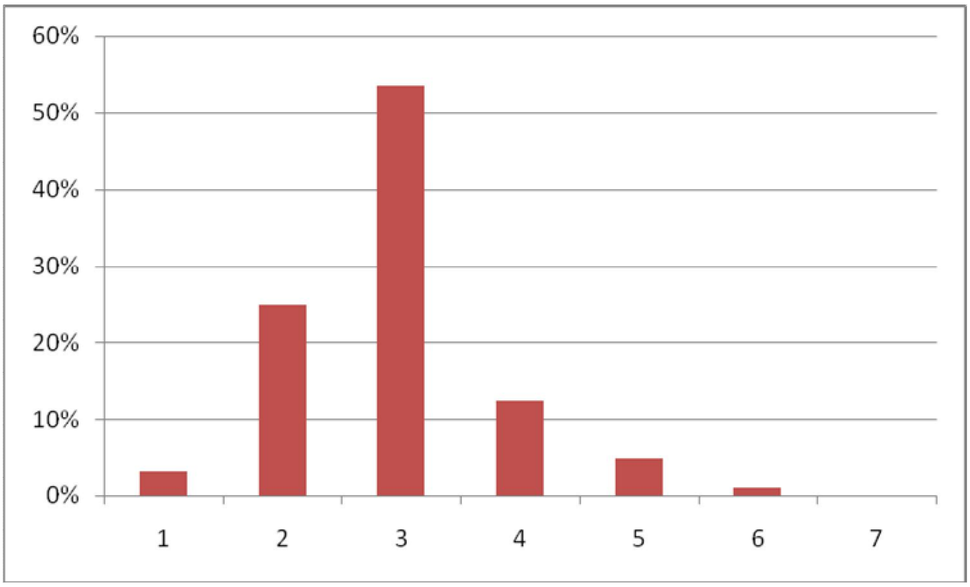


Fig. 5. Resources per level distribution.

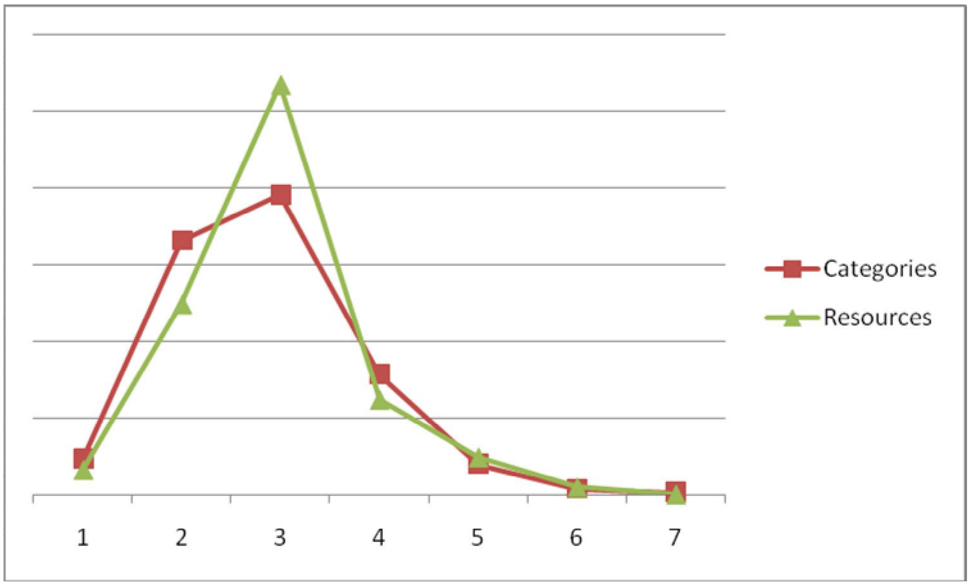


Fig. 6. Correlation between resource and category per level distribution.

