USEFULNESS OF SOCIAL TAGGING IN ORGANIZING AND PROVIDING ACCESS TO THE WEB: AN ANALYSIS OF INDEXING CONSISTENCY AND QUALITY

BY

YUNSEON CHOI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Linda C. Smith, Chair and Director of Research
Professor Allen H. Renear
Assistant Professor Miles J. Efron
Professor John M. Unsworth

ABSTRACT

This dissertation research points out major challenging problems with current Knowledge Organization (KO) systems, such as subject gateways or web directories: (1) the current systems use traditional knowledge organization systems based on *controlled* vocabulary which is not very well suited to web resources, and (2) information is organized by *professionals* not by users, which means it does not reflect intuitively and instantaneously expressed users' current needs. In order to explore users' needs, I examined social tags which are *user*-generated *uncontrolled* vocabulary. As investment in professionally-developed subject gateways and web directories diminishes (support for both BUBL and Intute, examined in this study, is being discontinued), understanding characteristics of social tagging becomes even more critical.

Several researchers have discussed social tagging behavior and its usefulness for classification or retrieval; however, further research is needed to qualitatively and quantitatively investigate social tagging in order to verify its quality and benefit. This research particularly examined the indexing consistency of social tagging in comparison to professional indexing to examine the quality and efficacy of tagging. The data analysis was divided into three phases: analysis of indexing consistency, analysis of tagging effectiveness, and analysis of tag attributes. Most indexing consistency studies have been conducted with a small number of professional indexers, and they tended to exclude users. Furthermore, the studies mainly have focused on physical library collections. This dissertation research bridged these gaps by (1) extending the scope of resources to various web documents indexed by users and (2) employing the Information Retrieval (IR) Vector Space Model (VSM) - based indexing consistency method since it is suitable for dealing with a large number of indexers. As a second phase, an analysis of tagging effectiveness with tagging exhaustivity and tag specificity was conducted to ameliorate the drawbacks of consistency analysis based on only the quantitative measures of vocabulary matching. Finally, to investigate tagging pattern and behaviors, a content analysis on tag attributes was conducted based on the FRBR model.

The findings revealed that there was greater consistency over all subjects among taggers compared to that for two groups of professionals. The analysis of tagging exhaustivity and tag specificity in relation to tagging effectiveness was conducted to ameliorate difficulties associated with limitations in the analysis of indexing consistency based on only the quantitative measures of vocabulary matching. Examination of exhaustivity and specificity of social tags provided insights into particular characteristics of tagging behavior and its variation across subjects. To further investigate the quality of tags, a Latent Semantic Analysis (LSA) was conducted to determine to what extent tags are conceptually related to professionals’ keywords and it was found that tags of higher specificity tended to have a higher semantic relatedness to professionals’ keywords. This leads to the conclusion that the term’s power as a differentiator is related to its semantic relatedness to documents. The findings on tag attributes identified the important bibliographic attributes of tags beyond describing subjects or topics of a document. The findings also showed that tags have essential attributes matching those defined in FRBR. Furthermore, in terms of specific subject areas, the findings originally identified that taggers exhibited different tagging behaviors representing distinctive features and tendencies on web documents characterizing digital heterogeneous media resources. These results have led to the conclusion that there should be an increased awareness of diverse user needs by subject in order to improve metadata in practical applications.

This dissertation research is the first necessary step to utilize social tagging in digital information organization by verifying the quality and efficacy of social tagging. This dissertation research combined both quantitative (statistics) and qualitative (content analysis using FRBR) approaches to vocabulary analysis of tags which provided a more complete examination of the quality of tags. Through the detailed analysis of tag properties undertaken in this dissertation, we have a clearer understanding of the extent to which social tagging can be used to replace (and in some cases to improve upon) professional indexing.

To Mom and Dad

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my dissertation advisor and committee chair, Dr. Linda C. Smith for all of her extraordinary support and guidance from the planning and development of this dissertation research. Without her great intelligence and excellent guidance, I would never have been able to start and complete my dissertation. She also provided me with unwavering encouragement and support in various ways.

I would also like to sincerely thank Dr. Allen H. Renear for his mentorship and insightful guidance not only during my work on the dissertation but throughout my graduate studies. I am deeply grateful to Dr. Miles J. Efron for his invaluable and constructive feedback on my dissertation. The progress of this dissertation could not be accomplished smoothly without his tremendous help. I am truly grateful to Dr. John M. Unsworth for his enthusiastic encouragement and useful critiques of this dissertation.

I would like to thank faculty members in the Graduate School of Library and Information Science. I am sincerely grateful to Dr. David Dubin for his sincere support throughout my doctoral program. I would also like to thank Dr. Kathryn La Barre, who gave me useful comments and guidance in the early stages of this dissertation.

I extend my thanks to my colleagues and friends who provided support and encouragement to me. Especially, I would like to thank Soohyung Joo for participation in intercoder reliability tests. I would never have been able to finish my dissertation without love and support from my family. My special appreciation goes to my husband Byunggil Yoo, who was always willing to help and offered useful suggestions. His technical support for collecting data and handling the software programs has been very helpful. I would like to thank my daughter Erin for her love and patience. Finally, I wish to thank my parents, my grandmother, my brother and my sister for their faith in me and unending encouragement in all my endeavors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Problem Statement

Effective searching and navigation of web resources is at the forefront of issues related to the area of information organization. As networked information resources on the web continue to grow rapidly, the need for effective access to better organized information has received a lot of attention. Morville (2005) points out that findability is the most important issue in an information overload environment. Given the growing number of web resources, tools for organization and providing access to the web have been developed. Subject gateways and web directories are such tools, designed to provide access to quality resources selected and indexed by experts or information professionals.

However, one of the problems with current organization systems for web resources is that they were developed using traditional library schemes for subject access based on controlled vocabulary. Nicholson et al. (2001) point out problems with controlled vocabularies including a lack of or excessive specificity in subject areas. Shirky (2005a) asserts that formal classification systems are not suitable for electronic resources. As Mai (2004a) notes, traditional classification schemes have difficulties with representing knowledge, and the problems of describing the subject matter of web documents have not received sufficient attention. Mai (2004a) posits the following two main obstacles for applying bibliographic classification principles to the classification of the web:

a. The principles are tied to the paper-based environment and,

b. The principles have been focused on organizing scientific or scholarly material.

The other problem with current approaches to organizing the web via gateways and directories is that web documents have been organized and indexed by professional indexers. Although there have been efforts to involve users in developing organization systems, they are not necessarily based on users' natural language.

Accordingly, social tagging has received significant attention since it helps organize contents by collaborative and user-generated tags. Users' tags reflect their language because they allow users to add their own tags based on their interests. Several researchers have discussed the impact of tagging on retrieval performance on the web (Bao et al., 2007; Choy and Lui, 2006; Golder and Huberman, 2006; Heymann et al., 2008; Kipp and Campbell, 2010; Sen et al., 2006; Yanbe et al., 2006). Choy and Lui (2006) have applied the statistical tool of Latent Semantic Analysis (LSA) to the evaluation of tag similarity by examining pairs of tags of singular and plural forms, and concluded that collaborative tagging has a great impact on retrieval. Yanbe et al. (2006) have explored an approach to enhancing search by proposing combining a link-based ranking metric with social tagging data, and investigated the utility of social bookmarking systems. Bao et al. (2007) have explored the use of social annotations to improve web search and stated that social annotations could be useful for web search by focusing on two aspects: similarity ranking (between a query and a web page) and static ranking. Kipp and Campbell (2010) have examined whether tags would be useful for information retrieval by limiting the scope of information to scholarly documents such as academic articles at CiteULike and Pubmed online journal database. Furthermore, several researchers have discussed the usefulness of social tagging for cataloging and classification by examining the linguistic aspects of user vocabulary (Makani and Spiteri, 2010; Spiteri, 2007). However, further research is needed to qualitatively as well as quantitatively investigate social tagging and to systematically verify its quality and accuracy, which is the first necessary step to utilize social tagging in digital information organization.

1.2 Purpose of the Study

To address identified problems with current web organization systems, this study aims to investigate whether user-generated tags through social tagging could be used to enhance access to web resources and provide additional access points beyond professionally-generated ones, and whether we could verify the usefulness of social tagging to obtain benefit from it. The main objective of this study is focused on examining the inter-indexer consistency of social tagging for systematically verifying its efficacy and quality. Traditionally, consistency in indexing is considered as an indication of its quality (Cooper, 1969; Rolling, 1981). Leonard (1977) asserts that indexing consistency has a positive influence on retrieval effectiveness.

Furthermore, vocabulary analysis of both users' and professionals' terms is necessary in order to determine the quality of terms in subject indexing. An analysis of exhaustivity and specificity in relation to tagging effectiveness was conducted to provide in-depth examination of tags.

Most indexing consistency studies have been conducted with a small number of professional indexers, and they tended to exclude users. Additionally, the studies mainly have focused on physical library collections, for example, physical books and periodicals rather than web resources. This dissertation research intends to bridge the following five research gaps: (1) no systematic research demonstrating the usefulness of social tagging as subject indexing in comparison to professional indexing, (2) no indexing consistency research measuring consistency on web resources across professionally indexed web directories, e.g., subject gateways, (3) insufficiency of studies on vocabulary analysis comparing user-generated tags with professional-generated index terms regarding web resources, (4) lack of comprehensive indexing consistency research including a large number of users and web resources, and (5) lack of extensive indexing consistency research covering a variety of kinds of web resources.

(1) *No systematic research demonstrating the usefulness of social tagging as subject indexing in comparison to professional indexing*

Several researchers have discussed social tagging behavior and its usefulness for classification or retrieval. However, there was no systematic research which employed quantitative as well as qualitative analysis for the comparison of social tagging with professional indexing in terms of subject indexing of documents.

(2) *No indexing consistency research measuring consistency on web resources across professionally indexed web directories*

No research has examined indexing performance on web resources across human expert-indexed web directories, i.e., subject gateways, which are clear candidates for discussing the professional’s point of view on web documents.

(3) *Insufficiency of studies on vocabulary analysis comparing user-generated tags with professional-generated index terms regarding web resources*

Little qualitative analysis has been conducted on users' tagging data by comparing tags with index terms which are created by professional indexers regarding web resources.

(4) *Lack of comprehensive indexing consistency research including a large number of users and web resources*

Most indexing consistency studies have been conducted with a small number like two or three groups of professional indexers, and they tended to exclude users. Additionally, the studies mainly have focused on physical library collections, for example, physical books and periodicals rather than web resources.

(5) *Lack of extensive indexing consistency research covering a variety of kinds of web resources*

Although there were indexing consistency studies (Kipp, 2010a; Wolfram and Olson, 2007), with a large number of indexers or taggers, they examined only a limited scope of documents, e.g., journal articles on CiteULike.org.

1.3 Research Questions

The following research questions are central in this study.

- Would social tagging be useful for subject indexing in organizing and providing access to the web? Could we verify the usefulness of social tagging to obtain benefit from it?

- How are web resources tagged or indexed at a social tagging site? What kinds of benefits could we obtain from tags?

The following specific research questions will be addressed when exploring the main areas of focus mentioned above.

1) How consistent is social tagging at Delicious regarding subject indexing of web resources? Is there a relationship between its indexing consistency and subject areas indexed?

2) How consistent is professional indexing between BUBL and Intute?

3) Are there various or alternative interpretations of the same web document between two professionally indexed subject gateways, BUBL and Intute?

4) How consistent is tagging/indexing between Delicious taggers and Intute professionals?

5) Would Delicious users' tags provide additional subject access points beyond index terms or keywords that Intute professionals provide?

6) What levels of tagging exhaustivity and tag specificity in Delicious characterize the indexing of web documents?

7) What are features and patterns of social tagging in describing a web document at Delicious? Do tags have other bibliographic attributes beyond describing subjects or topics of a document?

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Libraries have a long history in organizing and providing access to resources. As networked information resources on the web continue to grow rapidly, today's digital library environments have led librarians and information professionals to index and manage digital resources on the web. Thus, this trend has required new tools for organizing and providing more effective access to the web. Subject gateways and web directories are such tools for internet resource discovery. Yet, studies have shown that such tools based on traditional organization schemes are not sufficient for the web. Social tagging has received significant attention since it helps organize contents by collaborative indexing based on user-generated tags. Several researchers have discussed social tagging behavior and its usefulness for classification or retrieval. However, further research is needed to qualitatively as well as quantitatively investigate social tagging and to justify its efficacy and benefit. In order to increase the utilization of social tagging data, this dissertation research particularly examines the inter-indexer consistency of social tagging.

Section 2.2.1 provides the key definitions of subject gateways and their general background as tools for organizing the Web in order to address how professionally indexed web directories are characterized. The following sections present the details of BUBL and Intute which are examples of such tools and also are target subject gateways of this research for a comparison with a social tagging site. Section 2.2.2 discusses advantages of controlled vocabulary which has been traditionally used for subject indexing, and points out challenges of controlled vocabulary for the web with the intention to emphasize the need for social tagging data as natural language terms.

Section 2.3 discusses several points related to the issue of social tagging since it is a core concept of this dissertation research. Section 2.3.1 provides the definitions of the terms social tagging and folksonomy with the aim to provide a good understanding of the concepts. Section 2.3.2 describes an exemplary social tagging site, Delicious, which is a target social tagging site from which this research collected tags for analysis. Section 2.3.3 illustrates social tagging in subject indexing in order to provide appropriate context for the purpose of this study. Section 2.3.4 reviews related research which investigates tagging as a more accurate description of resources and reflection of more current terminology than subject headings assigned by the Library of Congress Subject Headings (LCSH). Section 2.3.5 briefly summarizes criticisms of folksonomy which should not be ignored. This leads to the following section covering the combination of controlled vocabulary and uncontrolled vocabulary.

Finally, section 2.4 discusses the measure used in this research to compare different forms of indexing.   Section 2.4.1 provides a brief overview of inter-indexer consistency. Section 2.4.2 explains a variety of methods of measuring inter-indexer consistency and justifies the choice of the approach used in this study for measuring three principal values: the inter-indexer consistency (1) among social tagging users, (2) between two groups of professional indexers, and (3) between social tagging users and professional indexers.

2.2 Organization of the Web

2.2.1 Subject gateways as organizing tools for the web

A growing number of web resources have required new tools for organizing and providing more effective access to the web. Subject gateways and web directories are such tools for internet resource discovery. Subject gateways can range from "loosely collated commercial directories" such as Yahoo! subject categories, to "collections of quality assessed web resources compiled by the academic or research community" (University of Kent, 2009). In this study, I will refer to the concept of the latter for further discussion.

The subject gateways emerged in response to the challenge of "resource discovery" in a rapidly developing Internet environment in the early and mid-1990s. The term "subject gateway" was commonly used in the UK Electronic Libraries Programme (eLib)[1] (Dempsey, 2000). Under the eLib project, Internet subject gateways were established to deal with Internet searching problems, such as finding good quality and relevant resources (Burton and Mackie, 1999). The EU project DESIRE[2] (Development of a European Service for Information on Research and Education) invented the term "subject-based information gateway (SBIG)" which looks like almost a synonym with the term "subject gateway" (Koch, 2000). Koch (2000) refers to

[1] eLib was a JISC-funded programme of projects in 1996 (initially £15m over 3 years but later extended to 2001). Projects included Digitisation, Electronic Journals, Electronic Document Delivery and On-Demand Publishing (Hiom, 2006).

[2] The DESIRE project (from July 1998 until June 2000) was a collaboration between project partners working at ten institutions from four European countries - the Netherlands, Norway, Sweden and the UK. The project focused on improving existing European information networks for research users in Europe in three areas: Caching, Resource Discovery and Directory Services (DESIRE Consortium, 2000).

“information gateways” by defining them as “quality controlled information services”. Sometimes, subject gateways are termed “quality gateways”, “subject directories” or “virtual libraries” (Bawden and Robinson, 2002).

Although there is no precise definition of subject gateways, they share several characteristics (Bawden and Robinson, 2002):

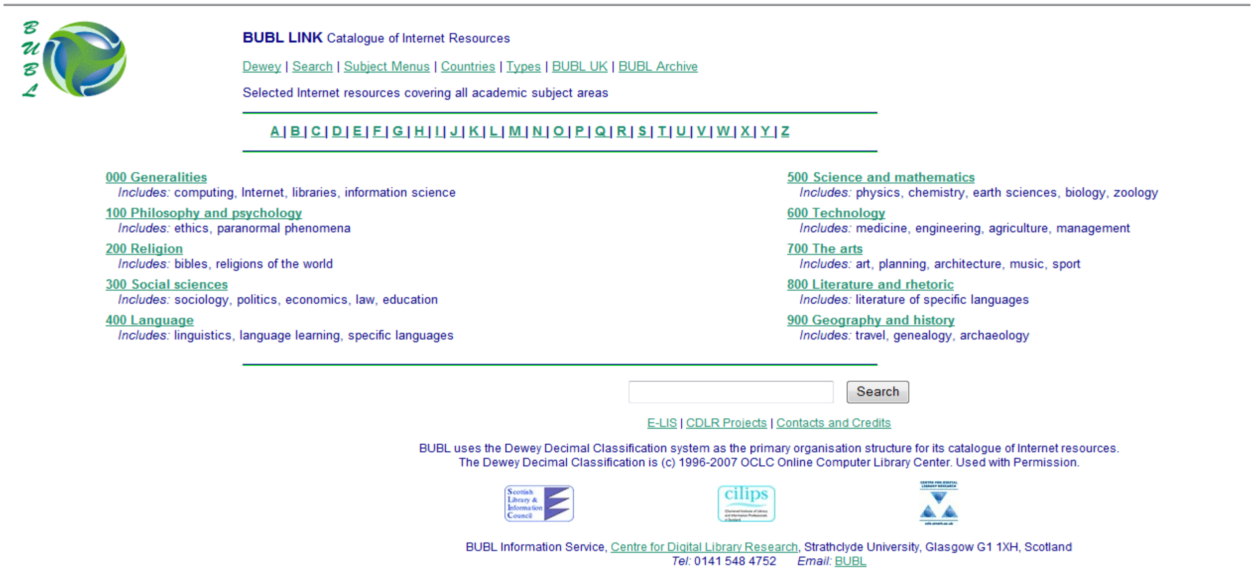
- “a clearly expressed subject scope, defining what resources may be considered for inclusion”,
- “explicitly defined criteria of quality, used to select resources for inclusion”,
- “some form of annotation or description of resources”,
- “some categorization, classification or indexing of the collection”,
- “clearly defined responsibilities for their creation and maintenance”

Subject gateways can be enumerated by the subject categories which they cover (University of Kent, 2009). For instance, Social Care Online (http://www.scie-socialcareonline.org.uk/) (professional development support portal), SocioSite (http://www.sociosite.net/) (the University of Amsterdam's social science information system), and SWAP (Social Policy and Social Work) (http://www.swap.ac.uk/) (subject portal providing resources to support teachers and lecturers in this subject) are subject gateways which provide resources in social science subjects. For a psychology subject area, there are CogNet (http://cognet.mit.edu/) (MIT portal for the brain sciences), PsychNet.UK (http://www.psychnet-uk.com/) (a comprehensive UK gateway to psychology information) and so on. Doctors.net.uk (http://www.doctors.net.uk/) (Peer led internet resource for UK doctors) and HON (Health On the Net) (http://www.hon.ch/) (international Swiss initiative to make quality guidance about medical treatments and health information available to patients and public) are examples for health and medicine subjects. As examples of subject gateways covering various subject areas, there are BUBL Link (http://www.bubl.ac.uk/index.html) and Intute (http://www.intute.ac.uk/). BUBL describes itself as ‘Free User-Friendly Access to selected internet resources covering all subject areas, with a special focus on Library and Information Science’ (Wikipedia). Intute is a free web service aimed at students, teachers, and researchers in UK further education and higher education (Wikipedia). In the following sections, more details about BUBL and Intute are presented.

2.2.1.1 BUBL

The BUBL Information Service is “an Internet link collection for the library and higher education communities, operated by the Centre for Digital Library Research at the University of Strathclyde, and its name was originally short for Bulletin Board for Libraries” (Wikipedia). Since 1993 the BUBL Information Service has been a structured and user-friendly gateway for web resources in order to direct librarians, information professionals, academics and researchers (Gold, 1996).

Many subject gateways provide controlled vocabularies: either “home-made” or “standard library/information tools” such as classification schemes, subject headings and thesauri (Bawden and Robinson, 2002). BUBL offers broad categorization of subjects based on the Dewey Decimal Classification scheme (BUBL Link Home) (See Figure 1). For each subject, subject specialists like librarians work on the maintenance and development of subject categories. However, it has been noted that BUBL is no longer being updated as of April 2011 (BUBL Link Home), as support for BUBL is being discontinued.



BUBL LINK Catalogue of Internet Resources

Dewey | Search | Subject Menus | Countries | Types | BUBL UK | BUBL Archive

Selected Internet resources covering all academic subject areas

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

000 Generalities
Includes: computing, Internet, libraries, information science

100 Philosophy and psychology
Includes: ethics, paranormal phenomena

200 Religion
Includes: bibles, religions of the world

300 Social sciences
Includes: sociology, politics, economics, law, education

400 Language
Includes: linguistics, language learning, specific languages

500 Science and mathematics
Includes: physics, chemistry, earth sciences, biology, zoology

600 Technology
Includes: medicine, engineering, agriculture, management

700 The arts
Includes: art, planning, architecture, music, sport

800 Literature and rhetoric
Includes: literature of specific languages

900 Geography and history
Includes: travel, genealogy, archaeology

E-LIS | CDLR Projects | Contacts and Credits

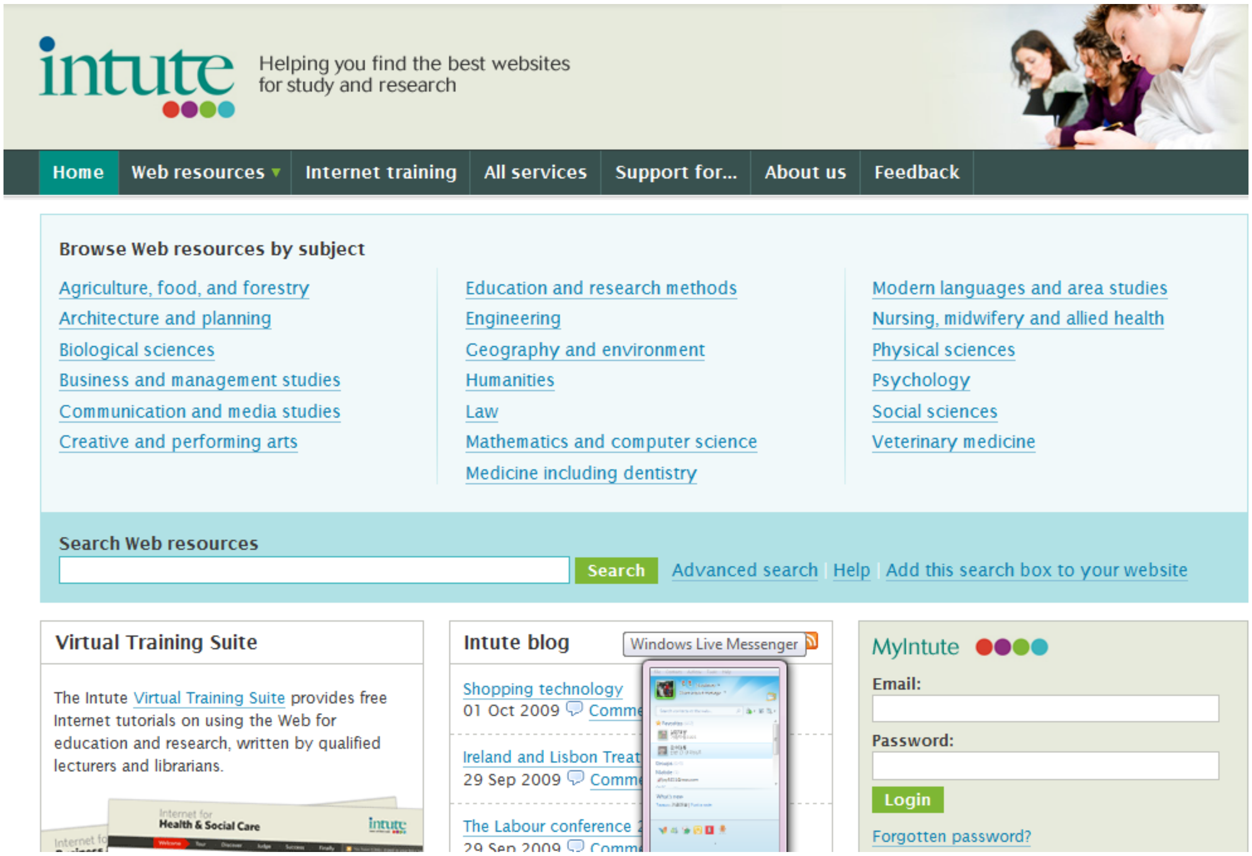
BUBL uses the Dewey Decimal Classification system as the primary organisation structure for its catalogue of Internet resources.
The Dewey Decimal Classification is (c) 1996-2007 OCLC Online Computer Library Center. Used with Permission.

BUBL Information Service, Centre for Digital Library Research, Strathclyde University, Glasgow G1 1XH, Scotland
Tel: 0141 548 4752 *Email:* BUBL

Figure 1. A screenshot of BUBL home page

2.2.1.2 Intute

Intute is funded by the Joint Information Systems Committee (JISC) which supports "education and research by promoting innovation in new technologies and by the central support of ICT services" in the UK higher and further education sectors (JISC Home). Intute offers a searchable and browsable database of web resources that subject specialists select, evaluate and describe (Joyce et al., 2008) (see Figure 2).



Browse Web resources by subject

- Agriculture, food, and forestry
- Architecture and planning
- Biological sciences
- Business and management studies
- Communication and media studies
- Creative and performing arts
- Education and research methods
- Engineering
- Geography and environment
- Humanities
- Law
- Mathematics and computer science
- Medicine including dentistry
- Modern languages and area studies
- Nursing, midwifery and allied health
- Physical sciences
- Psychology
- Social sciences
- Veterinary medicine

Search Web resources

Virtual Training Suite

The Intute Virtual Training Suite provides free Internet tutorials on using the Web for education and research, written by qualified lecturers and librarians.

Intute blog

Shopping technology
01 Oct 2009 Comme

Ireland and Lisbon Treat
29 Sep 2009 Comm

The Labour conference
29 Sep 2009 Comm

Email:

Password:

Figure 2. A screenshot of Intute home

Intute was formed in July 2006 after the Resource Discovery Network (RDN)[3]'s eight hubs were merged. These hubs respectively serve particular academic disciplines (Wikipedia):

- Altis - Hospitality, leisure, sport and tourism
- Artifact - Arts and creative industries
- Biome - Health and life sciences
- EEVL - Engineering, mathematics, and computing
- GEsource - Geography and the environment

[3] The Resource Discovery Network (RDN) is a JISC-funded national service. It is supported by the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC), in order to provide quality internet service for the education community. The RDN originated in the Electronic Libraries (eLib) Programme (Hiom, 2006).

- Humbul - Humanities
- PSIgate - Physical sciences
- SOSIG - Social sciences

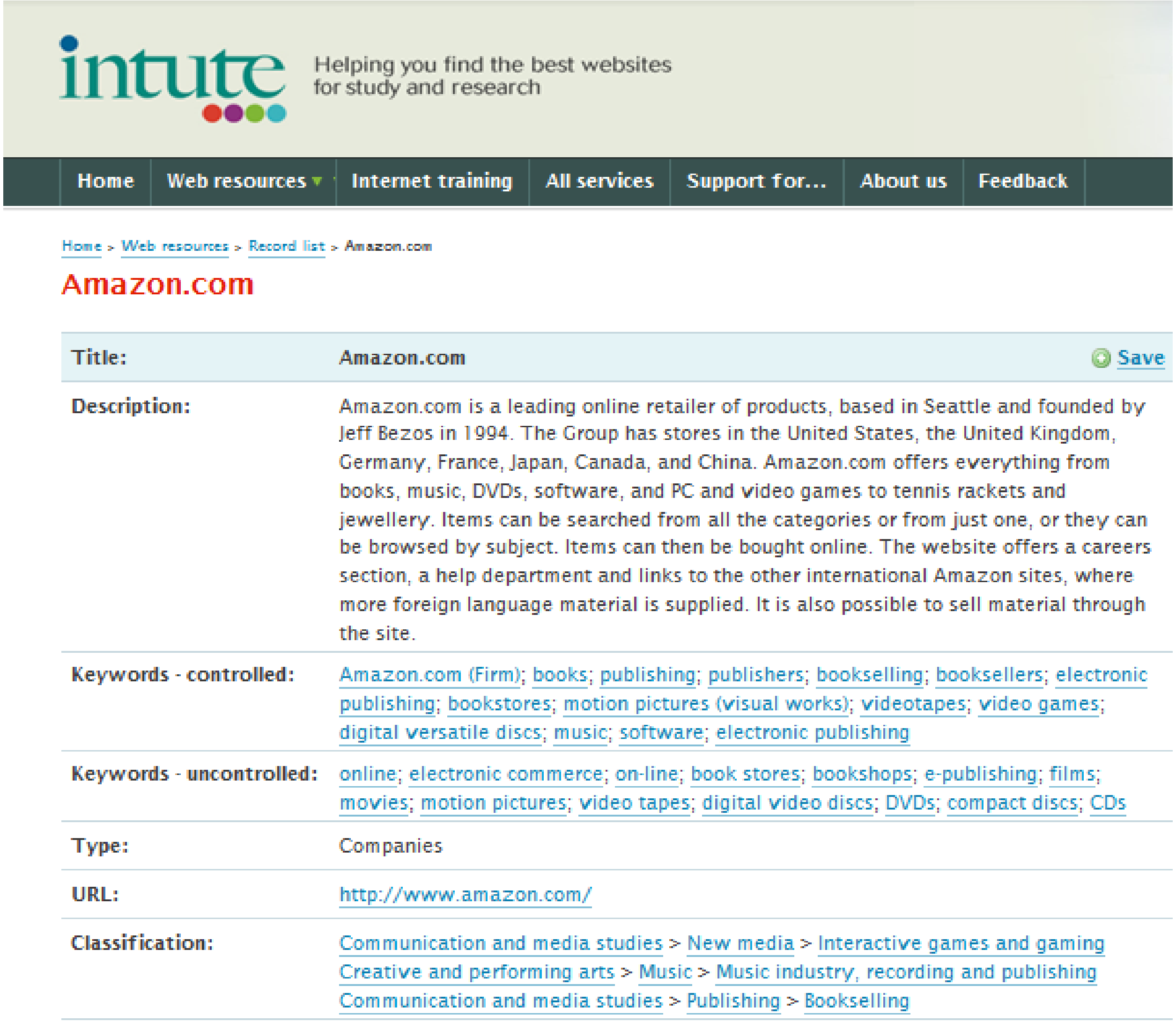
Intute is created by a consortium of seven universities and its service is offered by staff at those seven locations, i.e. University of Birmingham (Intute Social Sciences), University of Bristol (Intute Social Sciences and Intute Virtual Training Suite), Heriot-Watt University (Intute Science, Engineering and Technology), The University of Manchester (Intute Executive), Manchester Metropolitan University (Intute Science, Engineering and Technology), University of Nottingham (Intute Health and Life Science), and University of Oxford (Intute Arts and Humanities) (Intute Home).