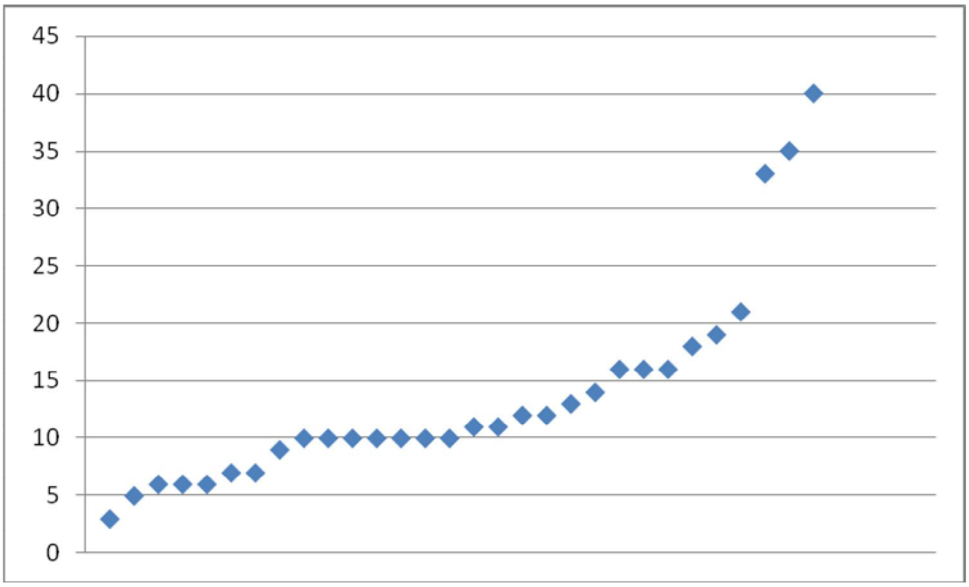
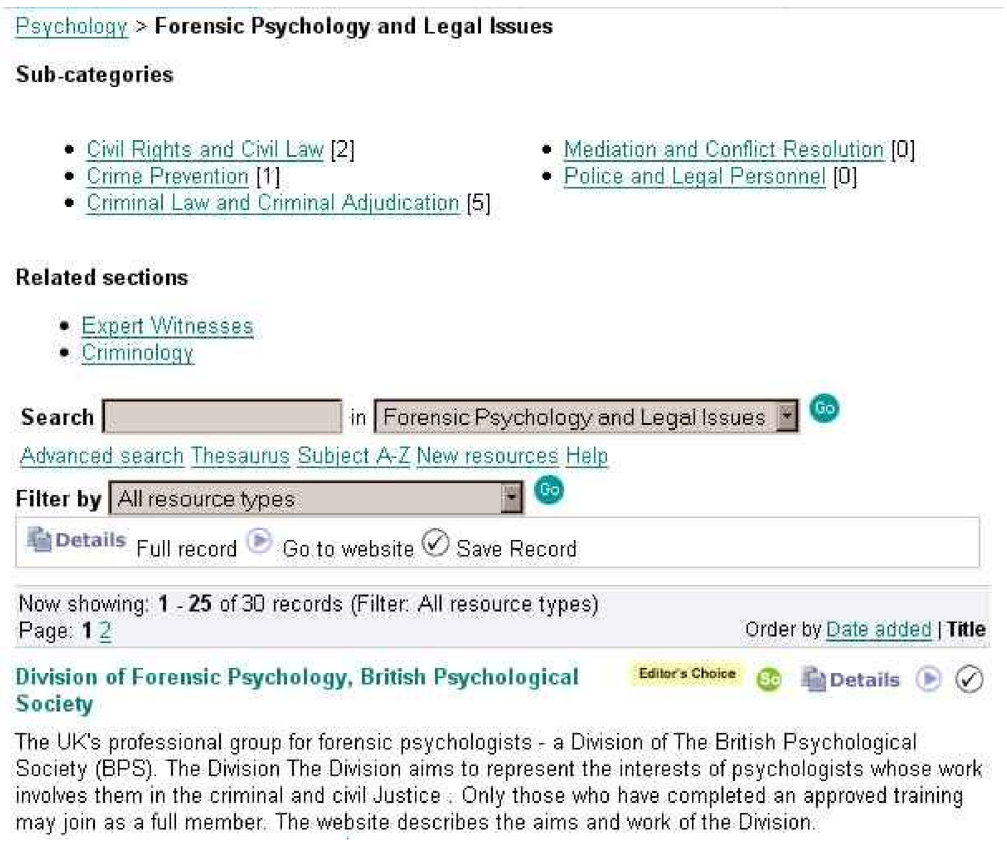


Fig. 7. First Cut - number of categories on the first level of division

Construction of categories in static hierarchical taxonomies is based on principles of literary warrant or “dummy categories”. Literary warrant “first introduced by Hulme as a means of class determination, the principle prescribes that controlled vocabulary usage be empirically derived from literature containing the vocabulary to be controlled.” (Svenonius, 2003) Applying the literary warrant principle as basis for category identification, there should not be any empty category (without posting). However this rule is not applied consistently as evidenced by the empty categories in many analyzed taxonomies. Another approach to category identification is using dummy categories. They are used as empty intermediaries in specific branch expansion. Dummy categories allow for creation of consistent structure. Dummy categories also show to the user that specific topic/concept is included in the scope of taxonomy but currently there is no posting to collection.



Psychology > **Forensic Psychology and Legal Issues**

Sub-categories

- Civil Rights and Civil Law [2]
- Crime Prevention [1]
- Criminal Law and Criminal Adjudication [5]
- Mediation and Conflict Resolution [0]
- Police and Legal Personnel [0]

Related sections

- Expert Witnesses
- Criminology

Search in Forensic Psychology and Legal Issues

Advanced search Thesaurus Subject A-Z New resources Help

Filter by All resource types

Details Full record Go to website Save Record

Now showing: **1 - 25** of 30 records (Filter: All resource types)
Page: **1** 2 Order by Date added | **Title**

Division of Forensic Psychology, British Psychological Society

The UK's professional group for forensic psychologists - a Division of The British Psychological Society (BPS). The Division The Division aims to represent the interests of psychologists whose work involves them in the criminal and civil Justice . Only those who have completed an approved training may join as a full member. The website describes the aims and work of the Division.

Fig. 8. Dummy categories in taxonomy

Applying appropriate first principle of division is important for whole taxonomy and its abilities for effective information exploration. Kwasnik calls this "the first cut" and argues that it is mostly important because this determines the shape and eventually the representational eloquence of the classification/ hierarchical structure/. If the first cut is a trivial one, the rest of the tree becomes awkward and does not reflect knowledge very well. (Kwasnik, 1999) The average number of categories on first level was 14 but this value ranges from 3 to 40. Only a few taxonomies had high number of categories on first level. The "first cut" in taxonomies is generally a result of topical/subject perspective, so these taxonomies present content-oriented approach. Commonly used approach, is however applying many principles of division on first level. Next to the topical/subject categories there is resource type and user oriented categories.

Information exploration in static hierarchical taxonomies includes browsing the content of taxonomy, analysis of postings, further exploration hierarchical relationships until relevant object is founded. There is a rule 1-many, which means that one resource (metadata record) can be assigned to many categories inside the taxonomy. The static nature of this tool is shown by weak response to the user exploration. The user must take many steps before he finds relevant category and therefore relevant postings.