The selection for inclusion of resources within the Intute collection considers the quality, relevance and provenance of resources (Robert Abbott, personal communication, May 21, 2009). It is reported that Intute mainly uses the Universal Decimal Classification (UDC) and DDC for classification and has adapted them for in-house use. Intute subject specialists collaboratively catalog web documents. A web document cataloged by one indexer is passed to another specialist for checking it according to their cataloguing guidelines before it is added to the database (Anne Reed, personal communication, July 14, 2010).

Intute also uses several thesauri for its subject relevance and comprehensiveness (A. M. Joyce, personal communication, June 2, 2009). For instance, the SCIE for keywords of Social Welfare subjects, the Hasset, IBSS, LIR for Law, and the NLM MeSH headings for Medicine. In some cases, e.g., Nursing, they index according to more than one thesaurus. Other subjects such as

Arts and Humanities apply similar principles (Robert Abbott, personal communication, May 21, 2009).

Intute offers index strings based on classification schemes and sometimes it provides keywords (controlled or uncontrolled or both) generated by professional indexers. Allocated keywords are reviewed by a group of subject indexers for consistent keywording (Anne Reed, personal communication, July 14, 2010). Uncontrolled keywords are added if indexers can find no suitable word in the above thesauri. They choose the uncontrolled keywords from among terms occurring in the titles and descriptions they write for the resources. They tend to select the uncontrolled keywords from among the words that the web sites themselves use (A. M. Joyce, personal communication, June 2, 2009). Figure 3 shows how Intute indexes a document, *Amazon.com* and how they present several types of information about the document including description, controlled keywords, uncontrolled keywords, type, URL, and category paths of classification. However, it has been recently noted that Intute is closing after July 2011 (Intute Home), as support for Intute is being discontinued.


intute
Helping you find the best websites for study and research
Home
Web resources
Internet training
All services
Support for...
About us
Feedback
Home > Web resources > Record list > Amazon.com
Save

Amazon.com

Title:	Amazon.com
Description:	Amazon.com is a leading online retailer of products, based in Seattle and founded by Jeff Bezos in 1994. The Group has stores in the United States, the United Kingdom, Germany, France, Japan, Canada, and China. Amazon.com offers everything from books, music, DVDs, software, and PC and video games to tennis rackets and jewellery. Items can be searched from all the categories or from just one, or they can be browsed by subject. Items can then be bought online. The website offers a careers section, a help department and links to the other international Amazon sites, where more foreign language material is supplied. It is also possible to sell material through the site.
Keywords - controlled:	Amazon.com (Firm); books; publishing; publishers; bookselling; booksellers; electronic publishing; bookstores; motion pictures (visual works); videotapes; video games; digital versatile discs; music; software; electronic publishing
Keywords - uncontrolled:	online; electronic commerce; on-line; book stores; bookshops; e-publishing; films; movies; motion pictures; video tapes; digital video discs; DVDs; compact discs; CDs
Type:	Companies
URL:	http://www.amazon.com/
Classification:	Communication and media studies > New media > Interactive games and gaming Creative and performing arts > Music > Music industry, recording and publishing Communication and media studies > Publishing > Bookselling

Figure 3. An example of an indexed document in Intute

2.2.2 Challenges of controlled vocabulary for the web

For effective indexing and retrieval, the indexing process needs to be controlled by using a so-called *controlled vocabulary* (Lancaster, 1972). Since the 19th century, controlled vocabularies have been developed and used for subject indexing. Lancaster (2003) identifies three major manifestations of controlled vocabulary: bibliographic classification schemes, subject heading lists and thesauri.

Controlled vocabulary has many advantages. One of the major advantages of controlled vocabulary is that it can increase the effectiveness of retrieval by providing unambiguous, standard search terms with a control of polysemy, synonymy, and homonymy of the natural language (Golub, 2006; Muddamalle, 1998).

Another benefit from controlled vocabulary is that it improves the matching process with its systematic hierarchies of concepts featuring a variety of relationships like “broader term,” “narrower term,” “related term,’ or “see” and “see also” (Golub, 2006; Olson and Boll, 2001).

However, as there are more and more resources available on the web, existing controlled vocabularies have been challenged in their ability to index the range of digital web resources. The challenges of controlled vocabulary for the web can be summarized as follows.

One of the major challenges of controlled vocabulary in the digital environment is the slowness of revision. Indexing web content requires an updated thesaurus, but usually subjects are rapidly evolving with new terminology, so it is hard to always keep up-to-date vocabulary (Muddamalle, 1998). Golub (2006) also addresses “improved currency” and “hospitality for new topics” as new roles which controlled vocabularies need to take.

The other problem is that the construction of controlled vocabularies and indexing are labor-intensive and expensive (Fidel, 1991; Macgregor and McCulloch, 2006). The process of indexing is conducted by professional efforts requiring expert knowledge (Olson and Boll, 2001).

Another obstacle of controlled vocabulary is that it has been developed with a focus on physical and traditional library collections. Traditionally, controlled subject headings have been employed for indexing physical resources, so they need to be flexible or expandable in order to encompass web resources (Golub, 2006; Nowick and Mering, 2003; Macgregor and McCulloch, 2006). For instance, LCSH is designed to describe monographs and serials, so it might not be specific enough for describing web resources (Nowick and Mering, 2003).

Furthermore, Nicholson et al. (2001) have discussed the problems with controlled vocabularies in indexing for describing online collections by identifying that “they have a lack of, or excessive, specificity in the subject areas”.

Last but not least, controlled vocabulary should be comfortable for users to use, and it should be able to meet the users’ interests and their needs (Golub, 2006). Golub mentions “intelligibility, intuitiveness, and transparency” as new challenges for controlled vocabulary.

Using free-text or natural language terms is one alternative to resolve identified problems with controlled vocabulary. Advantages of free-text terms are that they require only non-professional knowledge for searching techniques for users, and reflect up-to-date vocabulary (Dubois, 1987).

Social tagging data is one example of natural language terms, that is, uncontrolled vocabulary assigned by users. Social tagging is a promising way to complement the disadvantages of professional indexing because it is low-cost since a great number of users from everywhere contribute to the creation of tags. Thus, users' tags might be alternate terms with additional entry points of retrieval which are not easily attained using controlled vocabularies (Hayman, 2007; Maltby, 1975; Quintarelli, 2005). Tags are generally much more current than controlled vocabulary since they are constructed in the process of "sensemaking" in that users share their experiences in subject terms reflecting their interests in various communities (Smith, 2007). Unlike hierarchical structures (broader and narrower terms) of controlled vocabularies, folksonomies are inherently flat which allows great flexibility in indexing terms (Smith, 2007). Moreover, as investment in professionally-developed subject gateways and web directories diminishes (support for both BUBL and Intute, examined in this study, is being discontinued), understanding characteristics of social tagging becomes even more critical. In the next section, more details about social tagging and relevant issues will be described.

2.3 Social Tagging and Folksonomy

2.3.1 Definitions of terms

Social tagging is described as "user-generated keywords" (Trant, 2009). Since tags indicate users' perspectives and descriptions in indexing resources, they have been suggested as a means to improve search and retrieval of resources on the web. The term "social tagging" is frequently associated with the term "folksonomy" which was coined by Thomas Vander Wal from 'folk' and 'taxonomy' (Smith, 2004). Folksonomy consists of three elements: users, resources to be described, and tags for describing resources (Vander Wal, 2005a). Vander Wal (2007) describes "folksonomy" as "user-created bottom-up categorical structure development with an emergent thesaurus". Quintarelli (2005) defines folksonomy as "user-generated classification, emerging through bottom-up consensus." Examples of folksonomy sites include Flickr, Del.icio.us, and LibraryThing.

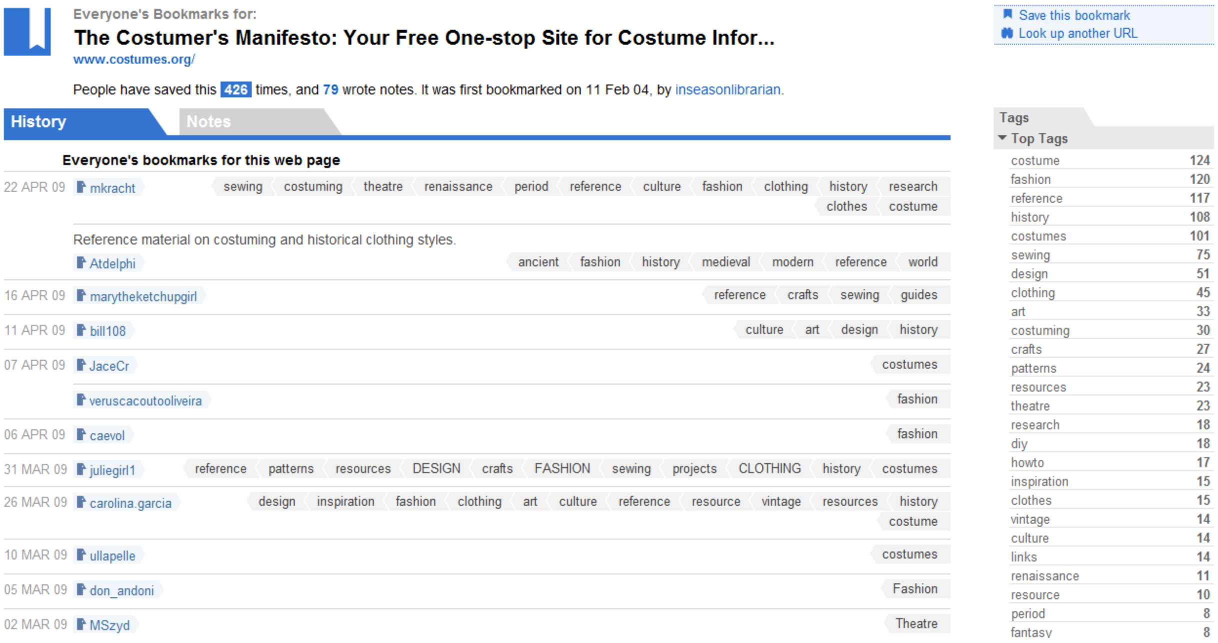
While Trant (2009) provides good reviews of the overall trends of research on social tagging and folksonomy, she distinguishes the two terms "social tagging" and "folksonomy" by providing short definitions:

- Tagging: "a process with a focus on user choice of terminology"
- Folksonomy: "the resulting collective vocabulary (with a focus on knowledge organization)"
- Social tagging: "a sociotechnical context within which tagging takes place (with a focus on social computing and networks)"

In addition, other terms have been used by several researchers like "social classification" (Furner and Tennis, 2006; Landbeck, 2007; G. Smith, 2004; Trant, 2006), "community cataloguing" and "cataloguing by crowd" (Chun and Jenkins, 2005), "communal categorization" (Strutz, 2004), and "ethnoclassification" (boyd, 2005; Merholz, 2004). These terms describing this phenomenon are not well defined yet, and they have often been selected depending on focal points, e.g., sociability, collaboration and cooperation (Vander Wal, 2005a; Weinberger, 2006). Sometimes, these terms are also regarded as synonyms. For example, Noruzi (2006) notes folksonomy as a synonym of social tagging while describing its characteristics.

2.3.2 An exemplary social tagging site: Delicious

Social tagging has been popularized by tagging sites such as Flickr, Technorati and Deli.cio.us. Deli.cio.us is one of the most popular social bookmarking services, allowing users to add or share and organize tags. Deli.cio.us now redirects to the new domain, Delicious. The site was established by Joshua Schachter in 2003 and acquired by Yahoo! in 2005 (Wikipedia). Figure 4 shows how a web document is tagged by users at Delicious. Delicious provides "Top Tags" lists at the right side of the screen, and these ranked tags are not checked for variant spellings, synonyms, singular vs. plural etc. For instance, "costume" and "costumes" are both ranked.



Everyone's Bookmarks for:

The Costumer's Manifesto: Your Free One-stop Site for Costume Infor...

www.costumes.org/

People have saved this 426 times, and 79 wrote notes. It was first bookmarked on 11 Feb 04, by inseasonlibrarian.

Everyone's bookmarks for this web page

22 APR 09	mkracht	sewing costuming theatre renaissance period reference culture fashion clothing history research clothes costume
	Reference material on costuming and historical clothing styles. Atdelphi	ancient fashion history medieval modern reference world
16 APR 09	marytheketchupgirl	reference crafts sewing guides
11 APR 09	bill108	culture art design history
07 APR 09	JaceCr	costumes
	veruscacoutooliveira	fashion
06 APR 09	caevol	fashion
31 MAR 09	juliegirl1	reference patterns resources DESIGN crafts FASHION sewing projects CLOTHING history costumes
26 MAR 09	carolina.garcia	design inspiration fashion clothing art culture reference resource vintage resources history costume
10 MAR 09	ullapelle	costumes
05 MAR 09	don_andoni	Fashion
02 MAR 09	MSzyd	Theatre

Tags

▾ Top Tags

costume	124
fashion	120
reference	117
history	108
costumes	101
sewing	75
design	51
clothing	45
art	33
costuming	30
crafts	27
patterns	24
resources	23
theatre	23
research	18
diy	18
howto	17
inspiration	15
clothes	15
vintage	14
culture	14
links	14
renaissance	11
resource	10
period	8
fantasy	8

Figure 4. An example of Delicious tags

2.3.3 Social tagging and subject indexing

Many researchers stress the need to add users to the development of controlled vocabularies for subject indexing (Abbott, 2004; Mai, 2004b; Quintarelli, 2005; Shirky, 2005b). Fidel (1991) asserts that online searchers use rules in an “intuitive way” to help their selection of search keys and these rules can be formalized.

Many researchers have suggested that social tagging has potential for user-based indexing (Golder and Huberman, 2006b; Lin et al., 2006; Tennis, 2006). It can be recognized that the participation of users in building controlled vocabulary is being realized in a social tagging environment where users create or generate search keywords based on their intuitive principles.

Olson and Wolfram (2006) posit that social tagging could be utilized to index web resources by adding keywords which are being used by users. They also describe the concept of tagging as indexing performance in that people create and share their identified terms to describe contents of web documents. Lin et al. (2006) describe “emerging characteristics of social classification” and the relationship between tags and index terms. Voss (2007) also argues that it is more acceptable to see that tagging is a common means of manual indexing on the web. In addition, Trant (2009) asserts that a folksonomy can be studied in relationship to other indexing vocabularies since it provides additional access points to resources.

2.3.4 Related research

There has been exploratory research investigating tagging as a more accurate description of resources and reflection of more current terminology. Smith (2007) has asserted that tagging is better than subject headings by investigating tags assigned in LibraryThing and the subject headings assigned by the Library of Congress Subject Headings (LCSH). LibraryThing is a website that allows users to manage a personal catalog with their own books (Wikipedia). Smith sampled five books including both fiction and nonfiction works published in the past five years. She analyzed the LCSH terms assigned to the book and the tag clouds and confirmed that the folksonomy has potential for augmenting subject analysis tools (see Figure 5).

LibraryThing	**LCSH**
Tags used to describe the book	England > Fiction
2005(42) Adventure(36) **boarding school**(22) british(69) children(136) children's fiction(42) children's literature(69) childrens(361) england(41) **fantasy**(1,309) favorites(58) **fiction**(967) hardcover(35) **harry potter**(590) **Hogwarts**(36) juvenile(33) juvenile fiction(16) **magic**(306) novel(60) own(62) **potter**(19) read(139) **rowling**(56) school(33) series(145) unread(16) witches(31) **wizardry**(31) **wizards**(115) young adult(314) youth(19)	England > Juvenile fiction
	Fantasy fiction > Juvenile
	Good and evil > Juvenile fiction
	Hogwarts School of Witchcraft and Wizardry (Imaginary place) > Juvenile fiction
	Intergenerational relations > Juvenile fiction
	Magic > Fiction
	Magic > Juvenile fiction
	Maturation (Psychology) > Juvenile fiction
	Potter, Harry (Fictitious character) > Juvenile fiction
	Schools > Fiction
	Schools > Juvenile fiction
	Wizards > Fiction
	Wizards > Juvenile fiction

Figure 5. Harry Potter tag cloud and subject headings (Source: Smith, 2007)

She hypothesized that LibraryThing would better represent the subject matter of fictional works whereas LCSH would be better at representing the subject of nonfiction works, and she concluded that LibraryThing is better at showing latent subjects when there are fewer synonym redundancies. She also noted that synonyms in the tag clouds allow for some natural language retrieval.

However, although social tagging or folksonomy has shown potential that it improves the retrieval of resources on the web, its problems also have been pointed out by several researchers. The problems of folksonomy are described in the next section.

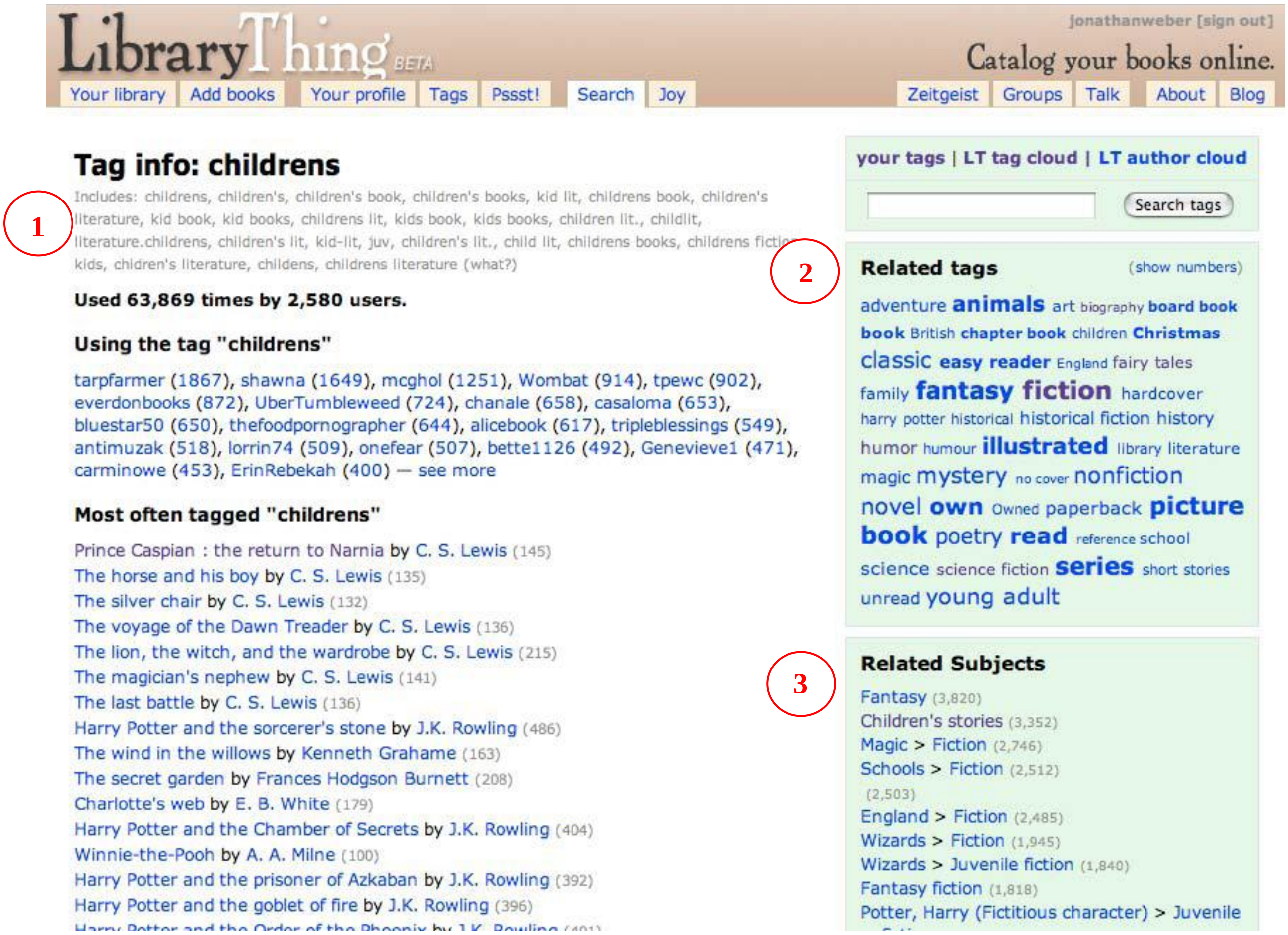
2.3.5 Criticisms of folksonomy

Folksonomy has been criticized with its ambiguity of terms, a large number of synonyms, a lack of hierarchy, unstable term specificity, and variations of spelling etc. (Quintarelli, 2005; Spiteri, 2005). Merholz (2004) also describes drawbacks of tags as synonyms and inaccuracy, and emphasizes the contribution of the traditional classification and vocabulary control. Furthermore, Peterson (2006) criticizes folksonomy in that it has an intrinsic defect caused by its inability to produce the accuracy of formal classification.

2.3.6 Combination of controlled vocabulary and uncontrolled vocabulary

As discussed, both controlled vocabulary and uncontrolled vocabulary have their own advantages and disadvantages. Several researchers suggest the combination of both approaches since both may complement each other. Macgregor and McCulloch (2006) argue that it is obvious that controlled vocabularies and collaborative tagging systems will coexist: what they describe as “the dichotomous co-existence”.

Knapp et al. (1998)’s study illustrates that combining both approaches produced more effective retrieval performance rather than using only one approach. They conducted an experimental study to identify whether the free-text search terms could add supplementary relevant documents which are not retrieved by the controlled vocabulary. Their study allowed humanities scholars to search using both controlled vocabulary and free-text terms. Its results showed that when controlled vocabulary and free-text terms work together, more relevant records are retrieved.



Tag info: childrens

Includes: childrens, children's, children's book, children's books, kid lit, childrens book, children's literature, kid book, kid books, childrens lit, kids book, kids books, children lit., childlit, literature.childrens, children's lit, kid-lit, juv, children's lit., child lit, childrens books, childrens fictio kids, chidren's literature, childens, childrens literature (what?)

Used 63,869 times by 2,580 users.

Using the tag "childrens"

tarpfarmer (1867), shawna (1649), mcghol (1251), Wombat (914), tpewc (902), everdonbooks (872), UberTumbleweed (724), chanale (658), casaloma (653), bluestar50 (650), thefoodpornographer (644), alicebook (617), tripleblessings (549), antimuzak (518), lorrin74 (509), onefear (507), bette1126 (492), Genevieve1 (471), carminowe (453), ErinRebekah (400) — see more

Most often tagged "childrens"

Prince Caspian : the return to Narnia by C. S. Lewis (145)
The horse and his boy by C. S. Lewis (135)
The silver chair by C. S. Lewis (132)
The voyage of the Dawn Treader by C. S. Lewis (136)
The lion, the witch, and the wardrobe by C. S. Lewis (215)
The magician's nephew by C. S. Lewis (141)
The last battle by C. S. Lewis (136)
Harry Potter and the sorcerer's stone by J.K. Rowling (486)
The wind in the willows by Kenneth Grahame (163)
The secret garden by Frances Hodgson Burnett (208)
Charlotte's web by E. B. White (179)
Harry Potter and the Chamber of Secrets by J.K. Rowling (404)
Winnie-the-Pooh by A. A. Milne (100)
Harry Potter and the prisoner of Azkaban by J.K. Rowling (392)
Harry Potter and the goblet of fire by J.K. Rowling (396)
Harry Potter and the Order of the Phoenix by J.K. Rowling (401)

Related tags (show numbers)

adventure animals art biography board book book British chapter book children Christmas classic easy reader England fairy tales family fantasy fiction hardcover harry potter historical historical fiction history humor humour illustrated library literature magic mystery no cover nonfiction novel own Owned paperback picture book poetry read reference school science science fiction series short stories unread young adult

Related Subjects

Fantasy (3,820)
Children's stories (3,352)
Magic > Fiction (2,746)
Schools > Fiction (2,512)
(2,503)
England > Fiction (2,485)
Wizards > Fiction (1,945)
Wizards > Juvenile fiction (1,840)
Fantasy fiction (1,818)
Potter, Harry (Fictitious character) > Juvenile

Figure 6. LibraryThing tag page for tag "childrens", showing (1) tag combinations, (2) related tags, (3) related subjects (Source: Weber, 2006)

Weber's report (2006) on LibaryThing demonstrates that folksonomies and controlled vocabularies can harmoniously coexist: the combination of both would obtain benefits, and there are useful correlations between the two. Figure 6 illustrates that LibraryThing supplies tag combinations including multiple aspects of the tagged objects, links to statistically related tags and subject headings.

2.4 Inter-Indexer Consistency

2.4.1 Inter-indexer consistency

Caras (1968) distinguishes *inter*-indexer consistency from *intra*-indexer consistency. *Inter*-indexer consistency means the agreement among a group of indexers on the same document and *intra*-indexer consistency means the agreement by the same indexer on the same document at different times.

Zunde and Dexter (1969) define inter-indexer consistency as "the degree of agreement in the representation of the essential information content of the document by certain sets of indexing terms selected individually and independently by each of the indexers in the group". Leonard (1977) also describes it as "a quantitative measure of the degree to which two or more indexers perceive the important information concepts contained in a document and represent these concepts using identical codes and/or terms." This study will use the term "indexing consistency" to refer to inter-indexer consistency.

Different indexers tend to assign different index terms to the same document. In terms of indexing consistency and quality, Cooper (1969) posits that "an increase in consistency can be expected to cause an improvement in indexing quality". Rolling (1981) also points out that if consistency is higher, the quality of indexing will be greater. However, Cooper concludes “Until a more general equation linking interindexer consistency with retrieval effectiveness has been derived, interindexer consistency cannot safely be used as a gauge of indexing quality”. He introduces a concept of *indexer-requester consistency*. That is, when indexers assign an index term to a document and then the index terms occur as search terms which requesters request, we would say that the indexer-requester consistency is high. It has been noted that there exists often a vocabulary mismatch between experts’ technical language and users’ common language (Furnas et al., 1987; Paek and Chandrasekar, 2005). Accordingly, in-depth analysis of users’ terms needs to be undertaken.

2.4.2 Measures of inter-indexer consistency

To evaluate indexers’ indexing performance, it is helpful to measure inter-indexer indexing consistency (David et al., 1995). The measure of inter-indexer consistency has been formulated differently by various researchers. Hooper (1965) has observed that “there is no standard measure of consistency”. Cooper (1969) points out that “this circumstance makes generalization about their findings difficult”. Tonta (1991) also notes that indexing consistency relies on the measure of evaluating consistency.

Leonard (1977) provides good reviews of the studies of inter-indexer consistency which were conducted from the mid-1950s through late-1960s. Chen (2008) describes measures of indexing consistency by dividing them into two categories: pair consistency and group consistency. Pair consistency is measured between two indexers or between different times by the same indexer. Hooper (1965) and Rolling (1981) use the pair consistency to calculate consistency. Where *a* and *b* are respectively the number of terms assigned by two indexers, and *c* designates the number of terms commonly assigned by the two indexers, the measures of Hooper and Rolling are as follows:

- Hooper (Hooper, 1965): Consistency = $c/(a+b-c)$
- Rolling (Rolling, 1981): Consistency= $2c/(a+b)$

These two well-known methods have been widely used for calculating indexing consistency. Horri and Neshat (2006) use Hooper's formula to compare indexing consistency between a pair of catalogers of the National Library of Iran (NLI). Chen (2008) employs both methods of Hooper and Rolling to calculate consistency between two Chinese bibliographic catalogues.

When more than two indexers are included in indexing, group consistency is calculated. Slamecka and Jacoby (1963) compared the indexing consistency among three indexers using controlled vocabulary. Zunde and Dexter (1969) measure group indexing consistency based on the concept of "fuzzy sets"; as they note, "it represents a higher consistency value if indexers agree on the more important terms than if they agree on less important terms". Tonta (1991) discusses the indexing consistency between Library of Congress catalogers and British Library catalogers assigning Library of Congress Subject Headings. David et al. (1995) use a cognitive approach to compare indexing consistency among four experienced indexers.

In today's social tagging environment, it has been acknowledged that traditional methods for assessing inter-indexer consistency need to be extended as a large group of users have been involved in indexing (Olson and Wolfram, 2006). Olson and Wolfram (2006)'s pilot study measured inter-indexer consistency on a large scale using informetric methods.