Dynamic model

The dynamic model of taxonomies in subject gateways involves faceted approach (see Gnoli, 2008; Mills, 2004) to knowledge organization. The dynamic approach to taxonomies construction in subject gateways includes analytic-synthetic methods for concept categorization and categories organization. The best example of this approach is the *Gateway to 21st Skills* taxonomy (http://thegateway.org). (Qin & Paling, 2001) There is a six dimensional concept categorization of concepts represented by categories. The facets constructed in this way are:

- Subject
- Resource Type
- Educational Level
- Keywords
- Mediator
- Beneficiary
- Price Code

These facets correspond to different views on organized set of concepts. Each facet contains categories representing one specific dimension. The dynamic nature of this taxonomy is based on automatic taxonomy reorganization every time user explores one of the categories. This means that when he picks any category from the set, the taxonomy management system presents him additional categories, but only those which are mutually related by semantic relationships or on the basis of co--occurrence in record description. Thus, user can dynamically reformulate query by picking several categories from different facets.

The dynamic taxonomies in subject gateways present different ways of hypertext usage as a tool for information visualization and exploration. Another approach presents CISMef (www.chu-rouen.fr/cismef/), French medicine related gateway. In this case the taxonomy is built from facets, here called "metaterms". (Neveol, Soulamia, & Douyere, 2004) They act as clusters grouping terms representing concepts related to specific dimension or point of view.

Information exploration with dynamic taxonomies gives many opportunities for categories organization and easy query modification by faceted browsing. From simple faceted approach to conceptually sophisticated schemes (see Sacco & Tzitzikas, 2009), analytic-synthetic model of taxonomy makes exploratory search more effective. Faceted model provides the possibility of "knowledge organization on demand" respectively on user's needs and taxonomy interaction.

Discussion

The results of study show different models of taxonomies in subject gateways. Along with increasing the complexity of relationships between categories in taxonomies, also the ability of effective exploratory mode of access to information increases. A characteristic feature of taxonomies in subject gateways is the usage of controlled vocabularies. Controlled vocabularies are used for naming categories and for semantic relationships adaptation. The naming issue covers usage of well established universal and domain specific lexical resources (for example LCSH, MeSH, Arts and Architecture Thesaurus). An evident issue in taxonomy construction in subject gateways is the adaptation of library classifications. In this case only specific sections, parts of classifications are accommodated in taxonomies. Bruno and Richmond argue that good taxonomies, based on the use of classification and controlled vocabularies, result in more efficient information retrieval. This ensures better productivity and less user frustration. (Bruno & Richmond, 2003).

It needs to be stressed that importance of faceted approach for information retrieval effectiveness was stated over 60 years ago by Classification Research Group. (Classification Research Group, 1955) Today, faceted approach to knowledge organization in networked environment takes many forms. One of them is taxonomy, especially dynamic model. Broughton argues that "facet analysis is significant for the clarity of the light it shines upon the relationships between objects and entities, and abstract concepts and their associated labels. It gives a rational, scientific, methodology for the construction of systems; it enables the full and precise description of objects of considerable structural complexity and of multi-dimensional semantic composition; it provides a flexible syntactical apparatus for the combination and ordering of concepts where this is required."(Broughton, 2006) In the case of taxonomies this approach is seeking the consensus between objective criteria (content), users' needs and context of application. However, dynamic taxonomies seem to be a good if not the best tool for multidimensional information exploration.

References

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. *ASLIB Proceedings: New Information Perspectives*, *58*(1/2), 49-72. Retrieved from http://www.fims.uwo.ca/people/faculty/Frohmann/LIS677/Documents/Subject Analysis/Need for a faceted classification 2006.pdf.

Bruno, D., & Richmond, H. (2003). The Truth About Taxonomies. *Information Management Journal*, *37*(2).

Classification Research Group. (1955). The need for a faceted classification as the basis of all methods of information retrieval. *Library Association Record,*, *57*(7), 262-268.

Day, M., Koch, T., & Neuroth, H. (2004). Searching and browsing multiple subject gateways in the Renardus service. RC33 Sixth International Conference on Social Science Methodology, Amsterdam, Netherlands, 16-20 August 2004.