On the other hand, an Information Retrieval (IR) model identifying similarity between documents has been applied for measuring indexing consistency. Medelyan and Witten (2006) use the cosine metric and calculate consistency between multiple-indexers in a vector space. They focus on the semantic relations between index terms such as RT (related terms) and BT/NT (broader/narrower terms). Wolfram and Olson (2007) propose a new method, the Inter-indexer Consistency Density (ICD) for comparing indexing consistency based on the vector space traditional Information Retrieval (IR) model.

The Vector-based ICD method has an advantage in comparing consistency among a large number of people. Wolfram and Olson (2007) applied the concept of document space in the vector space model into the terms assigned by a group of indexers to a document, and defined an Indexer/Tagger Space. Thus, the Vector-based ICD method represents indexing spaces among indexers, so it is able to deal with consistency analysis among a large number of people such as social tagging users. This dissertation research employs the method of the modified vector-based ICD with three different similarity measures: cosine similarity, dot product similarity, and euclidean distance metric. More details about this method will be described in Section 3.3.2 of the research design and methodology chapter.

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

3.1 Overview

This research investigates tags from a social tagging site, Delicious, which is one of the most popular social bookmarking services. It allows users to extract data from their accounts and to display this data on their own websites. This study collects Delicious tags assigned to web documents listed in two major subject gateways, BUBL and Intute, both of which cover various subjects. This study examines the relationship between indexing consistency of social tagging and the subject areas indexed, comparing indexing consistency of social tagging with that of professional indexing. This research also measures indexing consistency between two groups of professional indexers from BUBL and Intute. In order to measure indexing consistency, this research employs the method of Inter-indexer Consistency Density (ICD), based on the traditional Information Retrieval (IR) Vector Space Model. An analysis of tagging exhaustivity and tag specificity in relation to tagging effectiveness has been conducted to provide in-depth examination of the tags.

This research has been conducted in two stages: data collection and data analysis as shown in Table 1. The first stage of data collection is for the purpose of comparing users' and indexers' vocabulary. The main goal of the second stage is to investigate the efficacy and benefit of social tagging.

Table 1. Research processes and descriptions

Research processes	Descriptions
Data Collection	• Web document samples (commonly indexed in three locations: Delicious, BUBL and Intute) • Users' index terms tagged on the sampled documents (Delicious) • Indexers' index terms (index term strings) on the sampled documents at BUBL and Intute • Users' top ranked tags on the sampled documents at Delicious • Indexers' index keywords on the sampled documents at Intute
Data Analysis	• Analysis of users' tags (Delicious) • Analysis of professional indexers' index terms (BUBL and Intute) • Analysis of users' tags and professionals' keywords (Intute) • Analysis of indexing consistency • Analysis of indexing exhaustivity and specificity • Analysis of tag attributes and tagging behaviors using FRBR

Table 2 summarizes the methods of data analysis for each of the research questions. The data analysis incorporates quantitative and qualitative analyses (see Table 2). The methods are further detailed in subsequent sections.

Table 2. Research questions and methods

Data Analyses	Research Questions (1.-7.)	Methods
Analysis of Tagging/Indexing Consistency	1. How consistent is social tagging at Delicious regarding subject indexing of web resources? Is there a relationship between its indexing consistency and subject areas indexed?	• Measurement of indexing consistency on social tagging using VSM-based Inter-indexer Consistency • Inferential statistics using ANOVA and Kruskal-Wallis test to compare average Inter-Indexer (Tagger) Consistency Density among different subject areas • Normality test on tag data distribution
	2. How consistent is professional indexing between BUBL and Intute? 3. Are there various or alternative interpretations of the same web document between two professionally indexed subject gateways, BUBL and Intute?	• Measurement of indexing consistency on professional indexing using VSM-based Inter-indexer Consistency • Inferential statistics using ANOVA and Kruskal-Wallis test to compare indexing similarity between BUBL and Intute • Qualitative analysis: comparison of indexer's term strings between two subject gateways, BUBL and Intute
	4. How consistent is tagging/indexing between Delicious taggers and Intute professionals? 5. Would Delicious users' tags provide additional subject access points beyond index terms or keywords that Intute professionals provide?	• Measurement of indexing consistency on social tagging and professional indexing using VSM-based Inter-indexer Consistency • Qualitative analysis: comparison of Delicious tags with keywords (controlled or uncontrolled) from Intute
Analysis of Tagging Effectiveness	6. What levels of tagging exhaustivity and tag specificity in Delicious characterize the indexing of web documents?	• Descriptive statistics on the average number of tags per document by subject categories (for tagging exhaustivity) • Descriptive statistics on the number of documents indexed by a tag (for tag specificity) • Latent Semantic Analysis (LSA) on tags in various levels of specificity values • Correlation analysis between tag specificity and LSA values
Analysis of Tag Attributes and Tagging Behaviors	7. What are features and patterns of social tagging in describing a web document at Delicious? Do tags have other bibliographic attributes beyond describing subjects or topics of a document?	• Descriptive statistics of Delicious tag frequency by subject categories • Content analysis of tags based on the FRBR model • Inter-coder reliability for content analysis

3.2 Data Collection

3.2.1 Target social tagging site

This research extracted tags from a social bookmarking site, Delicious. Unlike other social bookmarking sites which provide the number of votes or users' comments, Delicious provides tagging data since they allow users to add or share and organize tags. Additionally, Delicious has a broad coverage of web resources, not limited to scholarly documents (e.g., journal articles on CiteUlike.org) or specific types of resources (e.g., photos and videos on Flickr).

According to Vander Wal's explanation of folksonomy, the broad folksonomy like Delicious has many people tagging the same object and every person can tag the object with their own tags in their own vocabulary while the narrow folksonomy such as Flickr is done by one or a few people providing tags that the person uses to get back to that information (Vander Wal, 2005b). He also claims that the tags in a narrow folksonomy tend to be singular, that is, only one tag with the term is used while many people assign the same tag in the broad folksonomy. Therefore, it is sensible to choose Delicious as a target social tagging site in that it allows investigation of the features and patterns of tags generated by many people on one object.

3.2.2 Target subject gateways

In order to examine professional indexers' vocabulary and compare it with users' vocabulary, this research investigates two major subject gateways: BUBL and Intute (see Table 3).

Table 3. BUBL vs. Intute

Site characteristics	**BUBL**	**Intute**
Classification	DDC	UDC and DDC
Keywords	N/A	*Controlled*: Several thesauri for their subject relevance and comprehensiveness, e.g., SCIE for Social Welfare, the Hasset, IBSS, LIR for Law, and the NLM MeSH headings for Medicine *Uncontrolled*: terms from web sites' titles and descriptions Intute indexers provide
Subjects covered	Various subjects	Various subjects
Database	Searchable and browsable	Searchable and browsable

The reason for these choices is because both BUBL and Intute subject gateways cover various subjects, and this feature allows one-to-one comparison on each subject area which is also dealt with by Delicious.

3.2.3 Sampling of web documents

Sampling documents was based on the 10 subject categories BUBL provides as top-level categories (see Table 4). In order to avoid potential bias in choosing documents at BUBL, a document was first randomly selected from the list of documents associated with a sub-category, and searched in turn at the other two sites, Intute and Delicious. The method of random sampling of documents was based on the True Random Number Generator (www.random.org). If the first document chosen randomly was not found in Intute or Delicious, then the next choice was made randomly until a web document satisfying the selection criteria was found. The selection criteria of sampling web documents are outlined as follows:

(1) Subject categorizations for selecting documents is based on the top-level category at BUBL

(2) A web document must be located at all three web sites, BUBL, Intute, and Delicious

(3) A web document indexed at Intute must include both classification category paths and keywords

(4) A web document having more than 50 taggers at Delicious will be selected in order to have a sufficient number of taggers for measuring consistency

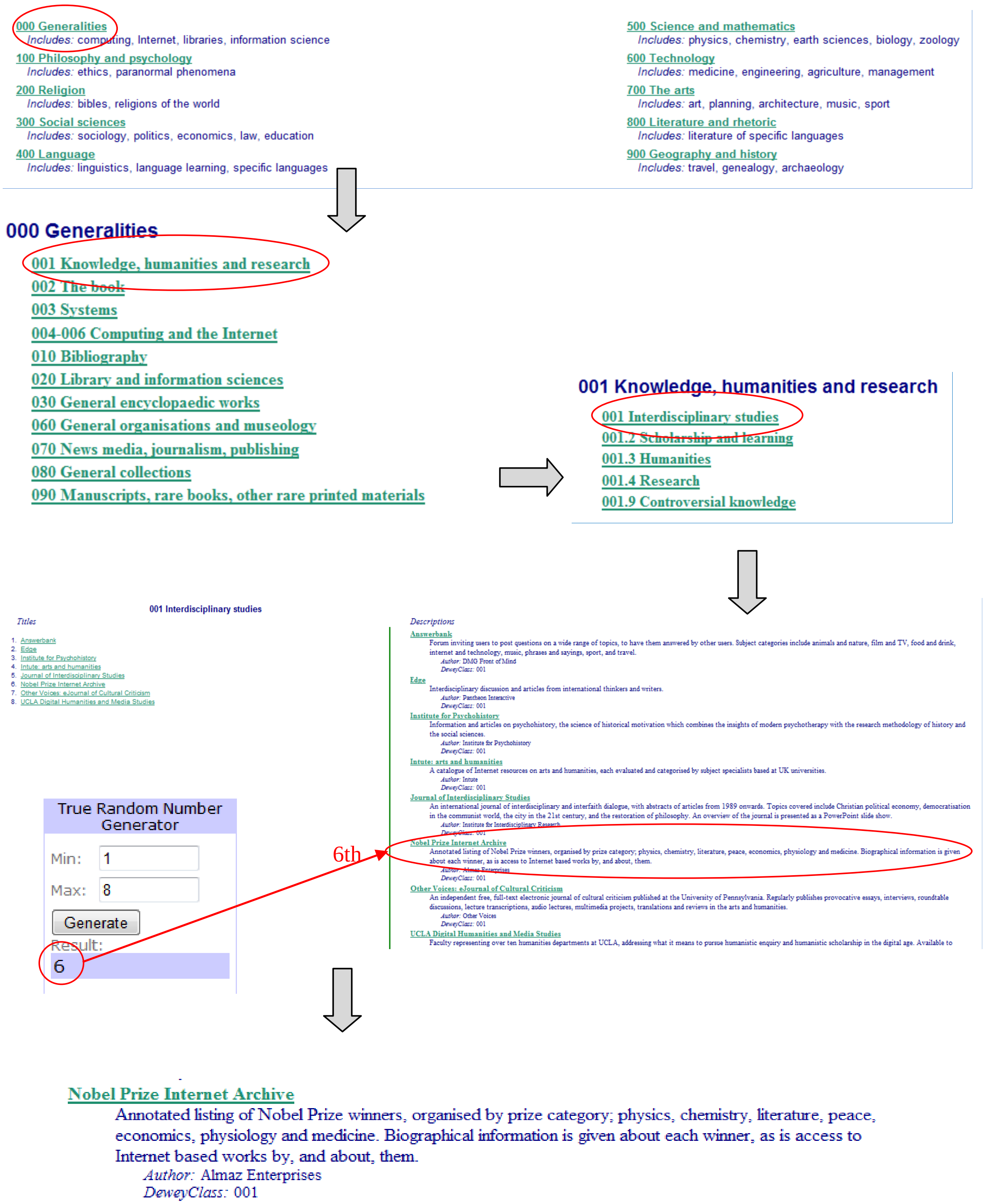
Table 4. BUBL subject categories

Top Categories	**Subjects covered**
000 Generalities	Computing, Internet, Libraries, Information Science
100 Philosophy and psychology	Ethics, Paranormal phenomena
200 Religion	Bibles, Religions of the world
300 Social sciences	Sociology, Politics, Economics, Law, Education
400 Language	Linguistics, Language learning, Specific languages
500 Science and mathematics	Physics, Chemistry, Earth Sciences, Biology, Zoology
600 Technology	Medicine, Engineering, Agriculture, Management
700 The arts	Art, Planning, Architecture, Music, Sport
800 Literature and rhetoric	Literature of specific languages
900 Geography and history	Travel, Genealogy, Archaeology

Based on 10 subject categorizations which BUBL provides as top-level categories, web documents were selected. Each top-level category is arranged by about 10 second level sub-categories, sometimes more than 10. For example, in the case of *700 The arts*, there are 22 sub-categories, so documents under sub-categories were randomly selected and searched in turn, and one document per sub-category was selected:

700 The arts
700 The arts: general resources
700 Fine and decorative arts: general resources
700 Fine and decorative arts: artists
701 Fine and decorative arts: philosophy
703 Fine and decorative arts: dictionaries and encyclopaedias
705 Fine and decorative arts: journals and magazines
706 Fine and decorative arts: organisations
707 Fine and decorative arts: art schools
708 Art galleries and museums in the UK
708 Art galleries and museums in the US
708 Art galleries and museums worldwide
709 History of art
710 Civic and landscape art
720 Architecture
730 Plastic arts, sculpture
740 Drawing and decorative arts
750 Painting and paintings
760 Graphic arts, printmaking and prints
770 Photography and photographs
780 Music
790 Recreational and performing arts
796 Sport and outdoor activities

This study collected Delicious's tags assigned to web documents which are also indexed at the other two web sites, BUBL and Intute. Only if a web document is found at all three locations (BUBL, Intute, and Delicious) were the tags assigned to the document at Delicious extracted. Regarding documents indexed at Intute, since index strings provided by Intute are necessary for comparing with BUBL and Intute's keywords are needed for comparing with tags, only when Intute offered both index strings and keywords to a web document was the web document selected for analysis. The process of random sampling a web document for collecting data is illustrated in Figure 7.



000 Generalities
Includes: computing, Internet, libraries, information science
100 Philosophy and psychology
Includes: ethics, paranormal phenomena
200 Religion
Includes: bibles, religions of the world
300 Social sciences
Includes: sociology, politics, economics, law, education
400 Language
Includes: linguistics, language learning, specific languages
500 Science and mathematics
Includes: physics, chemistry, earth sciences, biology, zoology
600 Technology
Includes: medicine, engineering, agriculture, management
700 The arts
Includes: art, planning, architecture, music, sport
800 Literature and rhetoric
Includes: literature of specific languages
900 Geography and history
Includes: travel, genealogy, archaeology

000 Generalities

001 Knowledge, humanities and research
002 The book
003 Systems
004-006 Computing and the Internet
010 Bibliography
020 Library and information sciences
030 General encyclopaedic works
060 General organisations and museology
070 News media, journalism, publishing
080 General collections
090 Manuscripts, rare books, other rare printed materials

001 Knowledge, humanities and research

001 Interdisciplinary studies
001.2 Scholarship and learning
001.3 Humanities
001.4 Research
001.9 Controversial knowledge

001 Interdisciplinary studies

Titles

1. Answerbank
2. Edge
3. Institute for Psychohistory
4. Intute: arts and humanities
5. Journal of Interdisciplinary Studies
6. Nobel Prize Internet Archive
7. Other Voices: eJournal of Cultural Criticism
8. UCLA Digital Humanities and Media Studies

Descriptions

Answerbank
Forum inviting users to post questions on a wide range of topics, to have them answered by other users. Subject categories include animals and nature, film and TV, food and drink, internet and technology, music, phrases and sayings, sport, and travel.
Author: DMG Front of Mind
DeweyClass: 001

Edge
Interdisciplinary discussion and articles from international thinkers and writers.
Author: Pantheon Interactive
DeweyClass: 001

Institute for Psychohistory
Information and articles on psychohistory, the science of historical motivation which combines the insights of modern psychotherapy with the research methodology of history and the social sciences.
Author: Institute for Psychohistory
DeweyClass: 001

Intute: arts and humanities
A catalogue of Internet resources on arts and humanities, each evaluated and categorised by subject specialists based at UK universities.
Author: Intute
DeweyClass: 001

Journal of Interdisciplinary Studies
An international journal of interdisciplinary and interfaith dialogue, with abstracts of articles from 1989 onwards. Topics covered include Christian political economy, democratisation in the communist world, the city in the 21st century, and the restoration of philosophy. An overview of the journal is presented as a PowerPoint slide show.
Author: Institute for Interdisciplinary Research
DeweyClass: 001

Nobel Prize Internet Archive
Annotated listing of Nobel Prize winners, organised by prize category; physics, chemistry, literature, peace, economics, physiology and medicine. Biographical information is given about each winner, as is access to Internet based works by, and about, them.
Author: Almaz Enterprises
DeweyClass: 001

Other Voices: eJournal of Cultural Criticism
An independent free, full-text electronic journal of cultural criticism published at the University of Pennsylvania. Regularly publishes provocative essays, interviews, roundtable discussions, lecture transcriptions, audio lectures, multimedia projects, translations and reviews in the arts and humanities.
Author: Other Voices
DeweyClass: 001

UCLA Digital Humanities and Media Studies
Faculty representing over ten humanities departments at UCLA, addressing what it means to pursue humanistic enquiry and humanistic scholarship in the digital age. Available to

True Random Number Generator

Min: 1
Max: 8
Result: 6

Nobel Prize Internet Archive
Annotated listing of Nobel Prize winners, organised by prize category; physics, chemistry, literature, peace, economics, physiology and medicine. Biographical information is given about each winner, as is access to Internet based works by, and about, them.
Author: Almaz Enterprises
DeweyClass: 001

Figure 7. The process of random sampling of a web document

The first step starts with looking at the category of *000 Generalities* which appears at the top category of BUBL. Under *000 Generalities*, the first sub-category is *001 Knowledge, humanities and research*. Clicking this leads to the list of all individual web documents covering the subject of *001 Interdisciplinary studies.* Under this sub-category, 8 web documents are presented in alphabetical order. In order to randomly select a document among 8 documents, the True Random Number Generator (www.random.org) was used. The True Random Number Generator produced number 6, and the 6th document, *Nobel Prize Internet Archive* is then searched at the other two locations, Intute and Delicious, to see if the document is indexed there. If the document searched was missing at either of these web sites, the next document was selected based on a new random number re-generated by the True Random Number Generator, and searched in the same way. The following category paths illustrate these steps:

```
000 Generalities > 001 Knowledge, humanities and research> 001
Interdisciplinary studies > Nobel Prize Internet Archive (a web document)
```

Table 5. The number of second-level categories at BUBL

Top Categories	**The number of second-level categories**	**The number of selected documents**
000 Generalities	11	8
100 Philosophy and psychology	14	6
200 Religion	24	12
300 Social sciences	13	12
400 Language	23	9
500 Science and mathematics	10	10
600 Technology	10	8
700 The arts	22	21
800 Literature and rhetoric	21	15
900 Geography and history	14	12
Total	**162**	**113**

Table 5 shows how many second level sub-categories exist per each top category. A total of 113 web documents were randomly selected for samples when choosing one document per sub-category (see Appendix A). As shown in Table 5, regarding three subject categories, Philosophy, Religion, and Language, the number of selected documents was smaller than expected. It turned out that second-level categories under those subject categories were too specific to yield web sites also covered by Intute and Delicious (see Table 6).

Table 6. Subject categories including very specific topics

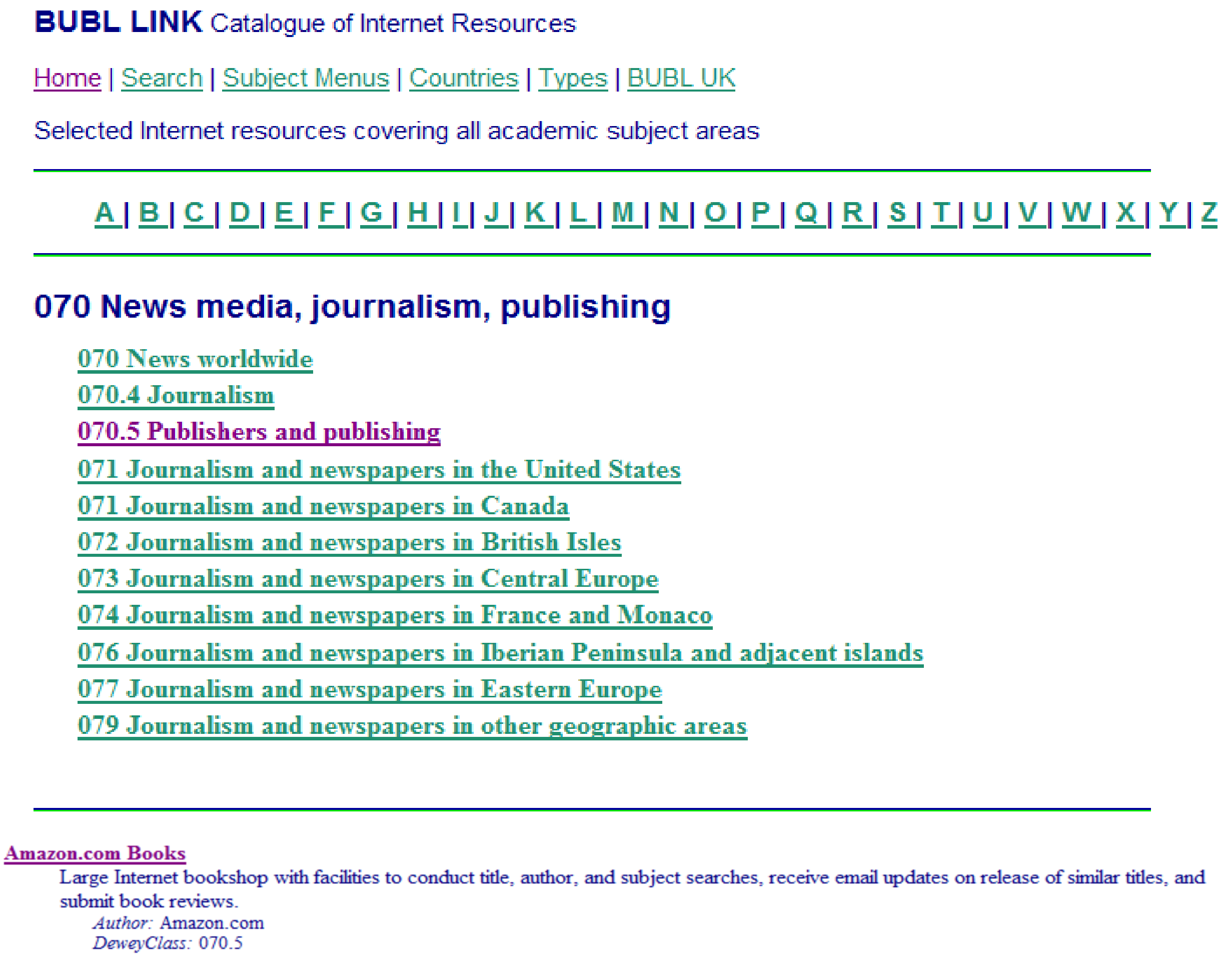
Subject categories	**Examples of second-level categories**
Philosophy	100 Philosophy: societies, 107 Philosophy education, 110 Metaphysics, 120 Epistemology, causation, humankind, 130 Paranormal phenomena…
Religion	234.161 Baptism, 238 Christian creeds and catechisms, 252 Texts of sermons, 268 Christian education…
Language	439.31 Dutch language, 439.7 Swedish language, 439.82 Norwegian language, 450 Italian language, 459 Romanian language, 460 Spanish language, 469 Portuguese language, 470 Latin language, 480 Greek language…

3.2.4 Collection of Delicious tags

A JAVA –based program was written for tag collection, tag pre-processing, and the inter-indexer consistency computation. Through the Delicious API, the program collected tags in a JSON (JavaScript Object Notation) format (Crockford, 2006) (see Appendix B). For the period from February to March in 2010, 31,319 Delicious tags in 113 web documents were collected for analysis.

3.2.5 Collection of BUBL and Intute index terms

In order to measure indexing consistency and to examine different points of view on the same document between professional indexers, indexers' index terms from BUBL and Intute were analyzed. BUBL assigns each document a classification number based on DDC as shown in Figure 8.


BUBL LINK Catalogue of Internet Resources

Home | Search | Subject Menus | Countries | Types | BUBL UK

Selected Internet resources covering all academic subject areas

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

070 News media, journalism, publishing

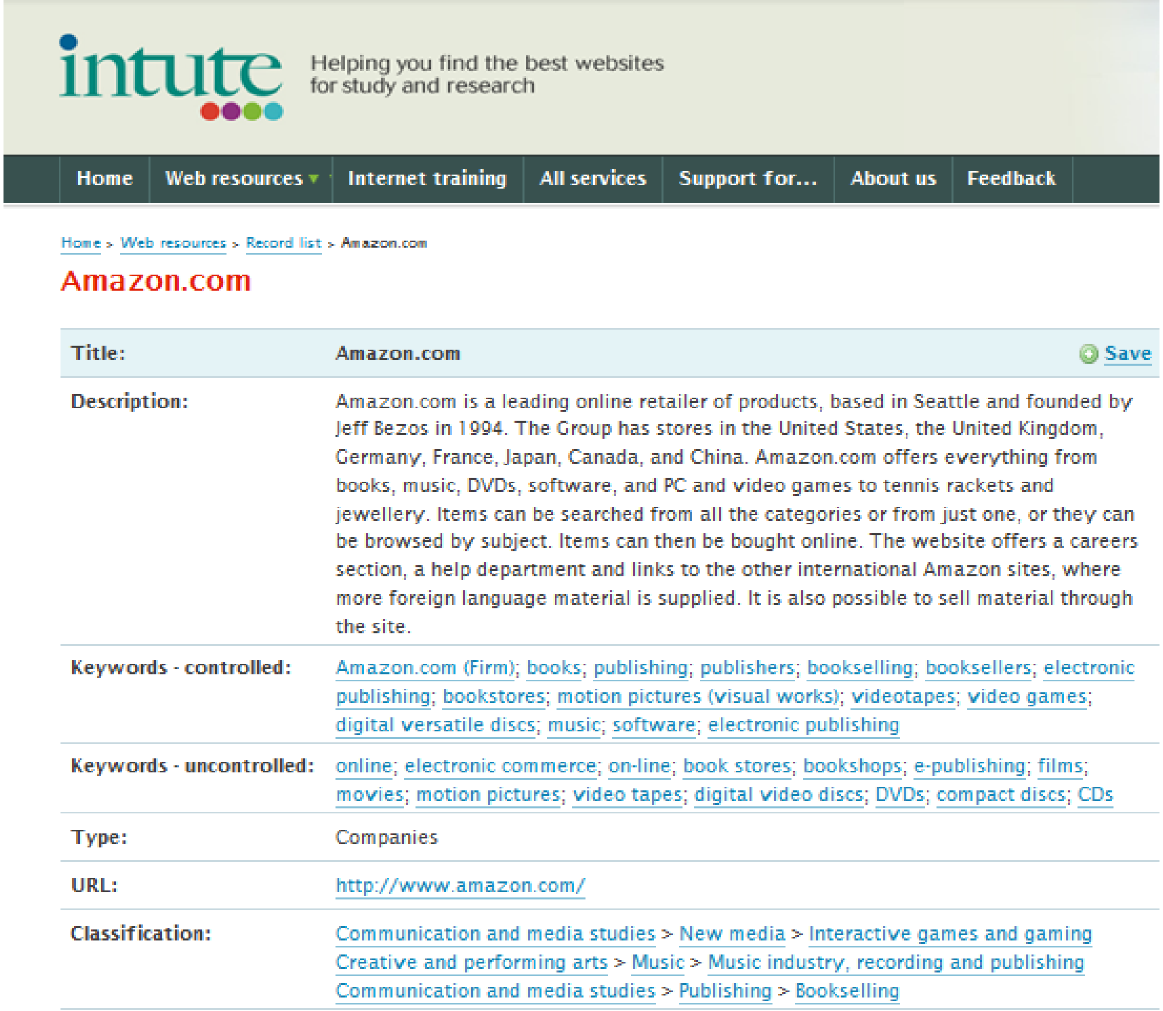
- 070 News worldwide
- 070.4 Journalism
- 070.5 Publishers and publishing
- 071 Journalism and newspapers in the United States
- 071 Journalism and newspapers in Canada
- 072 Journalism and newspapers in British Isles
- 073 Journalism and newspapers in Central Europe
- 074 Journalism and newspapers in France and Monaco
- 076 Journalism and newspapers in Iberian Peninsula and adjacent islands
- 077 Journalism and newspapers in Eastern Europe
- 079 Journalism and newspapers in other geographic areas

Amazon.com Books
Large Internet bookshop with facilities to conduct title, author, and subject searches, receive email updates on release of similar titles, and submit book reviews.
Author: Amazon.com
DeweyClass: 070.5

Figure 8. Amazon.com indexed at BUBL

On the other hand, Intute provides index strings based on classification schemes and sometimes it provides keywords (controlled or uncontrolled or both) generated by professional indexers (see

Figure 9).


intute
Helping you find the best websites for study and research
Home
Web resources
Internet training
All services
Support for...
About us
Feedback
Home > Web resources > Record list > Amazon.com
Save

Amazon.com

Title:	Amazon.com
Description:	Amazon.com is a leading online retailer of products, based in Seattle and founded by Jeff Bezos in 1994. The Group has stores in the United States, the United Kingdom, Germany, France, Japan, Canada, and China. Amazon.com offers everything from books, music, DVDs, software, and PC and video games to tennis rackets and jewellery. Items can be searched from all the categories or from just one, or they can be browsed by subject. Items can then be bought online. The website offers a careers section, a help department and links to the other international Amazon sites, where more foreign language material is supplied. It is also possible to sell material through the site.
Keywords - controlled:	Amazon.com (Firm); books; publishing; publishers; bookselling; booksellers; electronic publishing; bookstores; motion pictures (visual works); videotapes; video games; digital versatile discs; music; software; electronic publishing
Keywords - uncontrolled:	online; electronic commerce; on-line; book stores; bookshops; e-publishing; films; movies; motion pictures; video tapes; digital video discs; DVDs; compact discs; CDs
Type:	Companies
URL:	http://www.amazon.com/
Classification:	Communication and media studies > New media > Interactive games and gaming Creative and performing arts > Music > Music industry, recording and publishing Communication and media studies > Publishing > Bookselling