Dempsey, L., & Law, D. (2000). A policy context - eLib and the emergence of the subject gateways. *Ariadne*, (25).

Gnoli, C. (2008). Facets: A Fruitful Notion in Many Domains. *Axiomathes*, (January), 127-130. doi: 10.1007/s10516-008-9032-5.

Graef, J. (2001). Managing taxonomies strategically. *Montague Institute Review.*, (March 30).

Joyce, A., Kerr, L., Machin, T., & Williams, C. (2010). Intute Reflections at the End of an Era. *Ariadne*, (64). Retrieved from http://www.ariadne.ac.uk/issue64/joyce-et-al/.

Koch, T. (2000). Quality-controlled subject gateways: definitions, typologies, empirical overview. *Online Information Review*, *24*(1), 24-34.

Kuyng-Sun, K., Sei-Ching, J. S., & Soo-Jin, P. (2006). Facet Analyses of Categories Used in Web Ditrectories: A Coparative Study BT - Libraries: Dynamic Engines for the Knowledge and Information Society: Proceedings of World Library and Information Congress: 72nd IFLA General Conference and Council. Seul: IFLA.

Kwasnik, B. (1999). The role of classification in knowledge representation and discovery. *Library Trends*, *48*(1), 22-47.

Lambe, P. (2007). *Organising Knowledge : Taxonomies, Knowledge and Effectiveness*. Facet Publishing.

Lancaster, F. W. (2003). Do Indexing and Abstracting Have a Future? *Annales de Documentacion*, (6), 137-144. Servicio de Publicaciones, Universidad de Murcia (Spain) Commentary on: Commentary on UNSPECIFIED Alternative Locations: http://www.um.es/fccd/anales/.

MacLeod, R., & Bond, D. (2011). Pinakes : a subject launchpad. Retrieved April 22, 2011, from http://www.hw.ac.uk/libwww/irn/pinakes/pinakes.html.

Mills, J. (2004). Faceted classification and logical division in information retrieval. *Library Trends*, (Winter).

Neveol, A., Soulamia, L., & Douyere, M. (2004). Using CISMef MeSH "encapsulated terminology and a categorization algorithm for health resources. *International Journal of Medical Informatics*, (73), 57-64.

Qin, J., & Paling, S. (2001). Converting a controlled vocabulary into an ontology: the case of GEM. *Information Research*, *6*(2).

Sacco, G., & Tzitzikas, Y. (2009). Dynamic Taxonomies and Faceted Search. Springer.

Svenonius, E. (2000). *The Intellectual foundation of information organization*. Cambridge: MIT Press Massachusets Institute of Technology.

Svenonius, E. (2003). Design of Controlled Vocabularies. *Encyclopedia of Library and Information Science* (pp. 822-838). New York: Marcel Dekker. doi: 10.1081/E-ELIS.

Taylor, A. G. (1999). *The Organization of Information* (p. 280). Engelwood, Colorado: Libraries Unlimited, Inc.

Vinson, J. (2007). Organising Knowledge -- Taxonomies, Knowledge and Organisational Effectiveness. Retrieved March 21, 2011, from http://blog.jackvinson.com/archives/2007/07/25/organising_knowledge_taxonomies_knowledge_and_organisational_effectiveness.html.

Vizine-Goetz, D. (2002). Classification schemes for internet resources revisited. *Journal of Internet Cataloging*, *5*(4), 5-18.

Wheatley, A. (2000). Subject Trees on the Internet: A New Role for Bibliographic Classification? *Journal Of Internet Cataloging*, *2*(3/4), 115-141.

Zhonghong, W., Chaudhry, A. S., & Khoo, C. (2006). Potential and Prospects of Taxonomies for Content Organization. *Knowledge Organization*, *33*(3), 160-169.

Zins, C. (2002). Models for Classifying Internet Resources. *Knowledge Organization*, *29*(1), 20-28.