Figure 9. Amazon.com indexed at Intute

For indexer's index terms from BUBL, this study analyzed index strings which are category paths of classification (see Figure 8). The category paths were collected from top-level categories. Among top-level categories, only the term *Generalities* was excluded for comparison with Intute's category paths because Intute's top categories do not have an equivalent for the term *Generalities* as shown below:

Agriculture, food, and forestry
Architecture and planning
Biological sciences
Business and management studies
Communication and media studies
Creative and performing arts
Education and research methods
Engineering
Geography and environment
Humanities
Law
Mathematics and computer science
Medicine including dentistry
Modern languages and area studies
Nursing, midwifery and allied health
Physical sciences
Psychology
Social sciences
Veterinary medicine

Thus, for example, regarding a document, *Amazon.com*, the following paths were recognized and analyzed:

```
070 News media, journalism, publishing > 070.5 Publishers and
publishing > 070.5 Booksellers and bookshops
```

The collection of indexers' index terms from Intute was the same as BUBL. For a more accurate comparison between Intute and BUBL, at this point only index strings of category paths in classification schemes were analyzed:

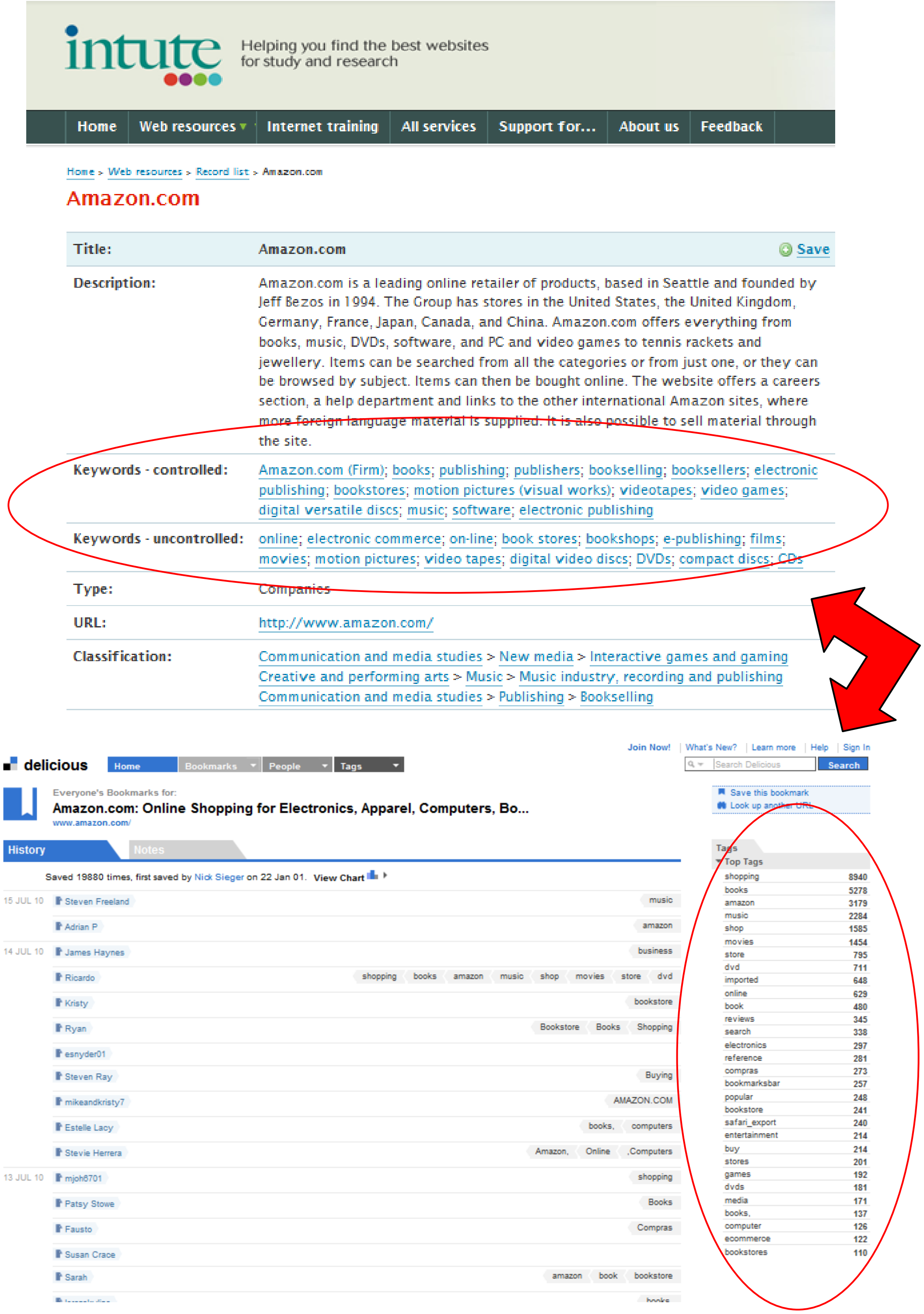
```
Communication and Media Studies > New Media > Interactive Games and
Gaming

Creative and Performing Arts > Music > Music Industry, Recording and
Publishing
Communication and Media Studies > Publishing > Bookselling
```

Regarding the analysis of terms in category paths, the rules for vocabulary analysis (see Section 3.3.1.1.) were applied for pre-processing.

3.2.6 Collection of Intute keywords and Delicious top 20 tags

To measure indexing consistency of social tagging users at Delicious compared with that of professional indexers of Intute, Intute keywords were compared with Delicious. The top ranked tags assigned to a document at Delicious were collected and normalized through checking spelling and word forms as described in Section 3.3.1.1 *Rules for vocabulary analysis*. The top 20 tags were compared with keywords (controlled or uncontrolled) from Intute (Figure 10). The keywords provided by Intute are useful and the most appropriate data in order to compare the professional indexer's point of view with the user's point of view in subject indexing on the same document.



Amazon.com

Title:	Amazon.com
Description:	Amazon.com is a leading online retailer of products, based in Seattle and founded by Jeff Bezos in 1994. The Group has stores in the United States, the United Kingdom, Germany, France, Japan, Canada, and China. Amazon.com offers everything from books, music, DVDs, software, and PC and video games to tennis rackets and jewellery. Items can be searched from all the categories or from just one, or they can be browsed by subject. Items can then be bought online. The website offers a careers section, a help department and links to the other international Amazon sites, where more foreign language material is supplied. It is also possible to sell material through the site.
Keywords - controlled:	Amazon.com (Firm); books; publishing; publishers; bookselling; booksellers; electronic publishing; bookstores; motion pictures (visual works); videotapes; video games; digital versatile discs; music; software; electronic publishing
Keywords - uncontrolled:	online; electronic commerce; on-line; book stores; bookshops; e-publishing; films; movies; motion pictures; video tapes; digital video discs; DVDs; compact discs; CDs
Type:	Companies
URL:	http://www.amazon.com/
Classification:	Communication and media studies > New media > Interactive games and gaming Creative and performing arts > Music > Music industry, recording and publishing Communication and media studies > Publishing > Bookselling

Everyone's Bookmarks for:

Amazon.com: Online Shopping for Electronics, Apparel, Computers, Bo...

www.amazon.com/

Saved 19880 times, first saved by Nick Sieger on 22 Jan 01.

15 JUL 10	Steven Freeland	music
	Adrian P	amazon
14 JUL 10	James Haynes	business
	Ricardo	shopping books amazon music shop movies store dvd
	Kristy	bookstore
	Ryan	Bookstore Books Shopping
	esnyder01	
	Steven Ray	Buying
	mikeandkristy7	AMAZON.COM
	Estelle Lacy	books, computers
	Stevie Herrera	Amazon, Online ,Computers
13 JUL 10	mjoh6701	shopping
	Patsy Stowe	Books
	Fausto	Compras
	Susan Crace	
	Sarah	amazon book bookstore

Tags

Top Tags

shopping	8940
books	5278
amazon	3179
music	2284
shop	1585
movies	1454
store	795
dvd	711
imported	648
online	629
book	480
reviews	345
search	338
electronics	297
reference	281
compras	273
bookmarksbar	257
popular	248
bookstore	241
safari_export	240
entertainment	214
buy	214
stores	201
games	192
dvds	181
media	171
books,	137
computer	126
ecommerce	122
bookstores	110

Figure 10. Intute keywords vs. Delicious top 20 tags

Table 7. Intute keywords vs. Delicious top 20 tags

Document	Keywords at Intute		Top 20 Tags at Delicious
Amazon.com	Keywords - controlled	Amazon.com (Firm); books; publishing; publishers; bookselling; booksellers; electronic publishing; bookstores; motion pictures (visual works); videotapes; video games; digital versatile discs; music; software	shopping, books, amazon, music, shop, movies, store, dvd, imported, online, book, reviews, search, electronics, reference, bookmarksbar, popular, bookstore, safari_export, entertainment
	Keywords - uncontrolled	online; electronic commerce; on-line; book stores; bookshops; e-publishing; films; movies; motion pictures; video tapes; digital video discs; DVDs; compact discs; CDs	

Basically, “controlled keywords” of Intute were compared with tags, and “uncontrolled keywords” were analyzed for the comparison only when “controlled keywords” were not available.

3.3 Data Analysis

The analysis of collected data is divided into three phases: analysis of indexing consistency, analysis of tagging effectiveness, and analysis of tagging attributes and tagging behaviors. To carry out the analyses, the process of data pre-processing on tags and index strings has been conducted.

3.3.1 Data pre-processing

Data pre-processing was conducted for the collected tags to exclude taggers who added non-English tags or no tags. The collected tags were checked for spelling, acronyms or singular and plural forms. That is, this step included removing misspelled terms and integrating terms which have different forms of words such as noun, adjective, adverb, and gerund.

3.3.1.1 An exact match between terms

Based on discussion by Lancaster and Smith (1983), this study used the following five rules for specifying an exact match between two terms.

- Exactly corresponding including singular/plural variations

  Ex) aurora to auroras, language to languages
- Variant spellings

  Ex) organization to organisation
- Word forms (adjectival, noun, or verbal forms)

  Ex) medicine to medical
- Acronyms or abbreviations and full terms

  Ex) National Center for Biotechnology Information to NCBI, biotechnology to biotech
- Compound terms

  human/body to humanbody to human_body to human, body etc.

In terms of tags, Delicious does not have the feature of adding a space between two terms for a compound term, so if there is a dash, slash, or underscore between two terms, or if two terms are found at the same time in the list of tags from a tagger, they were regarded as a compound term.

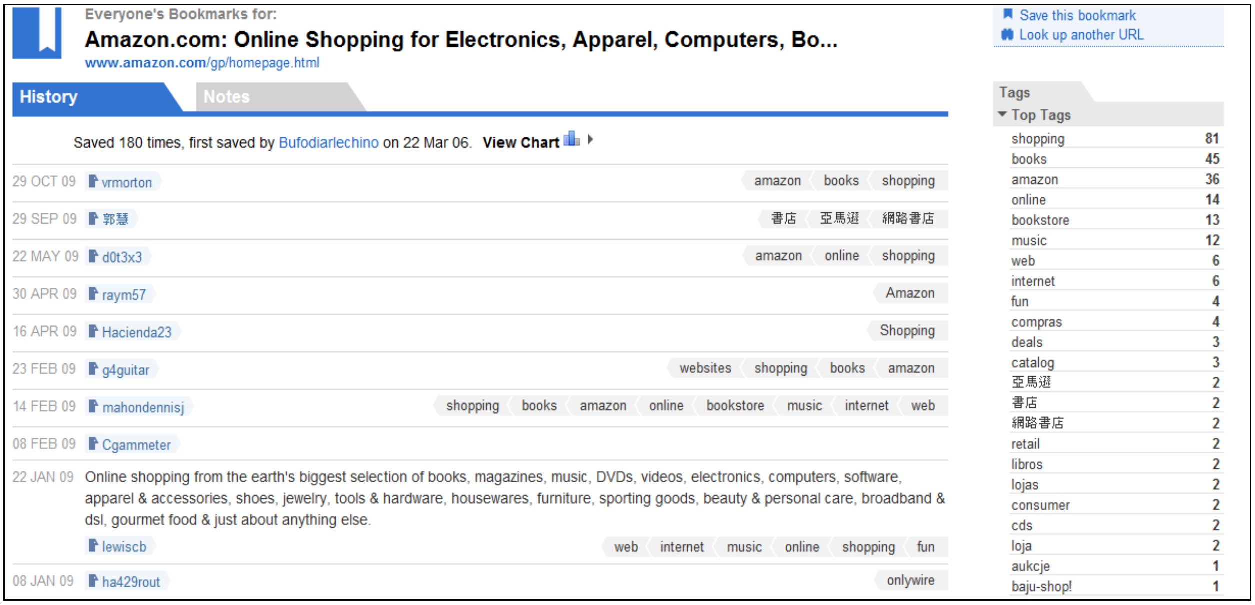
Regarding terms in category paths, two words having a space between them were regarded as a compound term, for example, “News media”, “Music Industry”, or “Interactive Games”.

Furthermore, if there is a connective like “And” among term strings in category paths, words before and after the connective “And” were processed separately. For instance, if a category path is “mathematics and computer science”, terms to be processed for analysis are “mathematics” and “computer science”.

The dragon toolkit (Zhou et al., 2007), which is a WordNet (http://wordnet.princeton.edu/) based lemmatization tool, was used for checking for English words and stemming which is for merging inflected forms of indexing words. Acronyms were checked in the *Acronyms, initialisms & abbreviations dictionary* (Reade and Romaniuk, 2005).

3.3.1.2 Term exclusion

Since users at Delicious are drawn from a worldwide audience, they might have different language backgrounds. For instance, Figure 11 shows that a document, *Amazon.com* is tagged in several languages. Thus, if assigned tags are not in English (e.g., in Spanish, Korean, Chinese etc.), they are excluded from the analysis.



Everyone's Bookmarks for:

Amazon.com: Online Shopping for Electronics, Apparel, Computers, Bo...

www.amazon.com/gp/homepage.html

Saved 180 times, first saved by Bufodiarlechino on 22 Mar 06. View Chart

29 OCT 09 vrmorton — amazon books shopping

29 SEP 09 郭慧 — 書店 亞馬遜 網路書店

22 MAY 09 d0t3x3 — amazon online shopping

30 APR 09 raym57 — Amazon

16 APR 09 Hacienda23 — Shopping

23 FEB 09 g4guitar — websites shopping books amazon

14 FEB 09 mahondennisj — shopping books amazon online bookstore music internet web

08 FEB 09 Cgammeter

22 JAN 09 Online shopping from the earth's biggest selection of books, magazines, music, DVDs, videos, electronics, computers, software, apparel & accessories, shoes, jewelry, tools & hardware, housewares, furniture, sporting goods, beauty & personal care, broadband & dsl, gourmet food & just about anything else.

lewiscb — web internet music online shopping fun

08 JAN 09 ha429rout — onlywire

Tags

Top Tags

shopping	81
books	45
amazon	36
online	14
bookstore	13
music	12
web	6
internet	6
fun	4
compras	4
deals	3
catalog	3
亞馬遜	2
書店	2
網路書店	2
retail	2
libros	2
lojas	2
consumer	2
cds	2
loja	2
aukcje	1
baju-shop!	1

Figure 11. An example of terms tagged in several languages

For the comparison of tags and professionals' indexing terms, this research developed a stoplist or a list of terms which can be excluded for processing (see Appendix C). All tags were checked against the stoplist. The stoplist included an explicit list of the terms that Sen et al. (2006) define as subjective and personal tags (see Table 8), since those types of tags are not meaningful for indexing subjects of documents. Table 8 provides the three types of tags and their definitions from Sen et al. and the related examples of tags identified.

Table 8. Sen et al. three types of tags

Types of tags	**Definitions**	**Examples of identified tags**
Factual tags	"identifies facts about" a resource e.g., people, places, or concepts	government, socialsecurity, finance etc.
Subjective tags	"express user opinions" related to a resource	good, worth, recommend, toRead, informative etc.
Personal tags	having "intended audience of tag applied themselves"	myDaughter, forSon, etc.

3.3.2 Analysis of Inter-indexer/tagger consistency

As a first phase of analysis, this dissertation research conducted an analysis of indexing consistency. This research employed Vector Space Model (VSM) based similarity measures. The vector model was discussed by Salton and colleagues in 1975. In the VSM, documents and queries are represented as vectors in the term space, and the documents are ranked by closeness to the query. Figure 12 shows a typical three-dimensional index space where each item is identified by up to three distinct terms (Salton et al., 1975a).

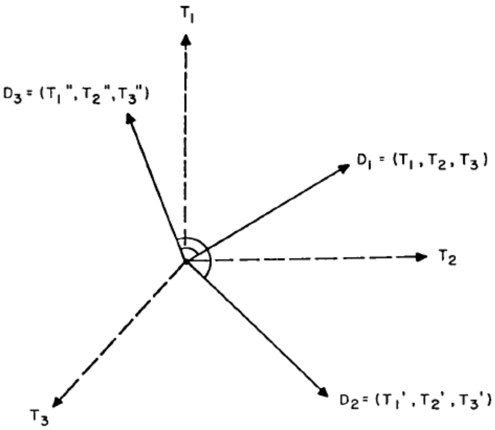


Figure 12. Vector representation of document space (Source: Salton et al., 1975a)

The three-dimensional space may be extended to *n* dimensions when *n* different index terms are present. A document matrix *V* for a document set consisting of *m* documents and *n* terms is as follows:

$$V = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix}$$

Figure 13. Document matrix V for a document (Source: Olson and Wolfram, 2006)

Vector representation does not consider the ordering of terms in a document. A document is represented by a bag of words where ordering is not significant, so it is called the bag of words model. For instance, the query "*white house rose garden*" is considered the same as the query "*white rose house garden*" even though they are not exactly the same query (Metzler and Croft, 2006). However, there are some issues with the vector space model. It assumes that terms are independent and there is an independence relationship among them. Furthermore, vector operations are not formal, e.g., weighting terms draws on heuristics (Wong et al., 1985).

Wolfram and Olson (2007) applied the concept of document space in the vector space model into the terms assigned by a group of indexers to a document, and defined an Indexer/Tagger Space. Wolfram and Olson calculated the distance between each indexer/tagger's resulting vector and the indexing centroid (or average vector across all indexers/taggers). In their measurement, high density space among indexers/taggers means more similarity and higher consistency (Figure 14):

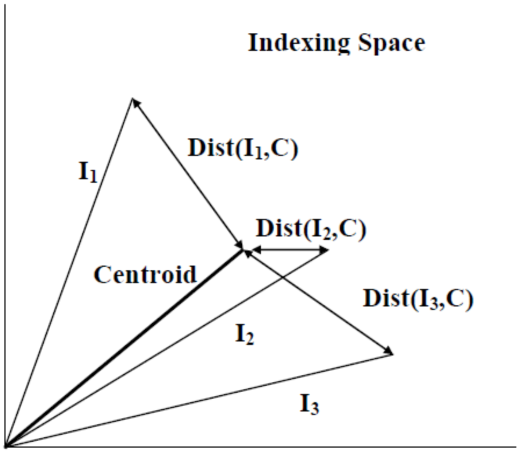


Figure 14. Indexer distances from the indexing centroid (Source: Wolfram and Olson, 2007)

Wolfram and Olson (2007)'s Inter-Indexer (Tagger) Consistency Density is calculated as follows:

$$ICD = \frac{\sum_{i=1}^{m} Sim(I_i, C)}{m}$$

Where *m* is the number of indexers/taggers, C denotes Centroid and I_i is an indexer vector.

$$Sim(I_i, C) = \begin{cases} \dfrac{1}{Dist(I_i, C)} & for\ \ Dist(I_i, C) \neq 0 \\ 1 & for\ \ Dist(I_i, C) = 0 \end{cases}$$

However, the formula for this similarity measure is rather problematic. When $Dist(I_i, C)$ is between 0 and 1, the value of $\dfrac{1}{Dist(I_i, C)}$ will be greater than 1 (see Figure 15).

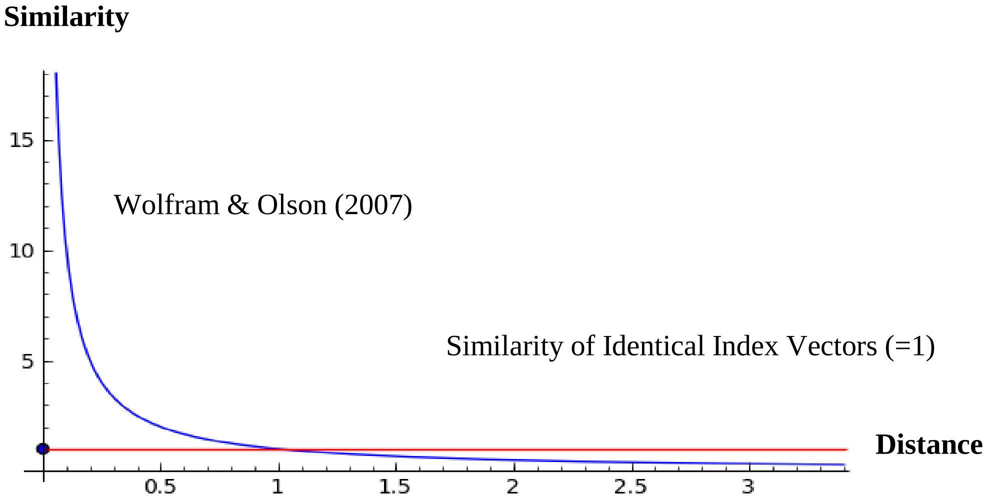


Figure 15. A problem with the existing formula

It explains that the similarity between two different vectors would be greater than the similarity between two identical vectors where a distance between two indexer vectors is 0, i.e., $Dist(I_i, C) = 0$. This leads to misinformation that the similarity where two indexers assigned the same index terms would be smaller than the similarity where two indexers assigned different index terms. Accordingly, this dissertation research adapted Wolfram and Olson's formula with different VSM based measures. To produce a more convincing and valuable analysis and to decrease possible bias by each measure, the following three different similarity measures were applied to the modified ICD.

1) Adjusted Euclidean distance metric
2) Dot product similarity
3) Cosine similarity

1) Adjusted Euclidean distance metric

The similarity as measured by the Euclidean distance metric (Kohonen, 1995) is inversely proportional to the Euclidean distance. Thus, sign minus one (-1) is put in front of the formula to make this metric proportional to the similarity:

$$Similarity(I_i, C) = -Dist(I_i, C)$$

This is equal to:

$-Dist(I_i, C) = -|C - I_i|$

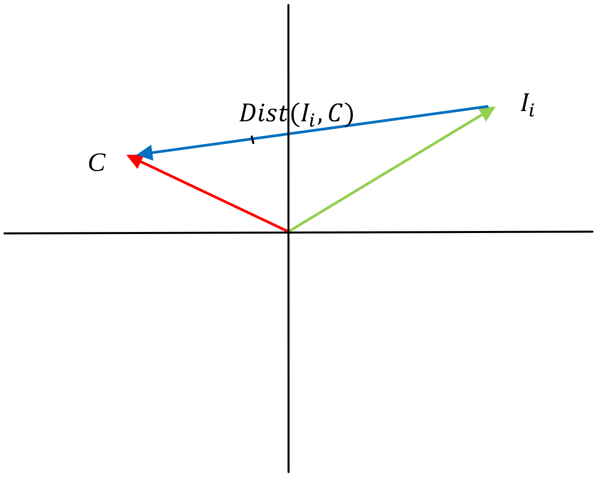


Figure 16. Euclidean distance metric

In this measure, similarity is not monotonic for |*Indexer vector* |. That is, for a given angle θ, similarity ($-Dist(I_i, C)$) decreases as |*Indexer vector* | increases when |*Indexer vector* | is bigger than |*Centroid* | *cos θ*. However, as Figure 17 shows, similarity also decreases as |*Indexer vector*| decreases when |*Indexer vector* | is smaller than |*Centroid* | *cos θ*.

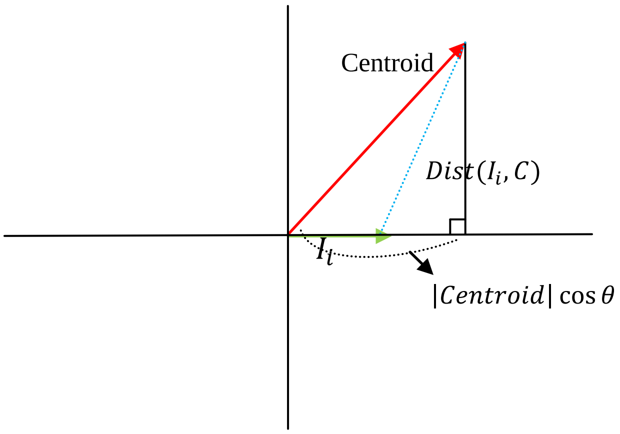


Figure 17. |*Indexer vector*| and Euclidean distance metric

2) Dot product similarity

Dot product based similarity is represented by:

$$Similarity\ (I_i, C) = I_i \cdot C$$

Dot product can be defined as:

$$I_i \cdot C = |I_i||C| \cos\theta$$

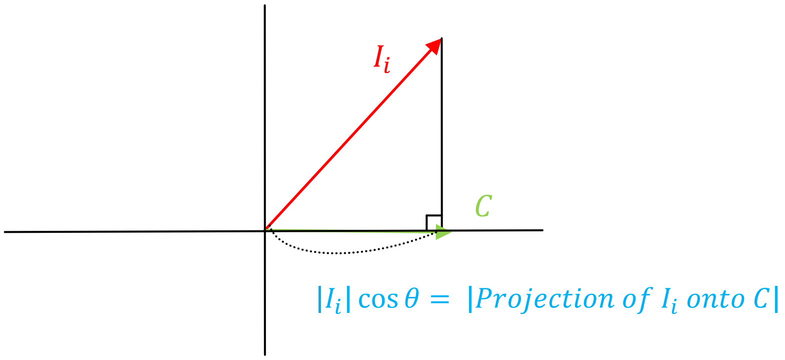


Figure 18. Dot product

As Figure 18 shows, Dot product also could be regarded as:

$$\left|I_i\right| * \left|Projection\ of C\ onto\ I_i\right| = |C| * \left|Projection\ of\ I_i\ onto\ C\right| = Constant * \left|I_i\right| \cos\theta$$

Since $\left|Projection\ of\ I_i\ onto\ C\right|$ is proportional to $\cos\theta$, the dot product will be bigger as the magnitude ($\left|I_i\right|$) of the indexer vector is bigger. That is, Dot product is proportional to not only $\cos\theta$ but also the magnitude of the indexer vector.

3) Cosine similarity

Cosine similarity is measured by the cosine of the angle between two vectors of the same dimensions. The cosine similarity (θ) is represented using a dot product and magnitude as:

$$Sim(I_i, C) = \frac{I_i \cdot C}{|I_i||C|}$$

$|I_i|$ = vector norm of I_i
|C| = vector norm of vector Centroid
θ = angle between vector I_i and vector C

where I_i and C are two vectors of attributes and $I_i \cdot$ C is the dot product of vectors (see Figure 19).

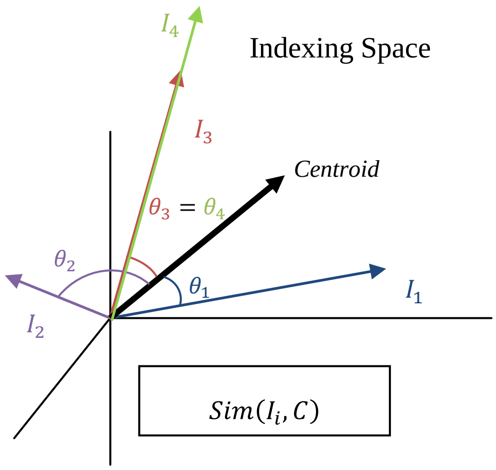


Figure 19. Cosine similarity

In cosine similarity, magnitude is not considered. That is, in Figure 19, when $\theta_3\ and\ \theta_4$ are the same angle, $Sim(I_3, C)$ is equal to $Sim(I_4, C)$ while $Dist(I_3, C)$ is different from $Dist(I_4, C)$.

With three similarity measures discussed above, in this research, indexing consistency is measured for the following:

(1) Inter-indexer/tagger consistency among Delicious users regarding subject areas

(2) Inter-indexer consistency between two groups of professional indexers from BUBL and Intute subject gateways

(3) Inter-indexer consistency between Delicious's users and Intute's professional indexers

(1) *Inter-indexer/tagger consistency among Delicious users*

First, inter-indexer consistency was measured on tags from more than 50 up to 100 taggers who assigned tags most recently since Delicious feeds up to 100 most recent bookmarks.

Table 9. Example of indexing space on a web document: Amazon.com

Tags	**Tagger 1**	**Tagger 2**	**Tagger 3**	**Tagger 4**	**Tagger 5**	**Tagger 6**	**Centroid**
shopping/shop	1	1	0	1	1	1	**5/6**
Books	1	0	1	0	1	1	**4/6**
Amazon	0	1	0	1	0	0	**2/6**
Music	0	1	0	1	0	0	**2/6**
Movies	0	0	0	0	0	1	**1/6**

Table 9 illustrates that regarding a document, *Amazon.com*, most of the taggers including tagger 1, tagger 2, tagger 4, tagger 5 and tagger 6 have assigned a tag, "shopping" or "shop" as an index term to the document. The number "1" and "0" will be respectively assigned as a result of the vocabulary analysis process. For instance, the number "1" means that a tagger assigned the tag, and the number "0" means that the tagger did not add the term as a tag. Table 9 also shows the values of the centroid for tags when 6 taggers assigned tags to the document.

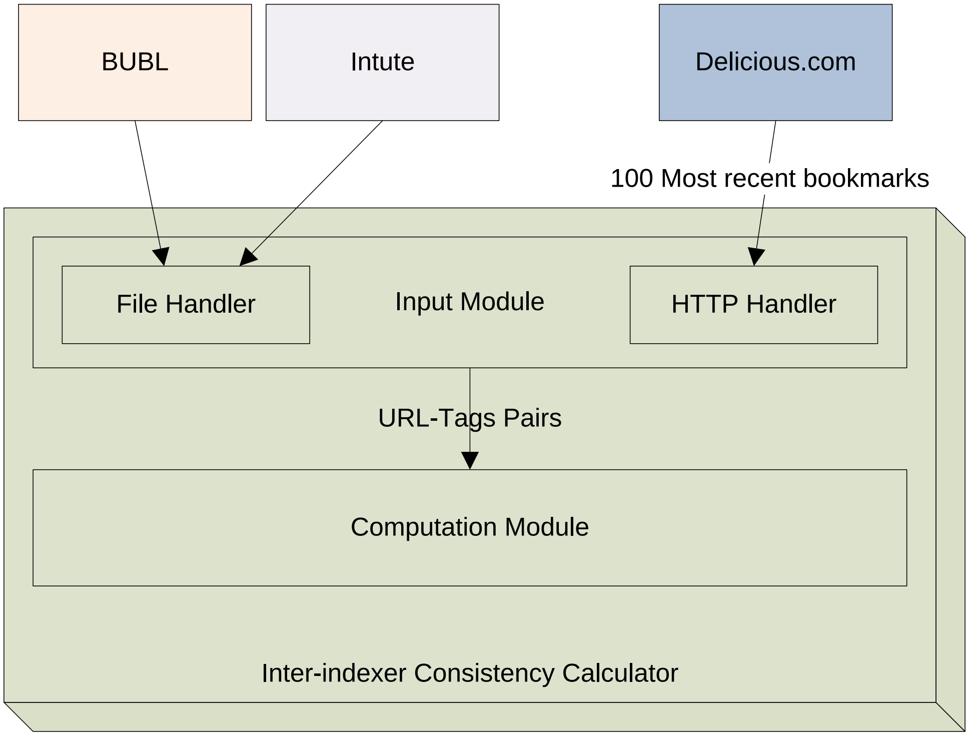


Figure 20. Overview of inter-indexer consistency calculator

All of these processes were automatically carried out with a JAVA-based program (see Figure 20). More details on the program are presented in Appendix B.

(2) *Inter-indexer consistency between two groups of professional indexers from BUBL and Intute subject gateways*

Indexing consistency between two groups, BUBL and Intute was measured with the aforementioned three similarity measures in order to compare with indexing tendency demonstrated in Delicious in terms of the 10 subject categories.

(3) *Inter-indexer consistency between Delicious's users and Intute's professional indexers*

The third indexing consistency is measured between Delicious's users and Intute's professional indexers.

It should be noted that ICD would not work properly for measuring indexing consistency between a small number of indexers since it is an average similarity between an index vector and the Centroid. For instance, even when there is no common index term between indexer 1 and indexer 2, ICD would generate a relatively large value (Table 10).

Table 10. Indexing space[4] on a document, "*Amazon.com*"

Tags	**Indexer 1**	**Indexer 2**	**Centroid**
Shop	1	0	**1/2**
Books	1	0	**1/2**
Amazon	0	1	**1/2**
Movies	0	1	**1/2**
ICD: 0.707			

So, indexing consistency between two groups of people, e.g., indexer 1 (A) and indexer 2 (B) —— in this research, between BUBL and Intute and between Delicious and Intute —— were measured without calculating the Centroid:

Euclidean distance:

$$Similarity\ (A, B) = -Dist(A, B)$$

Dot product:

$$Similarity\ (A, B) = A \cdot B$$

[4] The number "1" means that a tagger assigned the tag, and the number "0" means that a tagger did not add the term as a tag.

Cosine θ:

$$Cos\ \theta =$$

$$Similarity\ (A,B) = \frac{A \cdot B}{|A||B|}$$

$|A|$ = *vector norm of* A

$|B|$ = *vector norm of* B

θ = *angle between vector* A *and vector* B

To compare average Inter-Indexer (Tagger) Consistency Density among different topic areas, an analysis of variance (ANOVA) and Kruskal–Wallis one-way analysis of variance are used. The ANOVA method is for simultaneously comparing means of several groups, and originally it was developed by R.A. Fisher for data from agriculture experiments (Agresti and Finlay, 1999). If there is only one factor, a one-way ANOVA is used. In this analysis, there is only one factor (subject areas), so a one-way ANOVA is used. In ANOVA, we assume that the distribution of each group should be normally distributed. In the Kruskal-Wallis test, however, we do not make any assumption about the distribution. So the Kruskal-Wallis test is a distribution-free test.

The null hypothesis and alternative hypothesis for both ANOVA and Kruskal-Wallis test are as follows:

ANOVA: The null hypothesis: there is no difference in the average indexing similarity among 10 different subject areas. H_o: $\mu_1 = \mu_2 = ... = \mu_{10}$
The alternative hypothesis: the average indexing similarity for the 10 different subject areas are not the same. At least one pair of averages is different. $H_1: \ \mu_i \neq \mu_j$

Kruskal-Wallis test:

The null hypothesis: 10 different subject areas have the same distribution.

H_o: $\mu_1 = \mu_2 = ... = \mu_{10}$

The alternative hypothesis: at least one of the subject areas tends to yield larger values than at least one of the other subject areas.

$H_1: \mu_i \neq \mu_j$

3.3.3 Analysis of tagging effectiveness

3.3.3.1 Definitions of terms

This research has conducted an analysis of tagging effectiveness. The examination of social tagging considering both exhaustivity and specificity could ameliorate difficulties associated with limitations in the analysis of indexing consistency based on only the quantitative measures of vocabulary matching.

In terms of indexing effectiveness, there are two important notions of *exhaustivity* and *specificity* which are parameters to measure the effectiveness of indexing. *Exhaustivity* is defined as the number of different topics indexed for a document (Keen and Digger, 1972). Olson and Boll (2001) explain exhaustivity as “the number of concepts represented in the bibliographic record or the breadth of subject matter covered”. *Specificity* is related to the ability of the index terms to precisely describe the topics of a document (Keen and Digger, 1972). Olson and Boll (2001) describe specificity by dividing it into three factors:

- “The specificity and coextensiveness of the vocabulary”: the level of detail of the terminology in a vocabulary in hierarchical terms.

- “The specificity of its application”: the level of detail with which the vocabulary is applied.
- “The term specificity in the context of a given catalog”: how well a heading from a controlled vocabulary differentiates between topics in a particular catalog.

Concerning retrieval of documents, it has been well known that if the level of exhaustivity is higher, that is, all topics are indexed for a document, recall is higher. On the other hand, if the level of exhaustivity is lower, that is, if some of the topics are not indexed for a document, it results in lower recall. In the case of indexing specificity, high specificity leads to high precision (van Rijsbergen, 1979).

It has been known that it is not easy to quantify the levels of exhaustivity and specificity of indexing, but van Rijsbergen (1979) claims that it is important to be able to quantify these two notions since they have predictable effects on retrieval effectiveness. In terms of tagging effectiveness, tagging exhaustivity and tag specificity can be redefined: the number of tags assigned to one document and, the number of documents described by one tag (Spärck Jones, 1972; cited by Hassan-Montero and Herrero-Solana, 2006).

Thus, for measuring tagging exhaustivity, the average number of tags per document was calculated for the 10 subject categories which BUBL provides. For measuring tag specificity, we calculate the number of documents associated with one tag which is one of the tags listed among the top ranked tags (up to 20) in Delicious.

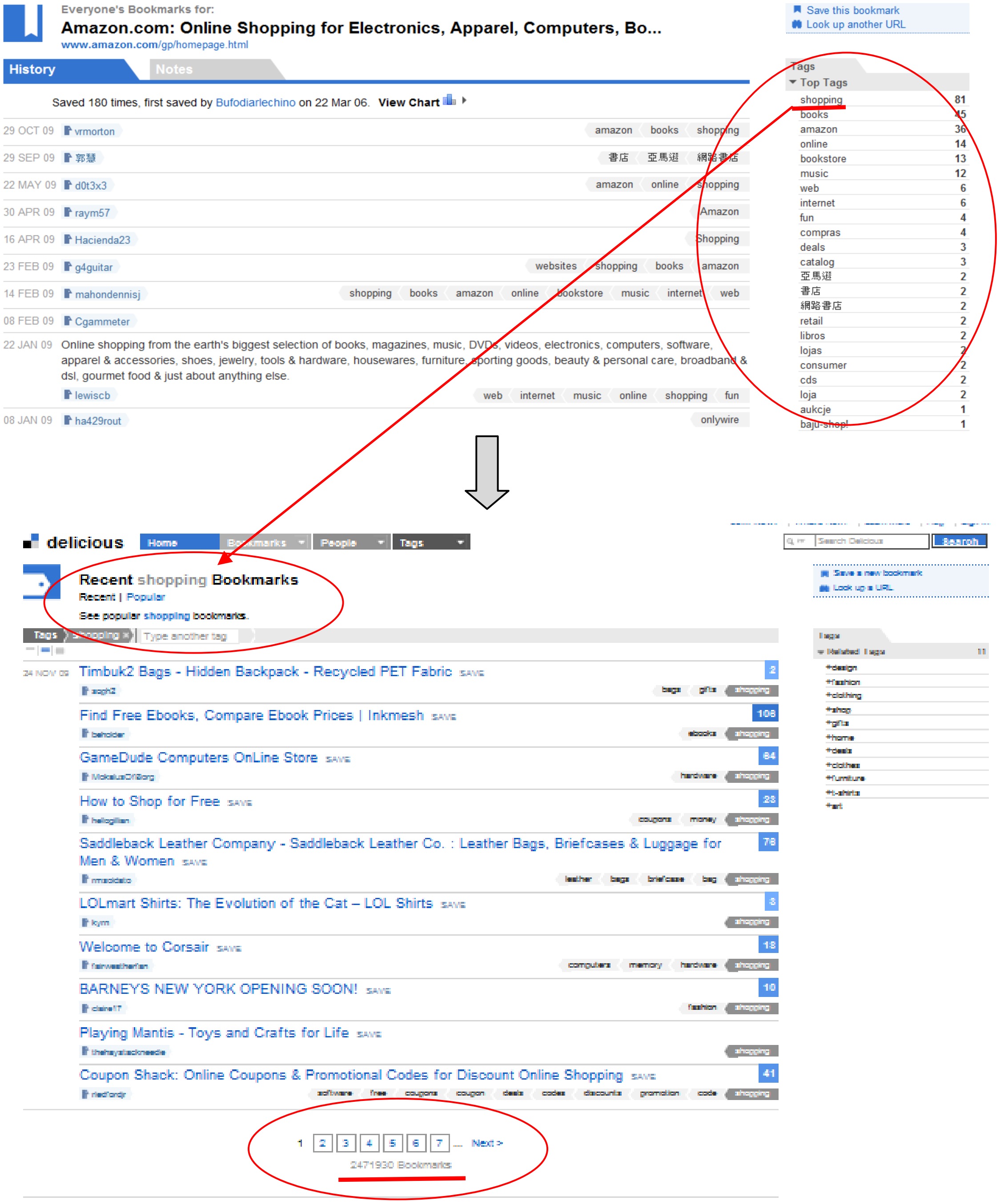
3.3.3.2 Tagging exhaustivity

The number of tags assigned to a document indicates in how much detail the topics in a document are represented. Thus, the average number of tags per document by subject categories can help demonstrate the exhaustivity of indexing across different subjects. For measuring tagging exhaustivity, descriptive statistics are calculated on the average number of tags per document by 10 subject categories BUBL provides.

3.3.3.3 Tag specificity

For measuring tagging specificity, descriptive statistics are calculated on the number of documents associated with a tag which is one of the tags listed among the top ranked tags in Delicious. Regarding the top 20 tags, the measure of tag specificity is determined.

For example, for a document *Amazon.com,* the top tags include shopping, books, amazon, online, bookstore, music, web, internet, fun, and deals, excluding the tag “compras” which is a Spanish word (see Figure 21). Figure 21 demonstrates that among the top tags, clicking the first ranked tag, “shopping” results in 2,471,930 documents indexed by the tag in Delicious. The process of figuring out the number of documents indexed by a tag continues in turn for tags ranked from 2nd to 20th.



Everyone's Bookmarks for:

Amazon.com: Online Shopping for Electronics, Apparel, Computers, Bo...

www.amazon.com/gp/homepage.html

Saved 180 times, first saved by Bufodiarlechino on 22 Mar 06. View Chart

29 OCT 09	vrmorton	amazon books shopping
29 SEP 09	郭慧	書店 亞馬遜 網路書店
22 MAY 09	d0t3x3	amazon online shopping
30 APR 09	raym57	Amazon
16 APR 09	Hacienda23	Shopping
23 FEB 09	g4guitar	websites shopping books amazon
14 FEB 09	mahondennisj	shopping books amazon online bookstore music internet web
08 FEB 09	Cgammeter	
22 JAN 09	Online shopping from the earth's biggest selection of books, magazines, music, DVDs, videos, electronics, computers, software, apparel & accessories, shoes, jewelry, tools & hardware, housewares, furniture, sporting goods, beauty & personal care, broadband & dsl, gourmet food & just about anything else. lewiscb	web internet music online shopping fun
08 JAN 09	ha429rout	onlywire

Save this bookmark
Look up another URL

Tags
Top Tags

shopping	81
books	45
amazon	36
online	14
bookstore	13
music	12
web	6
internet	6
fun	4
compras	4
deals	3
catalog	3
亞馬遜	2
書店	2
網路書店	2
retail	2
libros	2
lojas	2
consumer	2
cds	2
loja	2
aukcje	1
baju-shop!	1

Save a new bookmark
Look up a URL

Recent shopping Bookmarks

Recent | Popular

See popular shopping bookmarks.

Tags shopping Type another tag

24 NOV 09

Timbuk2 Bags - Hidden Backpack - Recycled PET Fabric SAVE azph2	bags gifts shopping	2
Find Free Ebooks, Compare Ebook Prices | Inkmesh SAVE beholder	ebooks shopping	108
GameDude Computers OnLine Store SAVE MokalusOfBorg	hardware shopping	84
How to Shop for Free SAVE hellogillian	coupons money shopping	28
Saddleback Leather Company - Saddleback Leather Co. : Leather Bags, Briefcases & Luggage for Men & Women SAVE rmaddato	leather bags briefcase bag shopping	78
LOLmart Shirts: The Evolution of the Cat – LOL Shirts SAVE kym	shopping	3
Welcome to Corsair SAVE fairweatherfan	computers memory hardware shopping	18
BARNEYS NEW YORK OPENING SOON! SAVE claire17	fashion shopping	10
Playing Mantis - Toys and Crafts for Life SAVE thehaystackneedle	shopping	
Coupon Shack: Online Coupons & Promotional Codes for Discount Online Shopping SAVE riedfordjr	software free coupons coupon deals codes discounts promotion code shopping	41

1 2 3 4 5 6 7 ... Next >

2471930 Bookmarks

Tags
Related Tags 11

+design
+fashion
+clothing
+shop
+gifts
+home
+deals
+clothes
+furniture
+t-shirts
+art

Figure 21. The documents indexed by a tag

3.3.4 Analysis of tag attributes and tagging behaviors

3.3.4.1 Content analysis based on Functional Requirements for Bibliographic Records (FRBR)

To provide in-depth investigation on the characteristics of tags, this research analyzed the bibliographic attributes of tags which are not limited to subject properties. The process of identifying bibliographic attributes of tags was based on the Functional Requirements for Bibliographic Records (FRBR) model. Since the attributes defined in the FRBR model were derived from "a logical analysis of the data that are typically reflected in bibliographic records" (IFLA, 1998), it supports a more systematic and meticulous analysis of the attributes of tags. The FRBR model is described in detail below.

FRBR is a conceptual model of the "bibliographic universe" (works, texts, editions, documents and the like) that was developed by the International Federation of Library Associations and Institutions (IFLA 1998). It is intended to guide the development of systems for creating and managing bibliographic records. FRBR identifies four "Group 1" entity types (*work, expression, manifestation*, and *item*), defines relationships between them (a work is realized through an expression; an expression is embodied in a manifestation; a manifestation is exemplified by an item), and assigns characteristic attributes to each entity - for instance, works have form, expressions may be in a particular language, manifestations may have a typeface, and items may have a provenance. Figure 22 depicts Group 1 entities and relationships between them.

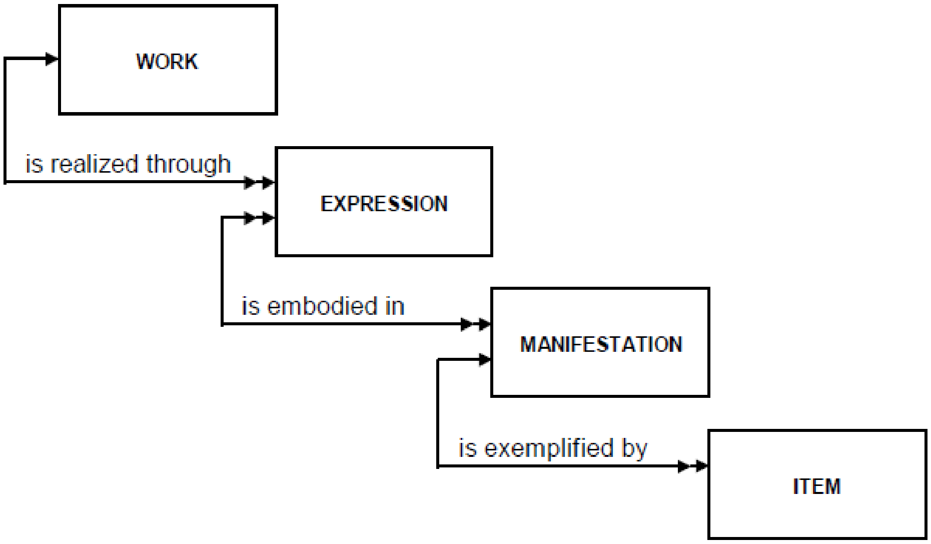


Figure 22. Group 1 entities and primary relationships (Source: IFLA, 1998)

The entity *work* is defined as “A distinct intellectual or artistic creation”, expression as “the intellectual or artistic realization of a work in the form of alphanumeric, musical, or choreographic notation, sound, image, object, movement, etc., or any combination of such forms”, manifestation as “the physical embodiment of an expression of a work” and item as “a single exemplar of a manifestation” (IFLA, 1998).

Each entity type is assigned a set of attributes. Works have attributes such as title and form; expressions have a language attribute (translations of the same work are different expressions); manifestations have attributes like typeface; and items have attributes such as condition and location. In this research, the scope of data analysis focuses on web documents, so consideration of manifestation and item has been excluded. Only the entities Work and Expression were considered and the attributes of both Work and Expression entities were investigated in order to map the attributes of tags to attributes defined for those two entities. Table 11 illustrates the attributes of Work and Expression among FRBR group 1 entities (IFLA, 1998). The attributes emphasized in bold face were only included for coding and other attributes were excluded for coding since it was determined that they are not applicable to web documents.

Table 11. FRBR Group 1 entities and logical attributes

Entities	Logical attributes
Work	**title of the work** **form of work** **date of the work** other distinguishing characteristic intended termination **intended audience** **context for the work** medium of performance (musical work) numeric designation (musical work) key (musical work) coordinates (cartographic work) equinox (cartographic work)
Expression	title of the expression **form of expression** **date of expression** **language of expression** other distinguishing characteristic extensibility of expression revisability of expression extent of the expression **summarization of content** context for the expression critical response to the expression **use restrictions on the expression** sequencing pattern (serial) expected regularity of issue (serial) expected frequency of issue (serial) type of score (musical notation) medium of performance (musical notation or recorded sound) scale (cartographic image/object) projection (cartographic image/object) presentation technique (cartographic image/object) representation of relief (cartographic image/object) geodetic, grid, and vertical measurement (cartographic image/object) recording technique (remote sensing image) special characteristic (remote sensing image) **technique (graphic or projected image)**

Table 12 shows the final list of FRBR attributes (IFLA, 1998) for coding and the coding scheme and coding instructions for tag attributes during content analysis are included in Appendix D. Since each attribute defined by FRBR is assumed to be disjoint (Renear & Choi, 2006), this research set up the principle that coding should not overlap.

Table 12. FRBR attributes and description

Entities	Logical attributes	Description
Work	title of the work (WT)	The title of the *work* is the word, phrase, or group of characters naming the *work*. There may be one or more titles associated with a *work*.
	form of work (WF)	The form of *work* is the class to which the *work* belongs (e.g., novel, play, poem, essay, biography, symphony, concerto, sonata, map, drawing, painting, photograph, etc.).
	date of the work (WD)	The date of the *work* is the date (normally the year) the *work* was originally created. The date may be a single date or a range of dates. In the absence of an ascertainable date of creation, the date of the *work* may be associated with the date of its first publication or release.
	intended audience (WI)	The intended audience of the *work* is the class of user for which the work is intended, as defined by age group (e.g., children, young adults, adults, etc.), educational level (e.g., primary, secondary, etc.), or other categorization.
	context for the work (WC)	Context is the historical, social, intellectual, artistic, or other context within which the *work* was originally conceived (e.g., the 17th century restoration of the monarchy in England, the aesthetic movement of the late 19th century, etc.).
Expression	form (EF)	The form of *expression* is the means by which the *work* is realized (e.g., through alpha-numeric notation, musical notation, spoken word, musical sound, cartographic image, photographic image, sculpture, dance, mime, etc.).
	date (ED)	The date of *expression* is the date the *expression* was created (e.g., the date the particular text of a *work* was written or revised, the date a song was performed, etc.). The date may be a single date or a range of dates. In the absence of an ascertainable date of *expression*, the date of the *expression* may be associated with the date of its publication or release.
	language of expression (EL)	The language of the *expression* is the language in which the *work* is expressed. The language of the *expression* may comprise a number of languages, each pertaining to an individual component of the *expression*.
	summarization of content (ES)	A summarization of the content of an *expression* is an abstract, summary, synopsis, etc., or a list of chapter headings, songs, parts, etc. included in the *expression*.
	use restrictions on the expression(EU)	Use restrictions are restrictions on access to and use of an *expression*. Use restrictions may be based in copyright, or they may extend beyond the protections guaranteed in law to the owner of the copyright.
	technique (graphic or projected image) (ET)	Technique is the method used to create a graphic image (e.g., engraving, etc.) or to realize motion in a projected image (e.g., animation, live action, computer generation, 3D, etc.).

3.3.4.2 Intercoder reliability test

Content analysis on tags extracted from Delicious top 20 tags is conducted for categorization based on FRBR attributes. The results of coded data can be trusted only when reliability can be demonstrated. Accordingly, a coder other than the researcher is recruited and the coded results are compared to measure the inter-coder reliability. Regarding the sub sample size for the inter-coder reliability test, Wimmer and Dominick (1987) recommend that between 10% and 25% of the data should be investigated to test intercoder reliability. In this research, 25% of the web document collection selected for data analysis is randomly sampled using the True Random Number Generator (www.random.org). For example, under 000 Generalities categories, the number of selected documents was 8, so sub-sample size in this category is 2. Thus, among 113 web documents, 29 web documents are selected for the intercoder reliability test (Table 13).

Table 13. The number of documents for intercoder reliability test

Top categories	**The number of selected documents**	**25%**	**The number of documents for inter-coder reliability**
000 Generalities	8	2	2
100 Philosophy and psychology	6	1.5	2
200 Religion	12	3	3
300 Social sciences	12	3	3
400 Language	9	2.25	2
500 Science and mathematics	10	2.5	3
600 Technology	8	2	2
700 The arts	21	5.25	5
800 Literature and rhetoric	15	3.75	4
900 Geography and history	12	3	3
Total	**113**	**28.25**	**29**

Thus, among 1,879 tags assigned to 113 documents, 442 tags assigned to 29 web documents are coded for intercoder reliability. The list of web documents for the intercoder reliability test is provided in Appendix E.

There are a number of measures of intercoder reliability. Lombard et al. (2004) describe several measures commonly used in social science and communication such as percent agreement, Holsti's method, Scott's pi (π) , Cohen's kappa (κ), and Krippendorff's alpha (α). The percent agreement index has advantages of simplicity and ease of calculation, but it records only agreements and disagreements. This index also has a flaw in that it does not account for agreement occurring by chance (Lombard et al., 2004). Holsti's method (1969) is a variation on the percent agreement index; it accounts for the situation in which the coders evaluate different units. But, when two coders evaluate the same units, the results by Holsti's method are the same as those by the percentage agreement index of reliability because it calculates percent agreement between two coders (Hayes, 2007; Lombard et al., 2004). Scott's pi (1955) takes into account both the observed proportion of agreement and the proportion that would be expected by chance. Yet, Scott's pi has a limitation to two coders and nominal data (Hayes, 2007). On the other hand, several researchers (Bakeman, 2000; Dewey, 1983) recommend Cohen's kappa (κ) (1960), one of the widely used measures for intercoder reliability. Cohen's kappa is identical to Scott's pi in that it accounts for agreement expected by chance. The equation for kappa (κ) is as follows:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

Pr(a): agreement, observed
Pr(e) : agreement, expected by chance

Unlike Scott's pi, the assumption of kappa is that the same two coders have coded all units, so it cannot be applicable to situations where different pairs of coders have coded different subsets of the units (Craig, 1981). Krippendorff (1978, 1987, 2004) also criticizes that Cohen's kappa (κ) is not appropriate for testing intercoder agreement. Krippendorff insists that since Cohen's kappa (κ) defines chance as "the statistical independence of two coders' use of categories", the categories one coder uses are not predictable from the categories the other coder uses.

Krippendorff's alpha (α) (1980) is also a commonly used measure for intercoder reliability. It is considered to be very flexible as it can account for different sample sizes and missing data, and can be applied to any number of observers, any number of categories, and any level of measurements, e.g., nominal, ordinal, interval, ratio, and more (Hayes, 2007; Lombard et al., 2004; Krippendorff, 2004). Alpha (α)'s general form is as follows (Krippendorff, 2004):

$$\alpha = 1 - \frac{D_o}{D_e}$$

D_o : disagreement, observed
D_e : disagreement, expected by chance

α = 1 means observers agree perfectly, i.e., perfect reliability and the value of D_o is zero. Also, α = 0 means the absence of reliability, and D_o=$D_{e.}$ Thus, α's range is explained by:

$$1 \geq \alpha \geq 0 \begin{cases} - \text{Systematic disagreement} \\ \pm \text{Sampling errors} \end{cases}$$

Although many reliability measures have been used and discussed by several researchers, there has been no consensus on a best measure for reliability, and each index has its own qualities and assumptions (Lombard et al., 2004; Taylor & Watkinson, 2007). In this research, therefore, four indices mentioned above − i.e., Holksti's method, Scott's pi (π), Cohen's kappa (κ) and Krippendorff's alpha (α) − are used to test intercoder reliability. Calculating and reporting reliability by using more than one index is a preferred approach that can take into account any bias or weaknesses caused by the results from one (Lombard et al., 2004).

CHAPTER 4: INDEXING CONSISTENCY

4.1 Overview

This chapter presents the results on the analyses of indexing/tagging consistency. The results illustrate the patterns and tendency of user tagging in comparison to professional indexing. The analysis of indexing consistency was conducted on the following:

(1) Inter-indexer/tagger consistency among Delicious users regarding subject areas

(2) Inter-indexer consistency between two groups of professional indexers from BUBL and Intute subject gateways

(3) Inter-indexer consistency between Delicious's users and Intute's professional indexers

Research questions from 1 to 5 are answered. Furthermore, the comparison of consistency between taggers (Delicious) and professionals (BUBL and Intute), and the comparison of consistency with three different measures are discussed.

4.2 Consistency in Social Tagging in Delicious

Research question 1 on consistency of social tagging was:

> Question 1) How consistent is social tagging at Delicious regarding subject indexing of web resources? Is there a relationship between its indexing consistency and subject areas indexed?

In three measures (Euclidean distance, dot product, and cosine), it was found that there was reasonable consistency of similarity among Delicious taggers over all subjects. Regarding the

Euclidean distance measure, since the values by the adjusted Euclidean distance formula represent minus figures (see Chapter 3, 3.3 Data Analysis, p. 52), the longer bar in Figure 23 means lower similarity. Therefore, Figure 24 is supplied to provide a better visualization for understanding the pattern of similarity.

Comparing figures 25 and 26, it is apparent that the cosine measure varies less across the ten subjects compared to the dot product measure.

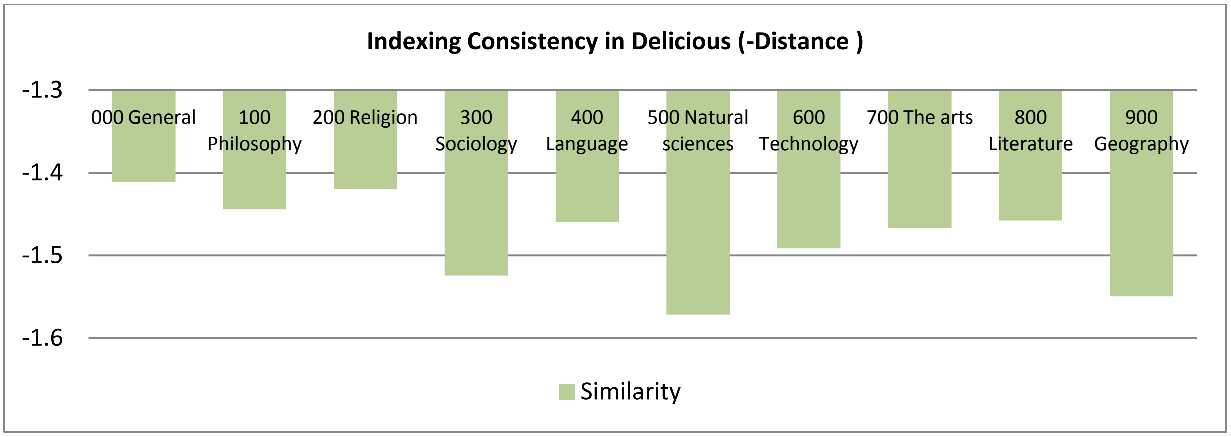


Figure 23. Indexing consistency in Delicious using the distance measure (bar graph)

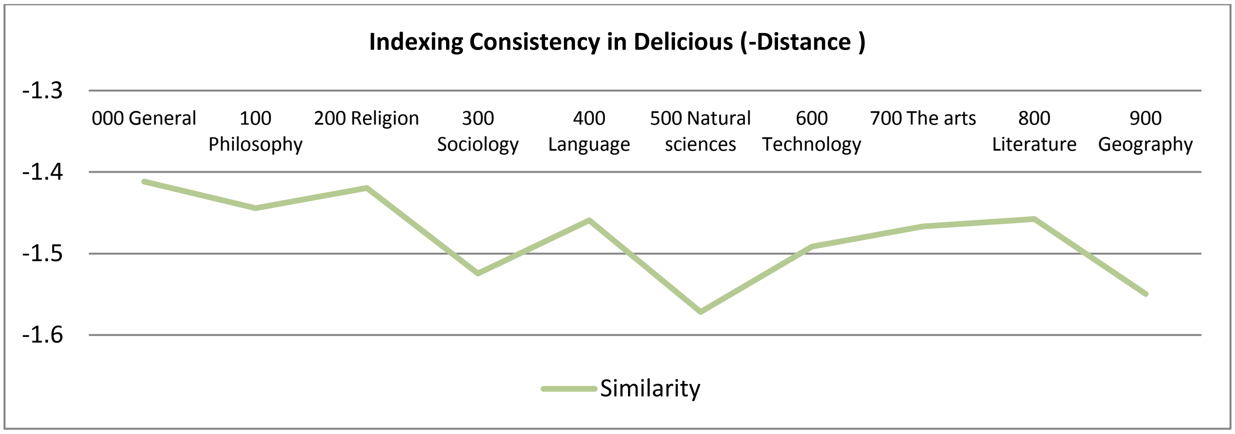


Figure 24. Indexing consistency in Delicious using the distance measure (line graph)

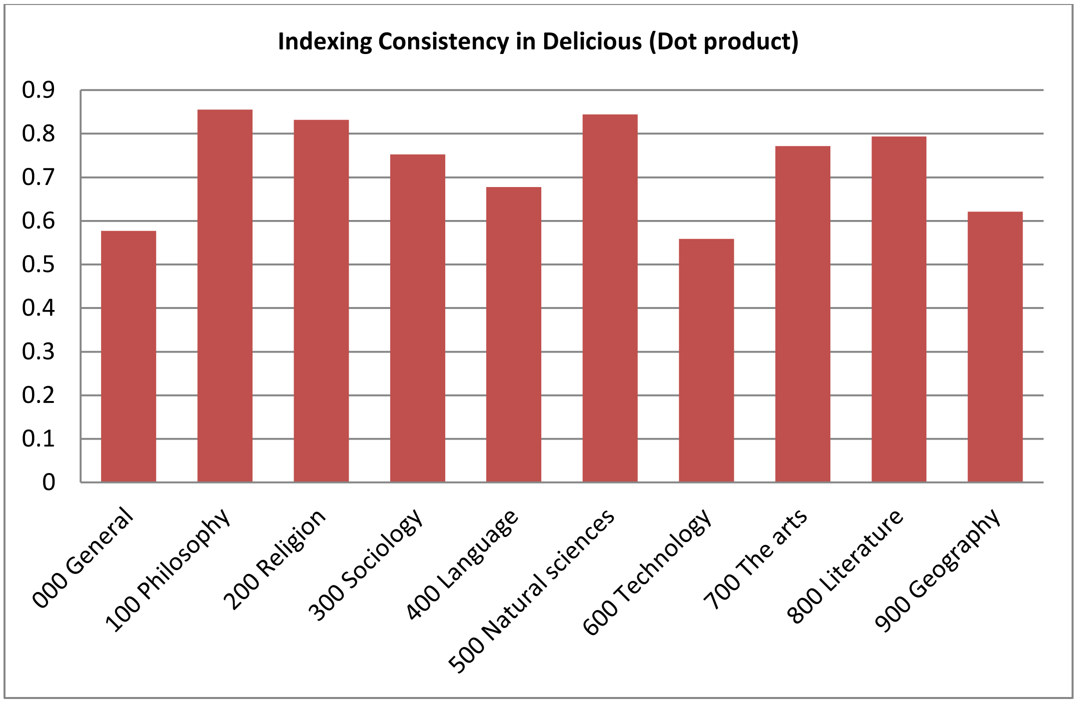


Figure 25. Indexing consistency in Delicious using the dot product measure

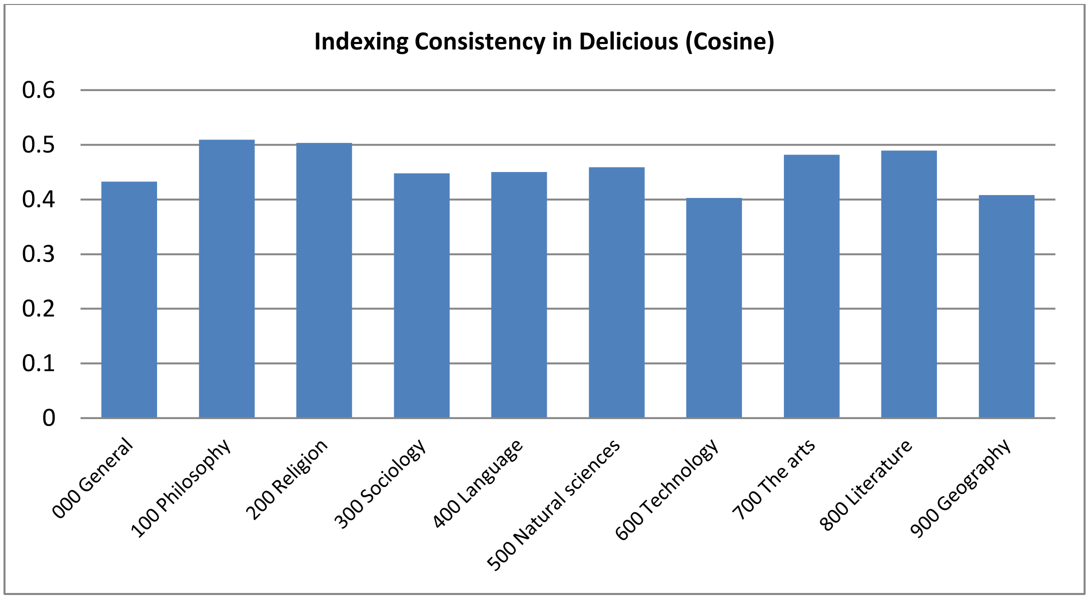


Figure 26. Indexing consistency in Delicious using the cosine measure

4.3 Consistency in Professional Indexing between BUBL and Intute

Research questions on indexing consistency between two professional groups were:

> Question 2) How consistent is professional indexing between BUBL and Intute?
>
> Question 3) Are there various or alternative interpretations of the same web document between two professionally indexed subject gateways, BUBL and Intute?

Comparison of indexing between two groups of professionals (BUBL vs. Intute) yielded inconsistency of similarity over all subjects with three measures (Euclidean distance, dot product, and cosine) (see Figures 27, 29, and 30). Figure 28 is supplied to provide a better display for understanding the pattern of similarity.

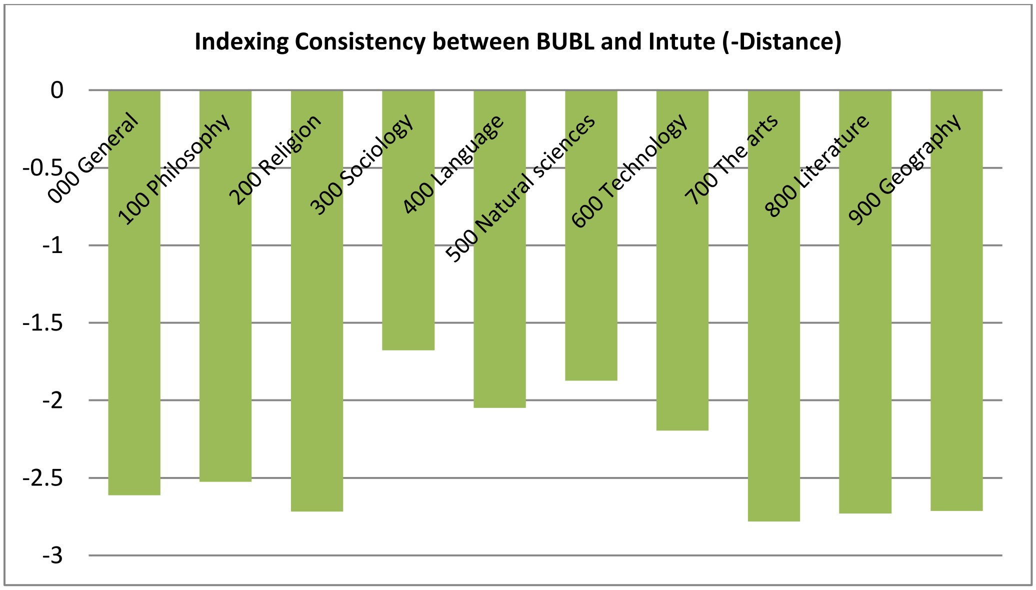


Figure 27. Indexing consistency between BUBL and Intute using the distance measure (bar graph)

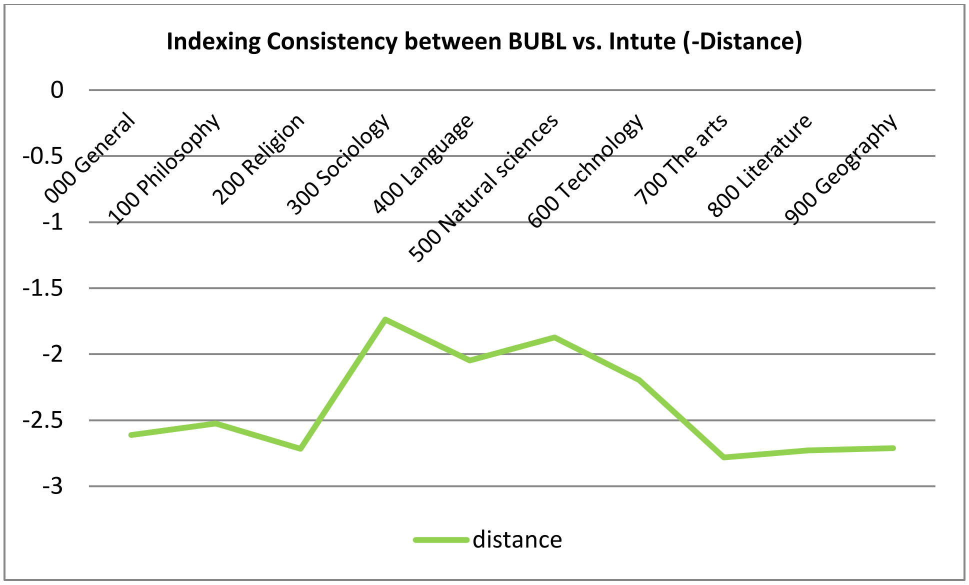


Figure 28. Indexing consistency between BUBL and Intute using the distance measure (line graph)

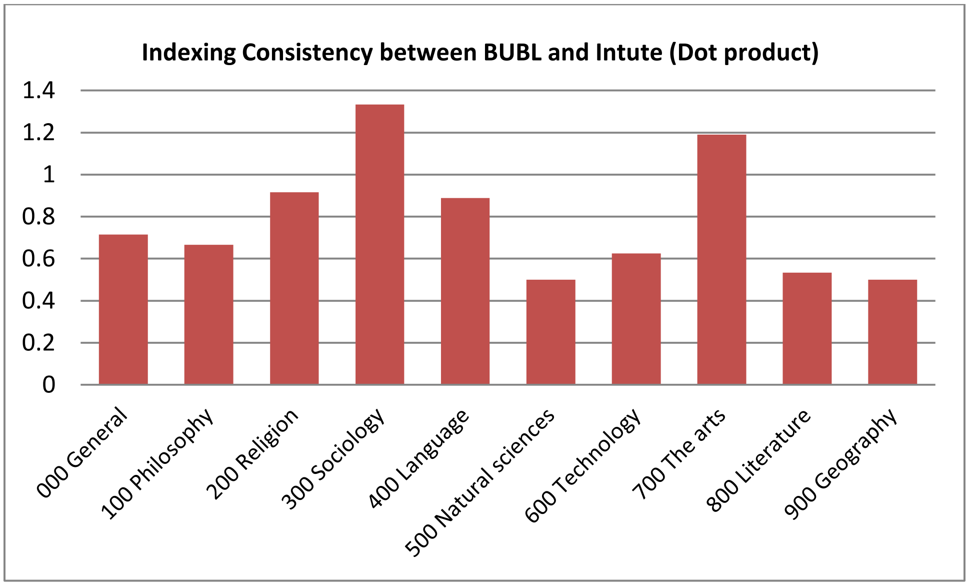


Figure 29. Indexing consistency between BUBL and Intute using the dot product measure

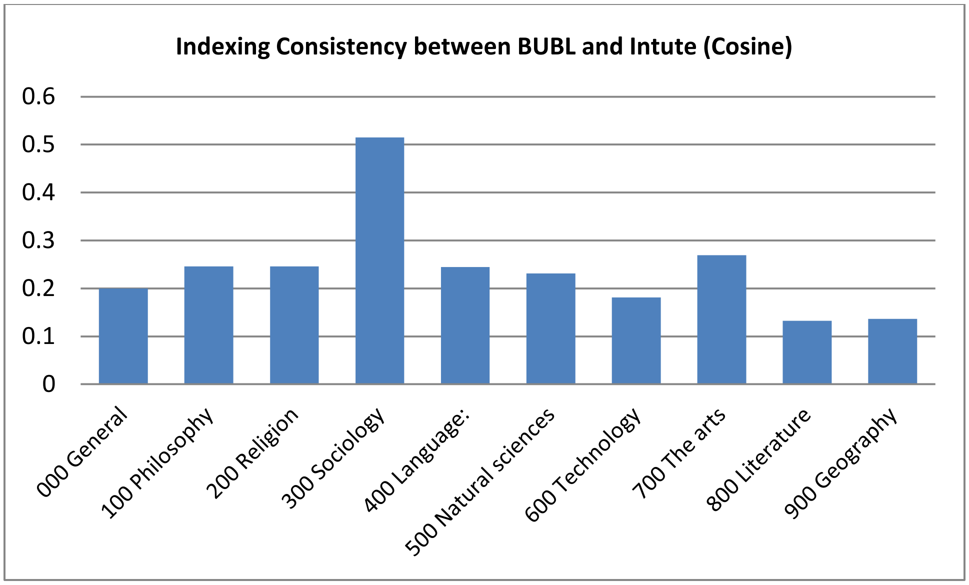


Figure 30. Indexing consistency between BUBL and Intute using the cosine measure

Low similarity in Literature and Geography in all three measures revealed that BUBL and Intute have different points of view on the same documents for those subjects. Regarding many documents in Literature, Intute categorized them into Languages and the Area Studies category (Table 14). On the other hand, BUBL maintains the Language category as a separate category.

Table 14. Indexing on Literature (BUBL vs. Intute)

Literature	Title	BUBL	Intute
808.8 Literature: general	Google Book Search, http://books.google.com/	Literature, Rhetoric, Collections	Modern Languages, Area Studies, Museums, Libraries, Archives
830 German literature	19th Century German Stories, http://www.fln.vcu.edu/menu.html	Literature, rhetoric, german_literature	Modern_languages, area_studies
880 Classical Greek literature	The Internet Classics Archives, http://classics.mit.edu/	Literature, rhetoric	Humanities, Classics, Modern_languages, area_studies

Also, in Intute, several documents in Geography were located in Modern Languages and Area Studies categories (see Table 15).

Table 15. Indexing on Geography (BUBL vs. Intute)

Geography	Title	BUBL	Intute
930 History of ancient world	English Heritage, http://www.english-heritage.org.uk	Geography, History, Ancient world	Humanities, History, Military_history Humanities, Museum, library, archive_studies, Architecture, planning, Modern_languages, area_studies, European, English
940 History of Europe	World War I Document Archive , http://wwi.lib.byu.edu/index.php/Main_Page	Geography, history, Europe	social_sciences, Government_policy, Military_science, Wars, World_War_One
980 History of South America	Latin American Network Information Center, http://lanic.utexas.edu/	Geography, history, South_America	Modern_languages, area_studies, media_studies, Modern_languages, area_studies, Philosophy, religion

In the dot product and cosine measures, regarding several documents on Natural sciences and Technology, the similarity between BUBL and Intute was relatively low. This indicates that BUBL and Intute have somewhat different points of view on the same documents in those subjects or use different terminology (see Table 16 and Table 17).

Table 16. Indexing on Natural Sciences (BUBL vs. Intute)

Natural Sciences	Title	BUBL	Intute
500 Natural sciences: national centres	National Science Foundation, http://www.nsf.gov/	Natural sciences	Engineering, Physical sciences
570 Life sciences, biology	BBSRC: Biotechnology and Biological Sciences Research Council: http://www.bbsrc.ac.uk/	Natural sciences, Life sciences, Biology	Biological sciences
580 Plants, general resources	Botanical Society of America Online Image Collection: http://images.botany.org/	Natural sciences, Plants	Biological sciences, Botany, Images
590 Animals, general resources	Animal Diversity Web: http://animaldiversity.ummz.umich.edu/site/index.html	Natural sciences, Mathematics, Animals	Biological sciences, Zoology

Table 16 illustrates that regarding most of the documents on Natural Sciences, Intute categorized them into “Biological Sciences” while BUBL has “Biology” in the 570 Life sciences, biology category. BUBL also separates "Biology" from "Plants" in two different categories. However, it is important to note that two pairs of terms indexed by BUBL and Intute, "Plants" versus "Botany" and "Animals" versus "Zoology" are different but comparable terms. In the Library of Congress Subject Headings, those two pairs of terms are defined as "Related Terms (RT)".

In terms of the Technology subject, Table 17 shows that Intute placed several documents on Technology into "Arts" or "Physical sciences".

Table 17. Indexing on Technology (BUBL vs. Intute)

Technology	Title	BUBL	Intute
670 Manufacturing	Bad Human Factors Designs, http://www.baddesigns.com/	Technology, Manufacturing	performing_arts, art, Design, Industrial_design, Ergonomics Psychology, organizational_psychology
600 Technology: general resources	EurekAlert, http://www.eurekalert.org/	Technology	Press_releases, news, Physical_sciences
620 Engineering: education and research	EDINA, http://edina.ac.uk/index.shtml	Technology, Engineering	Physical_sciences

4.4 The Results of Inferential Statistics on Indexing Consistency

To statistically test the results of indexing consistency among taggers in Delicious, inferential statistics were calculated. As discussed in Chapter 3, 3.3.2 *Analysis of Inter-indexer/tagger consistency*, ANOVA and Kruskal-Wallis are different in that in ANOVA, we assume that the distribution of each group should be normally distributed while in the Kruskal-Wallis test, we do not make any assumption about the distribution.

The tendency of the indexing consistency in Delicious can be more graphically explained as illustrated in Figure 31. A box plot is helpful in interpreting the distribution of a dataset, that is, in seeing how a variable is distributed. A box plot indicates the lower quartile (25%), the median (50%) and upper quartile (75%). A box plot also indicates “outliers”. Any data observation which lies more than 1.5 times the inter-quartile range (the difference between the upper and lower quartiles) is considered an outlier.

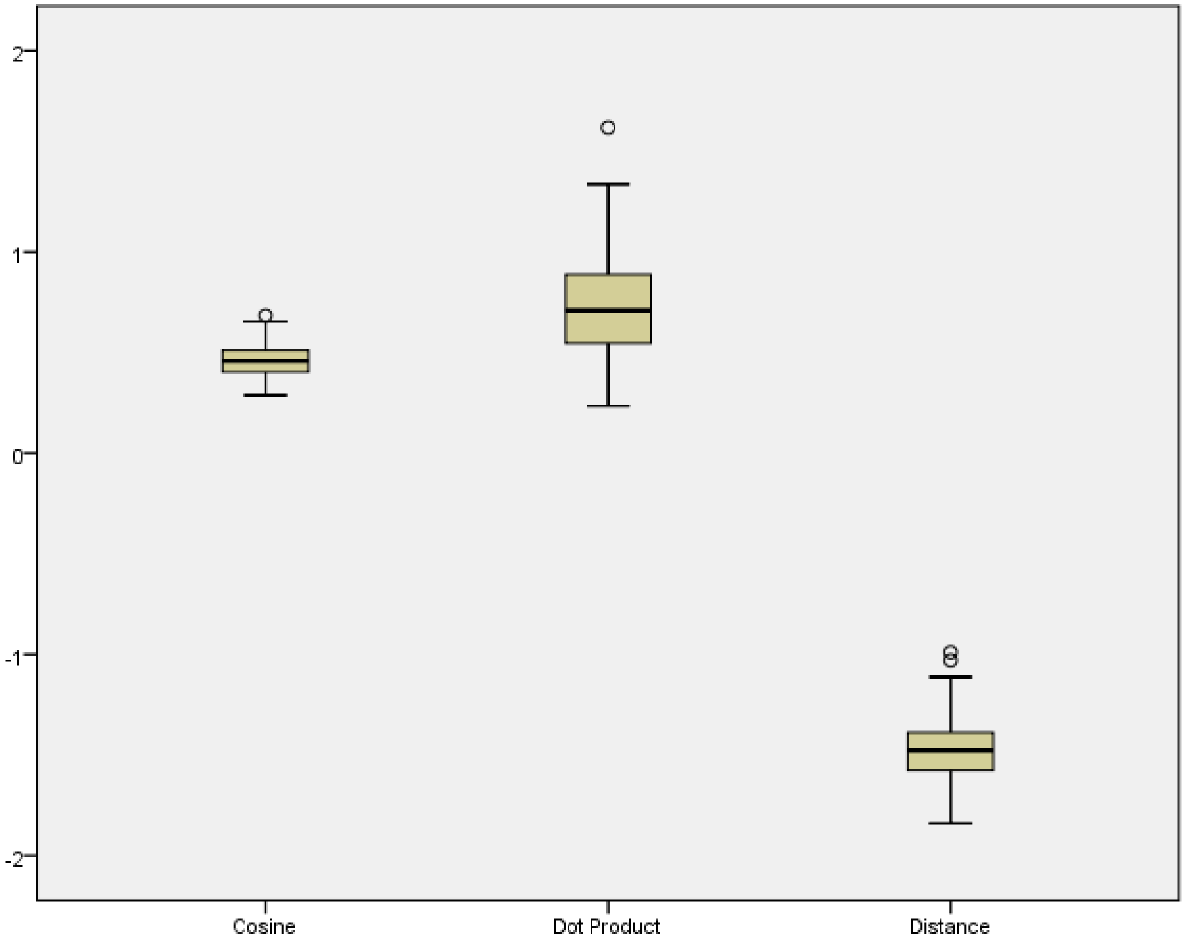


Figure 31. A box plot of indexing consistency in Delicious

In Figure 31, the middle black line is the median, the shaded region shows middle 50 % of consistency values, and outliers are represented by open dots. The box plot indicates the minimum value excluding outliers by connecting it to the box with a horizontal line or “whisker”.

Also, it indicates the maximum value excluding outliers by extending above the box with a "whisker". Furthermore, it is possible to understand the distribution of sets of data by looking at the lengths of the whiskers on both sides of the box and the position of the median within the box. It is noticeable that in the case of dot product measure, the whisker associated with the upper quartile is larger than the whisker associated the lower quartile. In these cases, therefore, we could detect possible skewness in the data.

To conduct in-depth investigation of the observed distribution of tag data, that is, to see whether tagging consistency in Delicious is normally distributed or not, the test of normality was conducted. This research used the Shapiro-Wilk test (1965) which is appropriate when the sample size is between 3 and 2000. The Shapiro-Wilk test calculates a *W* statistic (Pearson & Hartley, 1972):

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$

$x_{(i)}$: the ordered sample values ($x_{(1)}$ is the smallest)
a_i : constants generated from the means, variances and covariances of the order statistics of a sample of size *n* from a normal distribution

Table 18 shows the results of normality on tagging at Delicious. It was demonstrated that in dot product measure, tagging consistency was not normally distributed while tagging consistency in the cosine similarity and distance metrics measures derived from normal distributions. As shown in Table 19, in the Dot product measure, two subject categories (400 Language and 500 Natural Sciences) were not normally distributed while in the Cosine similarity measure, it turned out that consistency on all subject categories was normally distributed.

Table 18. The results of normality test on tagging consistency

	Shapiro-Wilk		
	Statistic	df	Sig.
Distance	.988	113	.433
Cosine	.989	113	.503
Dot product	.971	113	**.016**

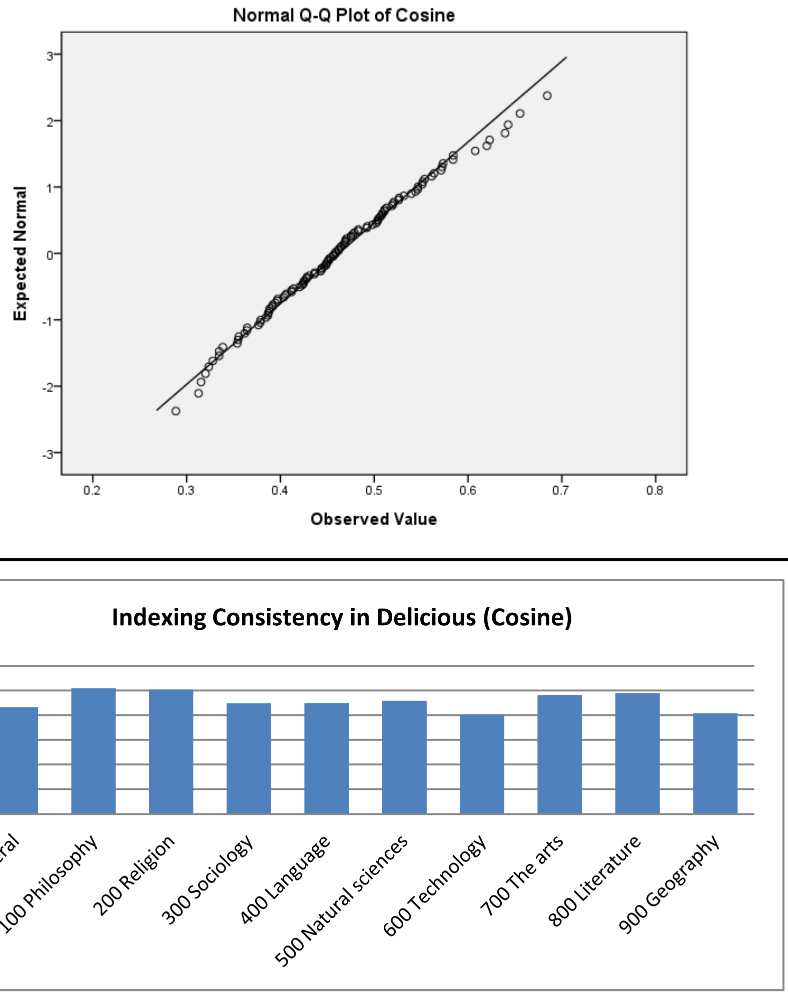
The results of the normality test presented in Table 18 can be graphically explained by using the Q-Q plot (Quantile-quantile plot). The normality Q-Q plots for each subject category (Table 19) are provided in Appendix F.

Table 19. The results of normality test on tagging consistency over all subjects

	Subject Category	**Shapiro-Wilk**		
		Statistic	**df**	**Sig.**
Cosine	0	.954	8	.755
	100	.916	6	.479
	200	.978	12	.975
	300	.951	12	.653
	400	.947	9	.652
	500	.962	10	.803
	600	.962	8	.827
	700	.960	21	.517
	800	.983	15	.987
	900	.949	12	.627
Dot Product	0	.944	8	.656
	100	.885	6	.294
	200	.954	12	.691
	300	.935	12	.434
	400	**.799**	**9**	**.020**
	500	**.736**	**10**	**.002**
	600	.877	8	.177
	700	.928	21	.126
	800	.942	15	.409
	900	.968	12	.892

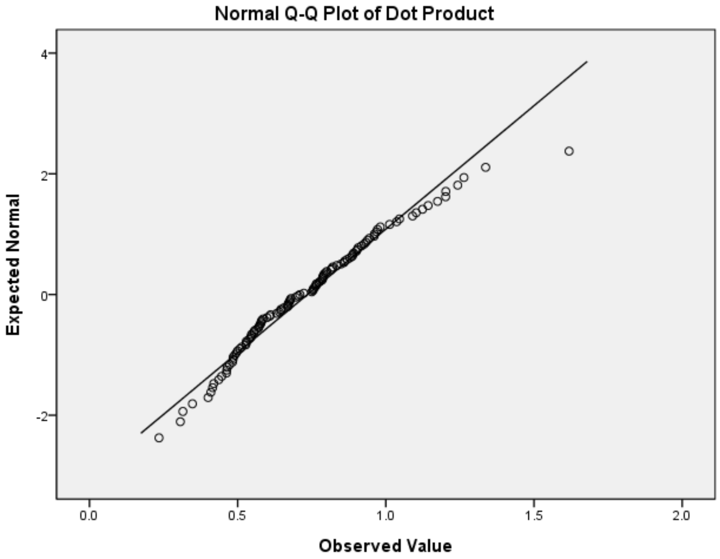
Table 19 (cont.)

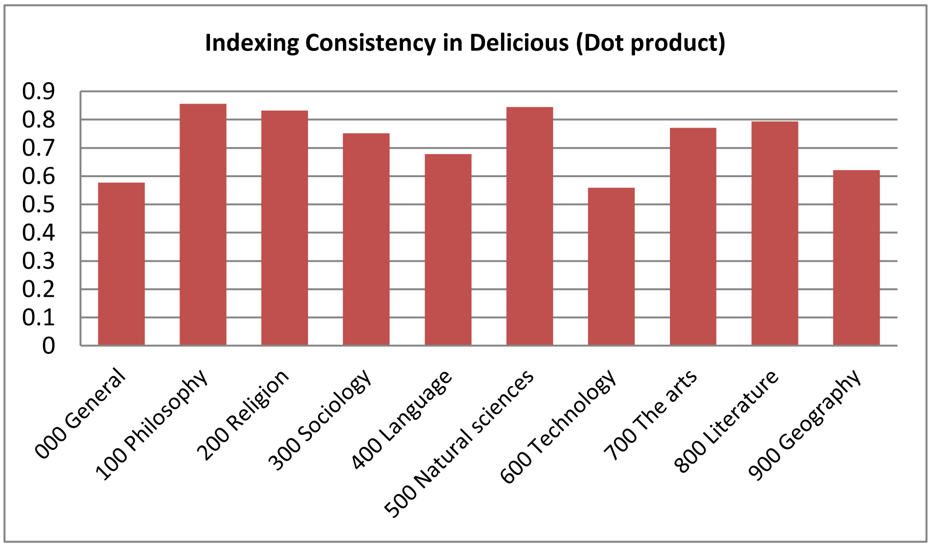
Distance	0	.928	8	.500
	100	.925	6	.540
	200	.960	12	.778
	300	.964	12	.843
	400	.946	9	.648
	500	.935	10	.500
	600	**.758**	**8**	**.010**
	700	.927	21	.118
	800	.936	15	.334
	900	.923	12	.313



Raw histogram of indexing consistency in Delicious using the cosine measure

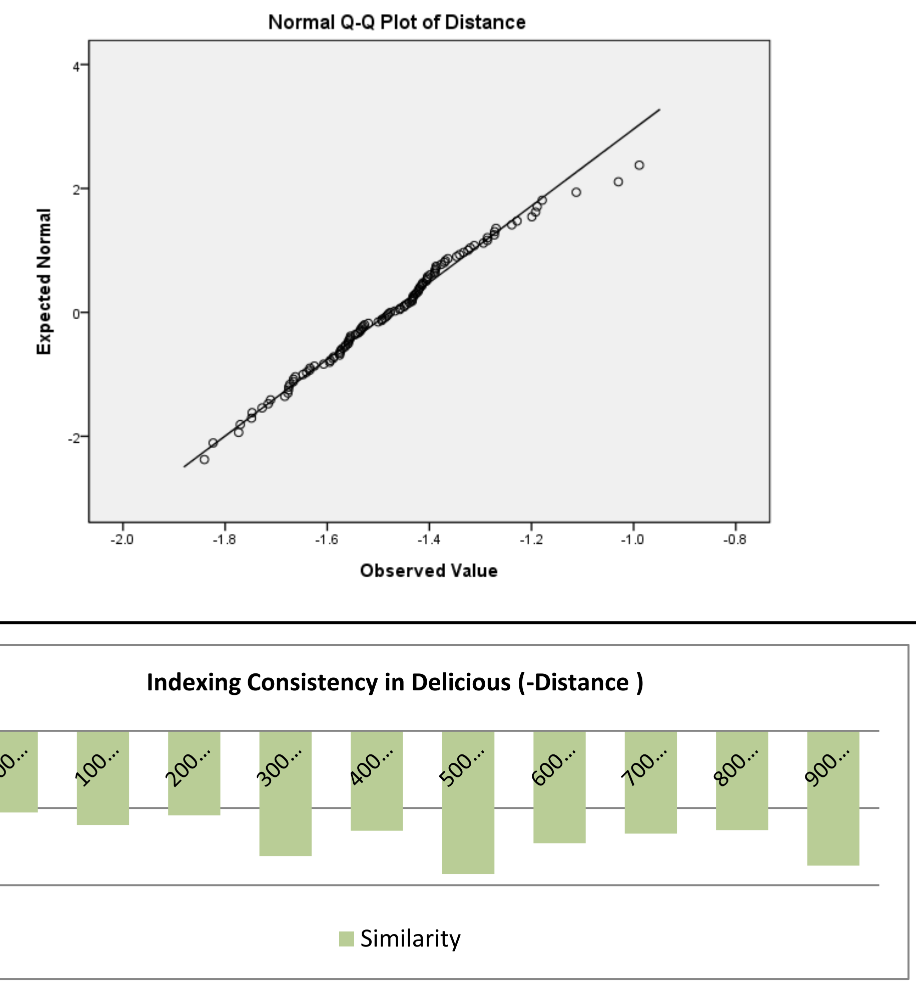
Figure 32. Q-Q plot of tagging consistency in the cosine similarity





Raw histogram of indexing consistency in Delicious using the dot product measure

Figure 33. Q-Q plot of tagging consistency in the dot product similarity



Raw histogram on indexing consistency in Delicious using the distance measure

Figure 34. Q-Q plot of tagging consistency in the distance measure

Figures 32-34 show the Q-Q plots and raw histograms for the three indexing consistency measures. In the Q-Q plots, when the observed values plot closely to the expected normal values, we can say the data are normally distributed. Figure 32 illustrates that the observed values in the cosine similarity measure are closest to the expected normal values.

As discussed above, the results of the normality test on tag data and Q-Q plots suggest that the analysis of consistency using the ANOVA could lead to misleading results since tag data did not follow a normal distribution in the dot product measure. In this research, therefore, the Kruskal-Wallis test was also conducted since it does not assume normal distribution among different groups. The Kruskal-Wallis test is appropriate when the data are very far from normally distributed. The Kruskal-Wallis is the non-parametric version of one-way ANOVA. Kruskal-Wallis is applied under the same conditions as one-way ANOVA, except that the dependent variable need not be normally distributed for Kruskal-Wallis. In this research, the null hypothesis and alternative hypothesis for both ANOVA and the Kruskal-Wallis test were as follows:

ANOVA:

The null hypothesis: there is no difference in the average indexing similarity among 10 different subject areas.

H_o: $\mu_1 = \mu_2 = ... = \mu_{10}$

The alternative hypothesis: the average indexing similarity for the 10 different subject areas is not the same. At least one pair of averages is different.

$H_1: \ \mu_i \neq \mu_j$

Kruskal-Wallis test:

The null hypothesis: 10 different subject areas have the same distribution.

H_o: $\mu_1 = \mu_2 = ... = \mu_{10}$

The alternative hypothesis: at least one of the subject areas tends to yield larger values than at least one of the other subject areas.

$H_1: \ \mu_i \neq \mu_j$

In Table 20, in the results of ANOVA, cosine p-value and dot product p-value suggest that there *is* a significant difference over all subjects for taggers at Delicious, but these p-values are more than p-values for professionals between two groups BUBL and Intute (Table 21). It means that indexing similarity among taggers is more consistent than indexing similarity between BUBL and Intute. In addition, in the results of the Kruskal-Wallis test, cosine p-value and dot product p-value rejected the null hypothesis that 10 different subject areas have the same distribution when the significance level is .05. Yet, these p-values are also more than p-values for professionals between two groups BUBL and Intute, which means there are more similar patterns of indexing among taggers at Delicious than that between two professional groups.

Table 20. ANOVA & Kruskal-Wallis (taggers)

Taggers (Delicious)		**ANOVA**					**Kruskal-Wallis**
		Sum of Squares	df	Mean Square	F	Sig.	Sig.
Cosine	Between Groups	.126	9	.014	2.285	**.022**	**.036**
	Within Groups	.630	103	.006			
	Total	.756	112				
Dot product	Between Groups	1.034	9	.115	2.094	**.037**	**.033**
	Within Groups	5.655	103	.055			
	Total	6.689	112				
Distance	Between Groups	.250	9	.028	1.070	.391	.253
	Within Groups	2.671	103	.026			
	Total	2.921	112				

The analysis of ANOVA confirmed that between BUBL and Intute's professional indexers, there is a difference in the average indexing consistency among 10 different subject areas when α is 0.05. Also, the results of the Kruskal-Wallis test rejected the null hypothesis, which means that the distribution of similarity over all subjects is not the same between BUBL and Intute (see Table 21).

Table 21. ANOVA & Kruskal-Wallis (professional groups)

Professional Groups (between BUBL and Intute)		**ANOVA**					**Kruskal-Wallis**
		Sum of Squares	df	Mean Square	F	Sig.	Sig.
Cosine	Between Groups	1.396	9	.155	3.681	**.000**	**.013**
	Within Groups	4.341	103	.042			
	Total	5.737	112				
Dot product	Between Groups	10.153	9	1.128	3.453	**.001**	**.002**
	Within Groups	33.652	103	.327			
	Total	43.805	112				
Distance	Between Groups	18.297	9	2.033	3.914	**.000**	**.001**
	Within Groups	53.496	103	.519			
	Total	71.793	112				

4.5 Consistency between Tagging in Delicious and Professional Indexing in Intute

Research questions on indexing consistency between Delicious taggers and Intute indexers were:

Question 4) How consistent is tagging/indexing between Delicious taggers and Intute professionals?

Question 5) Would Delicious users' tags provide additional subject access points beyond index terms or keywords that Intute professionals provide?

For all three measures, there was relatively high consistency concerning the Language subject but relatively low consistency regarding the Technology subject between Delicious tagging and Intute professional indexing (Figures 35, 37, and 38). Figure 37 is supplied to provide a better look for understanding the pattern of similarity.

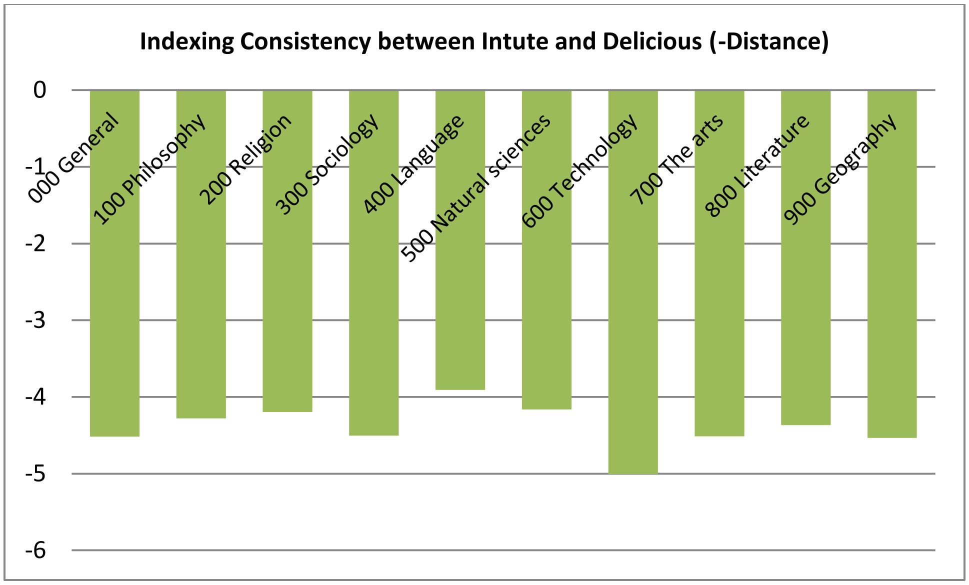


Figure 35. Indexing consistency between Intute and Delicious using the distance measure (bar graph)

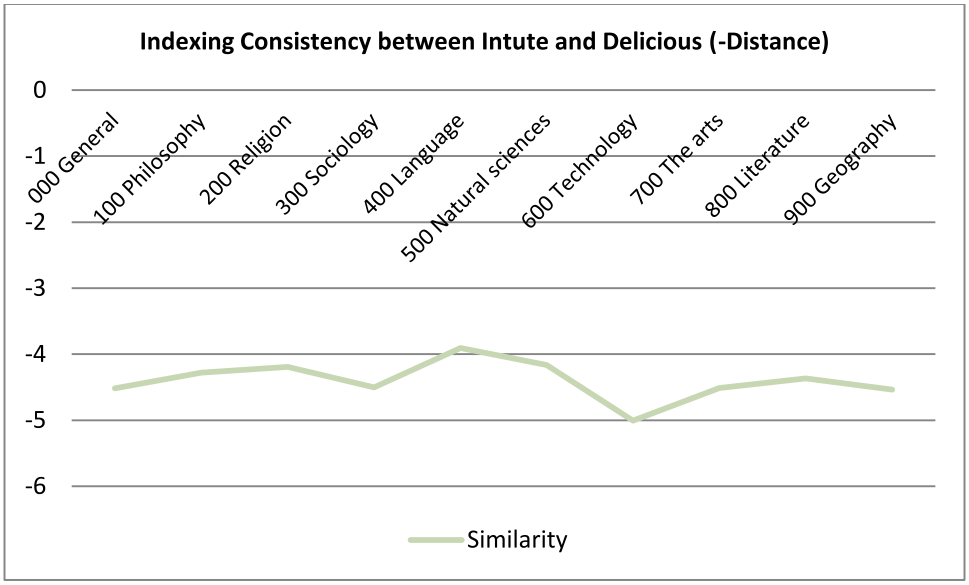


Figure 36. Indexing consistency between Intute and Delicious using the distance measure (line graph)

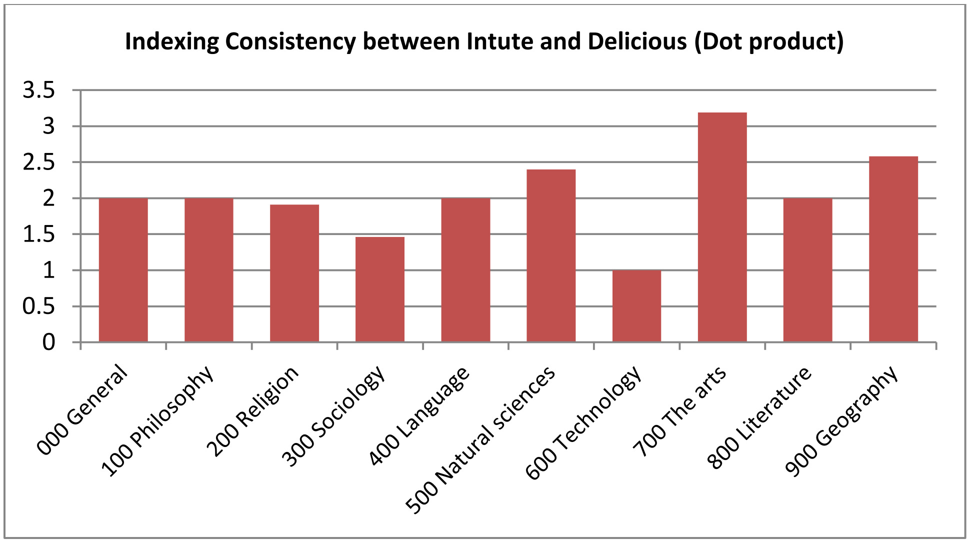


Figure 37. Indexing consistency between Intute and Delicious using the dot product measure

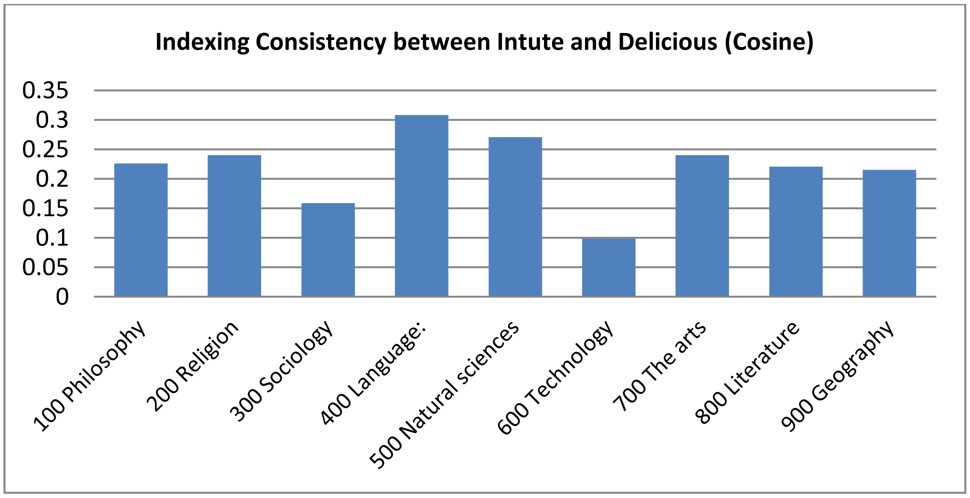


Figure 38. Indexing consistency between Intute and Delicious using the cosine measure

It is worthwhile to note that the Sociology subject, showing high similarity between two professional groups (BUBL and Intute) (Figures 26, 28, and 29), indicated low similarity between taggers and professionals (Delicious and Intute) (Figures 35, 37, and 38). Low similarity in Sociology and Literature between Delicious taggers and Intute professionals could be attributed to tags that included additional access points with many newly-coined terms such as ebook, online, web, web 2.0, e-guides, e-learning and cyberspace which reflect more accurate descriptions of the web documents (Table 22).

Table 22. Indexing on Sociology and Literature (Intute vs. Delicious)

Subject	Title	Intute	Delicious
Sociology (301 Sociology: general resources)	Sociological Tour Through Cyberspace, www.trinity.edu/~mkearl/index.html	death, euthanasia, families, homicide, mass media, time	sociology, links, resources, research, culture, web, science, resource, cyberspace, technology, web2.0, writing, social, internet, politics, reference, statistics
Sociology (370 Education)	Excellence Gateway , http://excellence.qia.org.uk/	numeracy, learning, key_skills, literacy	resources, education, e-learning, qia, teaching, learning, learning_resource , agency, elearning , quality, materials , jobs, qia_excellence, resource, e-guides, curriculum
Literature 808.8 Literature: general collections	Google Book Search, http://books.google.com/	writers, authors, books, search engines	books, google, search, ebooks, reference, book, library, research, tools, literature, search engine, web2.0, education, reading, resources, online, web, database
Literature 820 English, Scottish and Irish literature	Cambridge History of English and American Literature, http://www.bartleby.com/cambridge/	literature, poetry, fiction, drama, Renaissance, Restoration, English, American, poets, poems, Anglo_Saxon, plays, writings, encyclopedias, history	literature, history, reference, encyclopedia, ebooks, books, humanities, research , language, reading, criticism, academic, writing, resources, information, englishliterature

The Technology subject showed low consistency due to different levels of indexing between Intute indexers and Delicious taggers. For example, regarding the document 610 *Medical sciences, medicine*, Intute keywords tend to be broader terms, i.e., “disease” and “patient education,” but Delicious tags consist of terms in various semantic relationships, e.g., broader terms or narrower terms (Table 23). As shown in Table 23, tags on the document 610 *Medical sciences, medicine* include “health”, “medical”, “medicine”, “drugs”, “healthcare” etc. In the Library of Congress Subject Heading (LCSH), two terms “health”, and “medical” are represented as “narrower terms” of that term “medicine”. The term “healthcare” does not exist in the LCSH, but an alternative term “medical care” is represented as a narrower term of the term “health”.

Table 23. Indexing on Technology (Intute vs. Delicious)

Technology	Title	Intute	Delicious
610 Medical sciences, medicine	MedicineNet, http://www.medicinenet.com/script/main/hp.asp	Disease, Patient_Education	health, medical, medicine, reference, drugs, information, education, news, research, healthcare, dictionary, science, search, resources, doctors, diseases, biology
630 Agriculture and related technologies	AgNIC: Agriculture Network Information Center, http://www.agnic.org/	agricultural_sciences, agriculture, agricultural_education, information_centres,	agriculture, research, food, information, statistics, environment, plants, farming, libraries, international, database, library, agnic, science, associations, produce, portal, horticulture
660 Chemical engineering	American Institute of Chemical Engineers, http://www.aiche.org/	young_engineers	engineering, chemistry, chemical, aiche , organization, professional, associations, society, engineers american , education, institute, chemicalengine, job, research, science, work , usa

On the other hand, Natural Sciences, showing low similarity between two professional groups BUBL and Intute (see Figures 28 and 29), demonstrated relatively higher similarity between Delicious and Intute where its similarity reached the second highest peak in both Euclidean distance and cosine measures. Table 24 illustrates that while Delicious and Intute are including many common terms between them, for some terminology, Delicious tags also additionally supply users' preferred or up-to-date terms. Examples are "bioinformatics" and "biotech" for the term "biotechnology", and "cheminformatics" for "chemistry".

Table 24. Indexing on Natural Sciences (Intute vs. Delicious)

Natural Sciences	Title	Intute keywords	Delicious top ranked tags
500 Natural sciences: national centres	National Science Foundation, http://www.nsf.gov/	science-policy, USA	science, research, education, government, nsf, funding, reference, technology, news, grants, academic, foundation, usa, biology, national, information, resource
540 Chemistry	Linux4Chemistry, http://www.redbrick.dcu.ie/~noel/linux4chemistry/	software, Linux, computational_chemistry	linux, chemistry, software, science, visualization, simulation, reference, opensource, research, cheminformatics, bioinformatics, chemical, physics, modeling, tools, python, quantum, links, java
570 Life sciences, biology	BBSRC: Biotechnology and Biological Sciences Research Council: http://www.bbsrc.ac.uk/	research_support, research_institutes, biology, Biological_sciences, Research, Great_Britain, Biotechnology	research, science, biotechnology, funding, biology, uk, education, work, bioinformatics, bioscience, development, bbsrc, research, councils, research_councils, postgraduate, news, academic biotech, biological, researchcouncil
580 Plants, general resources	Botanical Society of America Online Image Collection: http://images.botany.org/	Botany, Plants	images, botany, plants, biology, science, research, photos, pictures, media, collection, horticulture, gardening, multimedia, flowers, botanica, biologyguide

4.6 Comparison of Taggers (Delicious) and Professionals (BUBL and Intute)

Using three measures (Euclidean distance, dot product, and cosine), it was shown that there was reasonable consistency over all subjects among taggers in Delicious. In contrast, indexing similarity between two groups of professionals (BUBL vs. Intute) illustrated more variation over all subjects in three measures (Figures 39, 40 and 41).

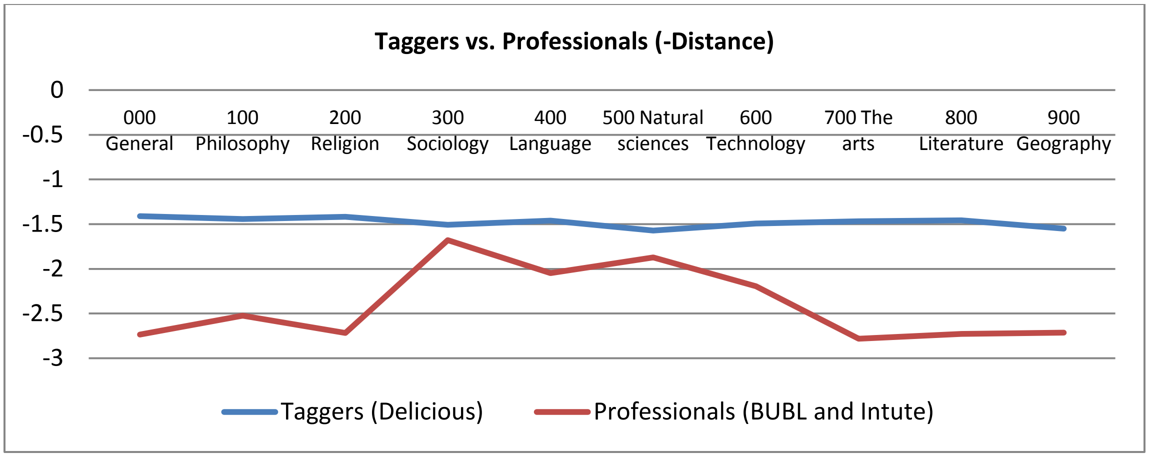


Figure 39. Indexing consistency between taggers (Delicious) and professionals (BUBL and Intute) using the distance measure

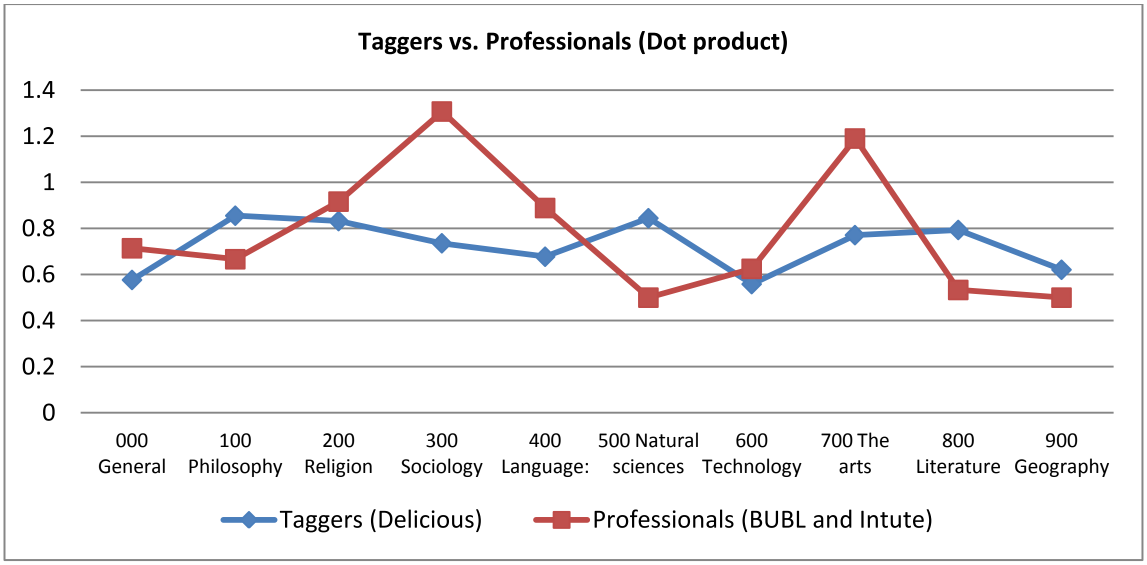


Figure 40. Indexing consistency between taggers (Delicious) and professionals (BUBL and Intute) using the dot product measure

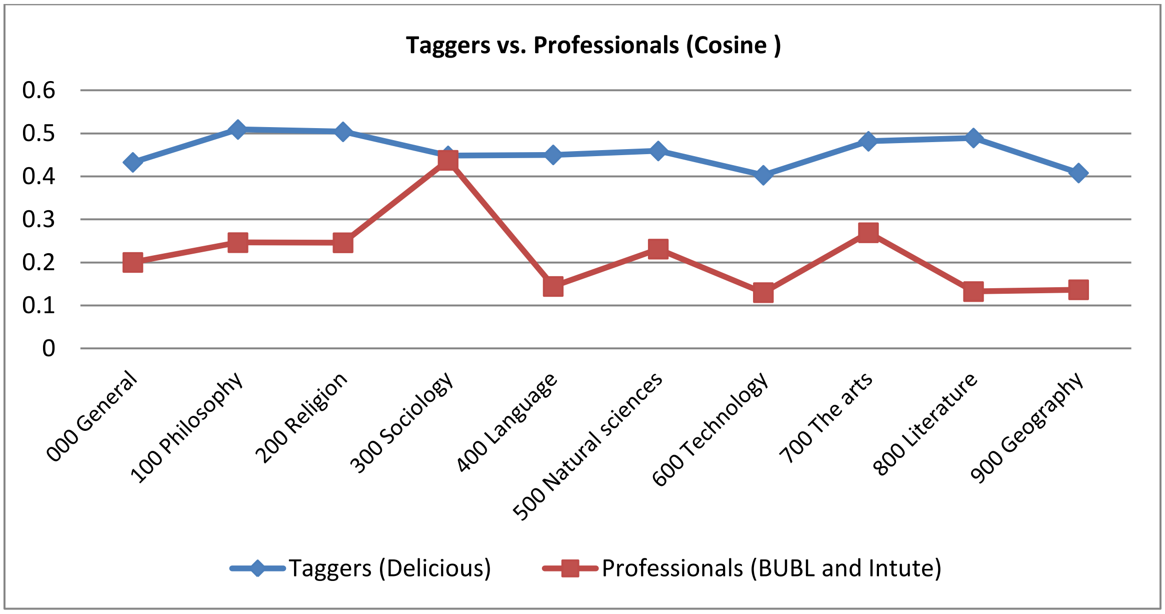


Figure 41. Indexing consistency between taggers (Delicious) and professionals (BUBL and Intute) using the cosine measure

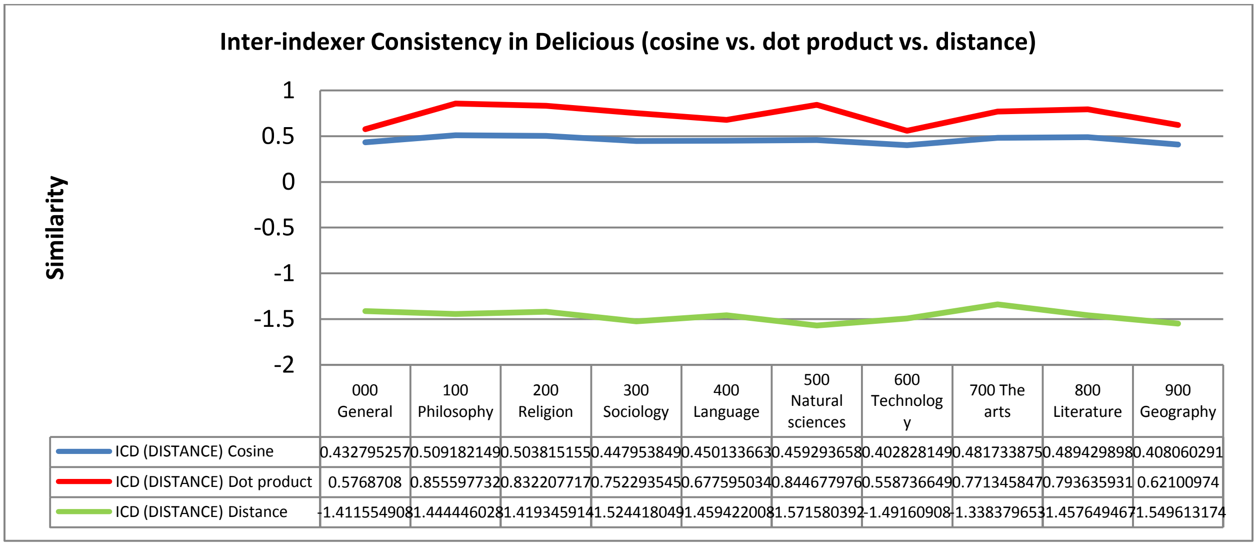
As illustrated in Figures 39, 40, and 41, the similarity for the Sociology subject in two professional groups reached the highest value in three measures. It was shown that both BUBL and Intute located most documents in that subject into “Social sciences” or “Sociology” categories (Table 25). Thus most documents on that subject were simply located in the existing categories.

Table 25. Indexing on Sociology between BUBL and Intute

Social sciences subject	Title	BUBL	Intute
301 Sociology: general resources	Sociological Tour Through Cyberspace, www.trinity.edu/~mkearl/index.html	Social sciences, Sociology	Social sciences, Sociology
310 International statistics	IDB Population Pyramids, International Data Base (IDB) - Pyramids, http://www.census.gov/ipc/www/idb/pyramids.html	Social sciences, Statistics	Social sciences, Statistics, data, Population
330 Economics: general resources	History of Economic Thought, http://cepa.newschool.edu/het/	Social sciences, Economics	Social sciences, Economics, Sociology
355 Military science: general resources	DOD Dictionary of Military Terms, http://www.dtic.mil/doctrine/dod_dictionary/	Social sciences, Military science	Social sciences, Government policy, Military science

4.7 Comparison of Three Similarity Measures

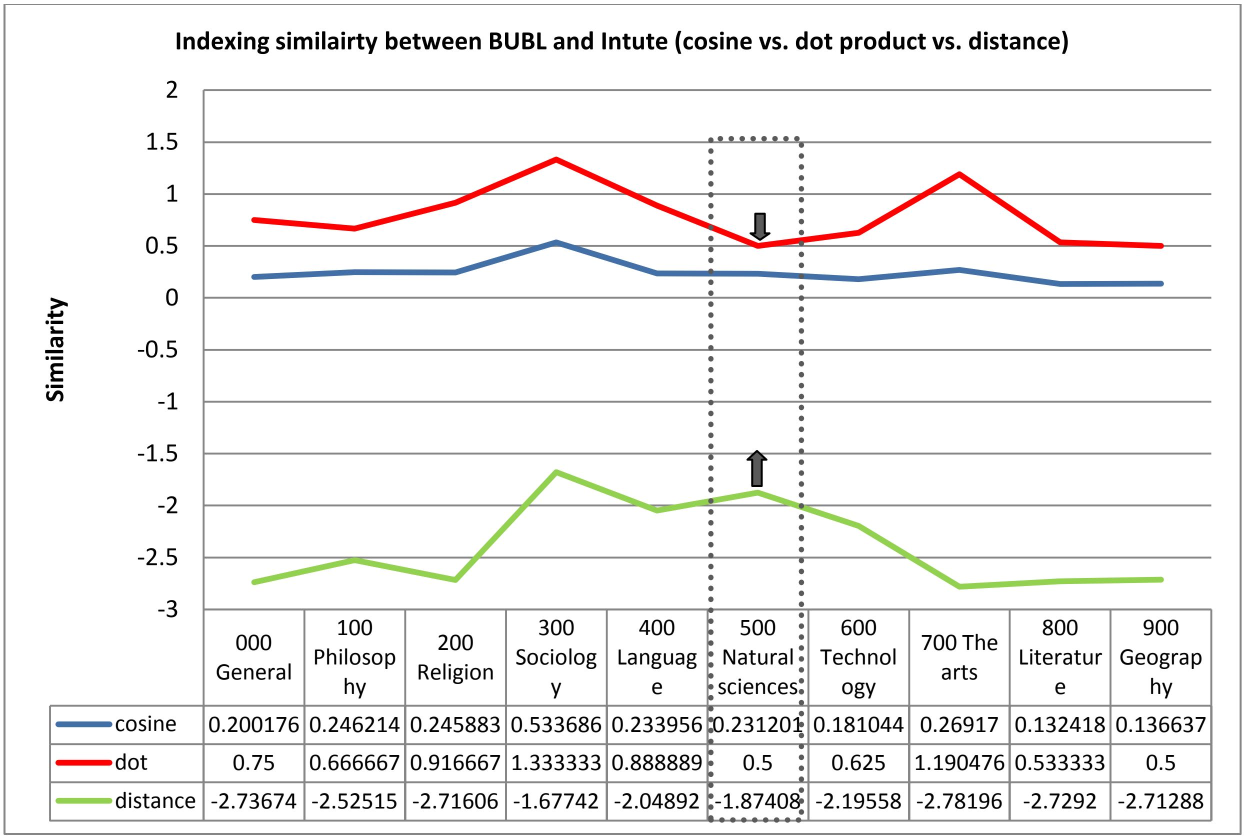
In all cases of consistency, the similarity with the distance measure and dot product measure showed more crooked curves, which means low consistency of similarity among 10 subject categories. In contrast, the similarity with the cosine measure showed a smoother curve over all subjects. Figure 42 illustrates that the similarities of three measures among taggers in Delicious tended to be parallel over all subjects with a slight difference.



	000 General	100 Philosophy	200 Religion	300 Sociology	400 Language	500 Natural sciences	600 Technolog y	700 The arts	800 Literature	900 Geography
ICD (DISTANCE) Cosine	0.432795257	0.509182149	0.503815155	0.447953849	0.450133663	0.459293658	0.402828149	0.481733875	0.489429898	0.408060291
ICD (DISTANCE) Dot product	0.5768708	0.855597732	0.832207717	0.752293545	0.677595034	0.844677976	0.558736649	0.771345847	0.793635931	0.62100974
ICD (DISTANCE) Distance	-1.411554908	1.444446028	1.419345914	1.524418049	1.459422008	1.571580392	-1.49160908	-1.338379653	1.457649467	1.549613174

Figure 42. Inter-indexer consistency in Delicious with three measures

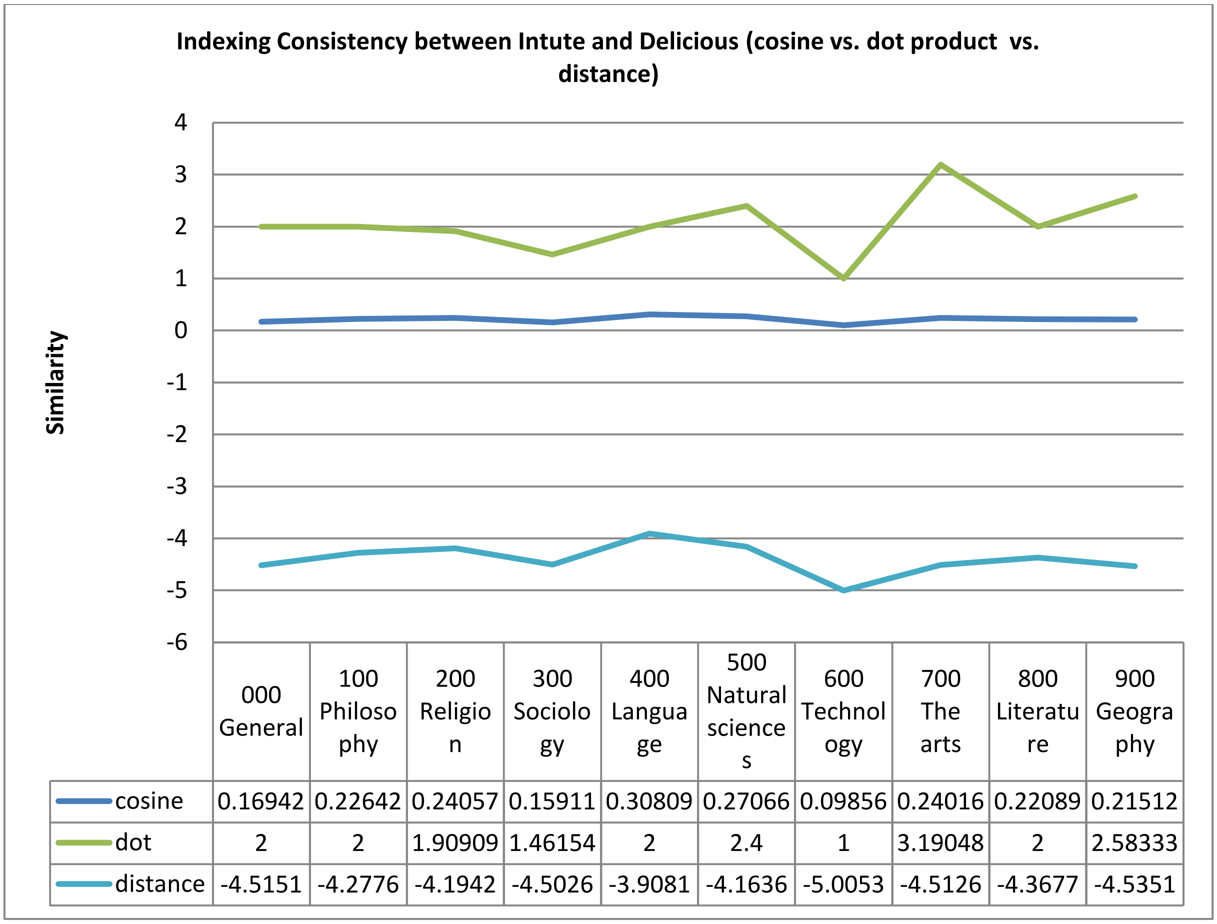
Regarding the similarity between BUBL and Intute professionals, it was revealed that the unique characteristics of three measures caused those somewhat different tendencies in three measures. For example, the similarity on Natural Sciences with the dot product measure reached the lowest point while the similarities with Euclidean distance and cosine showed relatively higher positions in both graphs.



	000 General	100 Philosophy	200 Religion	300 Sociology	400 Language	500 Natural sciences	600 Technology	700 The arts	800 Literature	900 Geography
cosine	0.200176	0.246214	0.245883	0.533686	0.233956	0.231201	0.181044	0.26917	0.132418	0.136637
dot	0.75	0.666667	0.916667	1.333333	0.888889	0.5	0.625	1.190476	0.533333	0.5
distance	-2.73674	-2.52515	-2.71606	-1.67742	-2.04892	-1.87408	-2.19558	-2.78196	-2.7292	-2.71288

Figure 43. Indexing similarity between BUBL and Intute professionals with three measures

In terms of the similarity between Delicious taggers and Intute professionals, Euclidean distance and cosine similarity measures tended to be parallel over all subjects. The Arts subject was the highest point in the dot product measure while in Euclidean distance and cosine measures, the Language subject reached the highest position in both graphs.



	000 General	100 Philosophy	200 Religion	300 Sociology	400 Language	500 Natural sciences	600 Technology	700 The arts	800 Literature	900 Geography
cosine	0.16942	0.22642	0.24057	0.15911	0.30809	0.27066	0.09856	0.24016	0.22089	0.21512
dot	2	2	1.90909	1.46154	2	2.4	1	3.19048	2	2.58333
distance	-4.5151	-4.2776	-4.1942	-4.5026	-3.9081	-4.1636	-5.0053	-4.5126	-4.3677	-4.5351

Figure 44. Indexing consistency for Intute professionals and Delicious taggers using three measures

The different patterns resulting from three measures can be explained by revisiting the formulas and characteristics of the three measures. As described in Section 3.3.2 Analysis of inter-indexer/tagger consistency, the adjusted Euclidean distance formula is expressed by:

$$Similarity\ (A, B) = -Dist(A, B)$$

Dot product based similarity is represented by

$$Similarity\ (A, B) = A \cdot B$$

(A · B is the dot product of vectors)

Also, the cosine similarity (θ) is represented using a dot product and magnitude as:

$$cos\theta = Similarity(A, B) = \frac{A \cdot B}{|A||B|}$$

$|A|$: vector norm of A

$|B|$: vector norm of B

θ: angle between vector A and vector B

Table 26. Comparison of similarity on Natural Sciences with three measures between BUBL and Intute

Natural Science	Title	BUBL	Intute	Distance	Dot product	Cosine	Total terms	Common terms
500 Natural sciences: national centres	National Science Foundation, http://www.nsf.gov/	Natural sciences	Engineering, Physical sciences	-1.73	0	0	3	0
510 Mathematics, general resources	MathSciNet: http://www.ams.org/mathscinet/	Natural sciences, Mathematics	Mathematics, Computer science	-1.41	*1*	*0.5*	4	1
520 Astronomy, general resources	Astronomy Picture of the Day, http://antwrp.gsfc.nasa.gov/apod/astropix.html	Natural sciences, Astronomy	Astronomy	-1	*1*	*0.707*	3	1
550 Earth sciences	GeoGuide, http://www.geo-guide.de/	Natural sciences, Mathematics, Earth sciences	Geography, environment, Physical sciences, Earth sciences	-2.2	*1*	*0.289*	7	1

In general, the magnitude of a vector becomes bigger as it has more elements. That is, the magnitude is proportional to the number of terms that the indexer tagged. In the Euclidean distance measure, $Dist(A, B)$ increases when there are more differences in each element. Thus, in the adjusted distance formula, a value ($-Dist(A, B)$) decreases. That is, when indexers assign more different terms, the similarity is smaller. On the other hand, in the dot product measure, the value of similarity is represented by the number of common terms between two indexers (see Table 26), and the magnitude of the vector is not considered. Cosine similarity is proportional to the angle between two vectors. Therefore, in both the Euclidean distance and cosine similarity measures, values will be reduced when the total number of terms assigned is larger. These unique characteristics of three measures caused those different tendencies in similarity depicted in Figures 42, 43, and 44.

As seen in Table 26, there was a slight difference between the dot product measure and the other two measures, Euclidean distance and cosine, depending on the number of index terms. For example, when there was only one common term between BUBL and Intute terms, the value of the dot product measure generated the same value, “1” for all cases regardless of the total number of terms while the values of both Euclidean distance and cosine similarity measures varied depending on the total number of terms. In both measures, the more terms, the lower values when the number of common terms is the same. Accordingly, the similarity on Natural Sciences with the dot product measure reached the lowest point while the similarities with Euclidean distance and cosine showed higher positions in both graphs (see Figure 43).

In terms of the Arts subject, Tables 27 and 28 also illustrate that both Euclidean distance and cosine measures varied depending on the total number of terms when the number of common terms is the same.

Table 27. Comparison of similarity on Arts with three measures between BUBL and Intute

Arts	Title	BUBL	Intute	Distance	Dot product	Cosine	TT[*]	CT[**]
700 The arts, general resources	Art deadlines list : http://artdeadlineslist.com	Arts	art, Awards, Competitions, performing_arts, Cross_disciplinary_studies, Collections, exhibitions	-2.449	1	0.183	8	1
700 Fine and decorative arts, general resources	Arts in Context http://www.artincontext.org/	arts, Fine_arts, decorative_arts	performing_arts, art, Cross_disciplinary_studies, Collections, exhibitions, Museum, gallery, exhibition, images, Visual_arts, contemporary_arts	-3.317	1	0.183	14	1
703 Fine and decorative arts, dictionaries and encyclopaedias	Artcyclopedia: http://www.artcyclopedia.com/	arts, Fine_arts, decorative_arts, dictionaries, encyclopaedias	performing_arts, art, Cross-disciplinary_studies, images, Architecture, planning, Architects ,Visual_arts, Art_history, Artists Contemporary_arts, photography, Photographers	-3.873	1	0.129	18	1
708 Art galleries and museums in the UK	Warhol, http://www.warhol.org/	arts, art galleries, museums	Communication, media_studies, Film_studies, performing_arts, art, Cross-disciplinary_studies, Collections, exhibitions, Museums, galleries, performing_arts, Visual_arts, Art_history, Artists, Painting, Painters	-3.742	2	0.298	19	2
708 Art galleries and museums in the US	The Metropolitan Museum of Art: http://www.metmuseum.org/	arts, galleries, museums	performing_arts, art, Design, 3D design, Museums, galleries, Visual_arts, Art_history, International, Visual_arts, Fine, contemporary arts, Museums, Cross-disciplinary_studies, Collections, exhibitions, The americas, United States, New York	-4.243	2	0.265	22	2
709 History of art	Futurism: http://www.unknown.nu/futurism/	arts, art_history	performing_arts, art, Visual_arts, Art_history, Periods, styles, movements, 20th century, Futurist Modern_languages, area_studies, Europe, Europe by region / Country, Western, Europe, Italy, Art / architecture history, Area / diaspora_studies	-3.742	2	0.354	20	2

*. TT means the number of total terms

**. CT means the number of common terms

As shown in Figure 44, regarding the consistency between Delicious and Intute, the similarity with the dot product measure looked considerably different from the similarities with the other two measures: Euclidean distance and cosine. It can be explained in Table 28 which includes examples of cases when the similarity with the dot product measure increased e.g., from 6 to 7, the similarity with both Euclidean distance and cosine measures rather decreased, e.g., Euclidean distance (from -4.0 to -4.796 or -4.359) and cosine (from 0.447 to 0.382 or 0.424).

Table 28. Indexing on Arts with three measures between Intute vs. Delicious

Arts	Title	Intute	Delicious	Distance	Dot product	Cosine	TT*	CT**
700 The arts, general resources	Art deadlines list : http://artdeadlineslist.com	American, art, mailing_lists, competitions, events, scholarships, contests	art, grants, deadlines, contests, resources, opportunities, competitions, photography, artists, competition, jobs, reference, news, list, design, arts , resource, funding, submissions, artist	-4.123	3	0.283	27	3
700 Fine and decorative arts, general resources	Art in Context http://www.artincontext.org/	fine_art, art_history, artists, photography, drawing, painting, performing_arts, art, video_art, textiles, sculpture, Film_studies, performing_arts, art, drawing, design, ceramics, architecture, images, exhibitions, art dealers, art galleries, museums, visual_works	art, artists, reference, museum, database, gallery, galleries, directory, resources, design, links, exhibitions, news, photography, dealers, architecture, databases	-4.796	7	0.382	41	7

*. TT means the number of total terms
**. CT means the number of common terms

Table 28 (cont.)

703 Fine and decorative arts, dictionaries and encyclopaedias	Artcyclopedia: http://www.artcyclopedia.com/	photography, sculpture, painting (image-making), art galleries, galleries (display spaces), museums, sculptors, photographers, artists, painters (artists), indexes (reference sources), decorative arts, architecture	art, reference, encyclopedia, artists, images, search, culture, history, research, resources, museum, education, gallery, painting, photography, design, database, artist	-4.899	3	0.209	31	3
708 Art galleries and museums in the UK	Warhol, http://www.warhol.org/	Andy_Warhol_ Museum, art_galleries, artists, museums, galleries, painters, artists, Pop, popular_culture, Warhol, Andy	art, warhol, museum, pittsburgh , design, artists, photography, culture, artist, museums, gallery, popart, pop, andy, reference, fashion, education, graphics, music	-4	6	0.447	30	6
708 Art galleries and museums in the US	The Metropolitan Museum of Art: http://www.metmuseum.org/	libraries, buildings, armors, museums, art_galleries, exhibitions, painting, photography, sculpture, New_York_City, drawing, Metropolitan_Museum_of_Art, NY	art, museum, nyc, museums, culture, history, gallery, travel, design, education, reference, met, usa, images, photography, research, metropolitan	-4.899	3	0.202	30	3
709 History of art	Futurism: http://www.unknown.nu/futurism/	Futurist, art, artists, modern_Italian_styles, movements, Art_history, manifestoes, Italy, Balla, Giacomo, Boccioni, Umberto,	futurism, art, history, manifesto, philosophy, reference, literature, architecture, design, music, theory, culture, future , technology, artist, modernism, archive, painting, resource	-5.099	3	0.194	31	3

Table 28 (cont.)

720 Architecture	American Institute of Architects, http://www.aia.org/	Western_Association, Architects, American_Institute_of Architects, AIA, buildings, architecture, architectural_ education, architectural_training, competitions, awards, directories, conferences, design, construction, architects, United_States, American	architecture, design, aia, architects, reference, professional, organization, institute, art, construction, building, associations, usa, american, green, engineering, business, marketing	-4.359	7	0.424	35	7
796 Sport and outdoor activities	International Rugby Board - Home , http://www.irb.com	equipment, regulations, sports_organizations, football, international_ organizations, rugby,	rugby, sport, irb, world, international, news, reviews, education, fun, game, england , cup, abs	-4.243	1	0.109	19	1

In the Arts subject, there were many more terms assigned than other subjects, which resulted in low similarity with Euclidean distance and cosine measures. The similarity with the dot product measure, which is not affected by the total number of terms, thus demonstrated relatively high consistency (see Figures 43 and 44).

4.8 Summary and Discussion

Using three measures (Euclidean distance, dot product, and cosine), it was demonstrated that there was greater consistency over all subjects among taggers in Delicious when compared to indexing consistency between two groups of professionals (BUBL vs. Intute). When comparing tagging in Delicious and professional indexing in Intute, indexing consistency varied by subject area.

Through examination of specific examples in different subject areas, it was shown that the unique characteristics of the three measures affected the values computed for indexing consistency. In both the Euclidean distance and cosine similarity measures, the values of similarity were reduced when the number of common terms was the same but the total number of terms assigned was larger. However, in the dot product measure, when there was only one common term, the similarity was always "1" regardless of the total number of terms. So, when more emphasis is placed on the "exactness" of similarity, the dot product measure would be more appropriate. In contrast, when the "difference" between two group's perspectives is of greater interest, the Euclidean distance or cosine similarity measure would be more suitable. This result also would contradict a natural argument that indexing at Delicious would reach more agreement because many tags were applied to most documents. The results here suggest that the observed level of agreement is not simply due to an increased number of common terms (i.e., matching opportunity) but due to a large number of assigned terms.

On the other hand, it is important to note that the concept of inter-indexer consistency discussed in this chapter is different from that of "*indexer-requester consistency*" introduced by Cooper (1969) which is discussed in Chapter 2. Literature Review, 2.4.1 Inter-indexer consistency. The concept of inter-indexer consistency is concerned with the agreement of terms among different indexers, while indexer-requester consistency is associated with the correspondence between two groups of terms, e.g., terms used in searching by requesters vs. terms assigned by professional indexers. To more fully assess indexing "quality" beyond what was analyzed in this study, assessing the impact of indexing on retrieval performance would be helpful.

CHAPTER 5: TAGGING EFFECTIVENESS

5.1 Overview

This chapter presents the results of the analyses of tagging exhaustivity and tag specificity in relation to tagging effectiveness. Research question 6 was answered:

> Research Question 6) What levels of tagging exhaustivity and tag specificity in Delicious characterize the indexing of web documents?

This chapter discusses the levels of tagging exhaustivity and tag specificity in a social tagging site in indexing web documents. Tag exhaustivity is discussed by associating the degree of indexing consistency. To provide an in-depth analysis of tagging accuracy, BUBL and Intute indexers' descriptions are reviewed and compared with tags at various levels of specificity. This dissertation research employed the Latent Semantic Analysis (LSA) method to examine whether the tags accurately describe documents, i.e., how well those tags represent important concepts related to the documents. LSA with term-to-term comparison (between professional keywords and user tags) was conducted concerning documents that showed the lowest similarity between Intute professionals' keywords and Delicious users' tags. In addition, to investigate statistically whether there is a relationship between tag specificity and LSA values between tags and professionals' keywords, a correlation analysis was conducted.

5.2 Relationship between Indexing Similarity and Tagging Exhaustivity

Figure 45 demonstrates that the average number of tags per documents over all subjects ranges from 2.5 to 3.3 tags, and also provides an analogous plot using median tags instead of mean. In term of the average number of tags, it should be noted that since there were a number of documents tagged by one term or sometimes tags were not in English, the number of tags per documents was not many. The graph indicates that there is not a wide variation among the 10 subject categories.

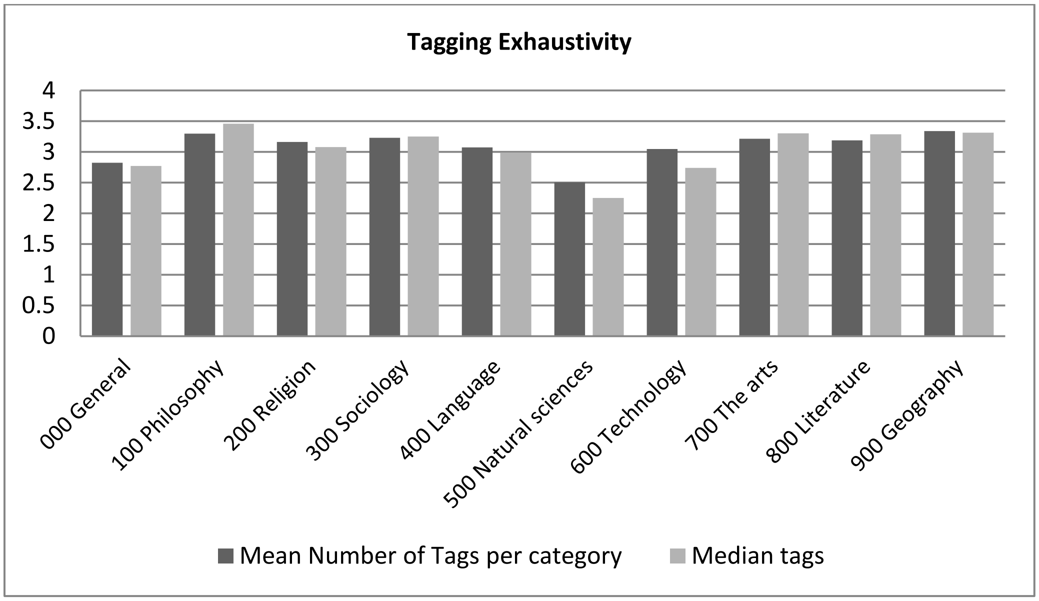
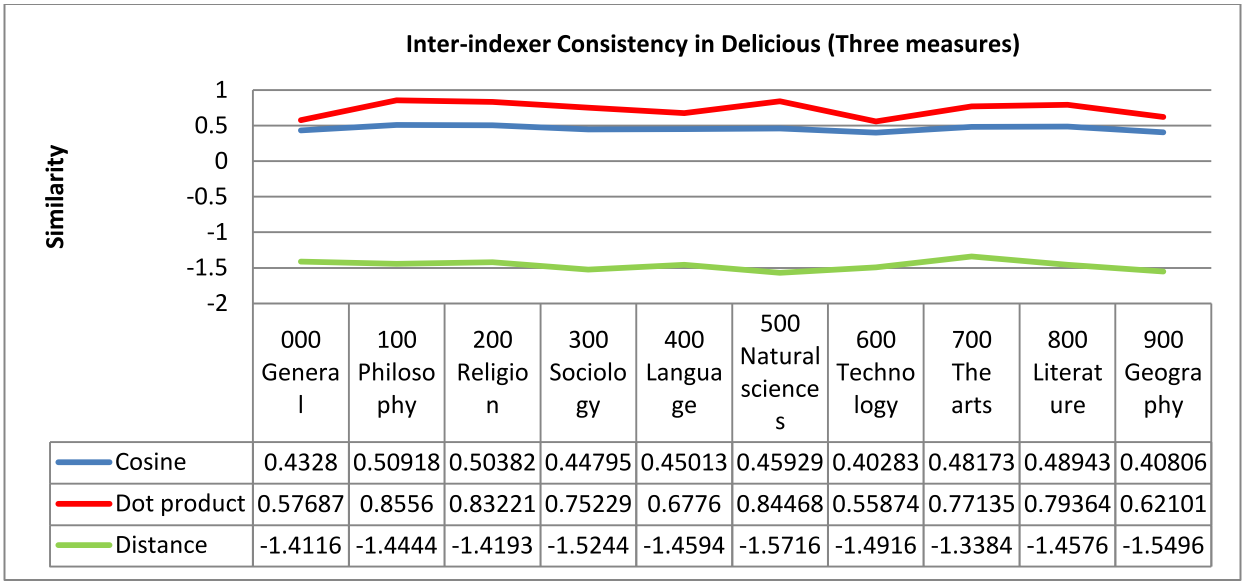


Figure 45. Tagging exhaustivity: Mean number of tags per category

Reviewing the measures of indexing consistency that were analyzed in Chapter 4, for all three measures, the subjects of Philosophy, Religion, Arts, and Literature showed relatively higher indexing similarity in Delicious (Figure 46).



	000 General	100 Philosophy	200 Religion	300 Sociology	400 Language	500 Natural sciences	600 Technology	700 The arts	800 Literature	900 Geography
Cosine	0.4328	0.50918	0.50382	0.44795	0.45013	0.45929	0.40283	0.48173	0.48943	0.40806
Dot product	0.57687	0.8556	0.83221	0.75229	0.6776	0.84468	0.55874	0.77135	0.79364	0.62101
Distance	-1.4116	-1.4444	-1.4193	-1.5244	-1.4594	-1.5716	-1.4916	-1.3384	-1.4576	-1.5496

Figure 46. Indexing consistency in Delicious (three measures)

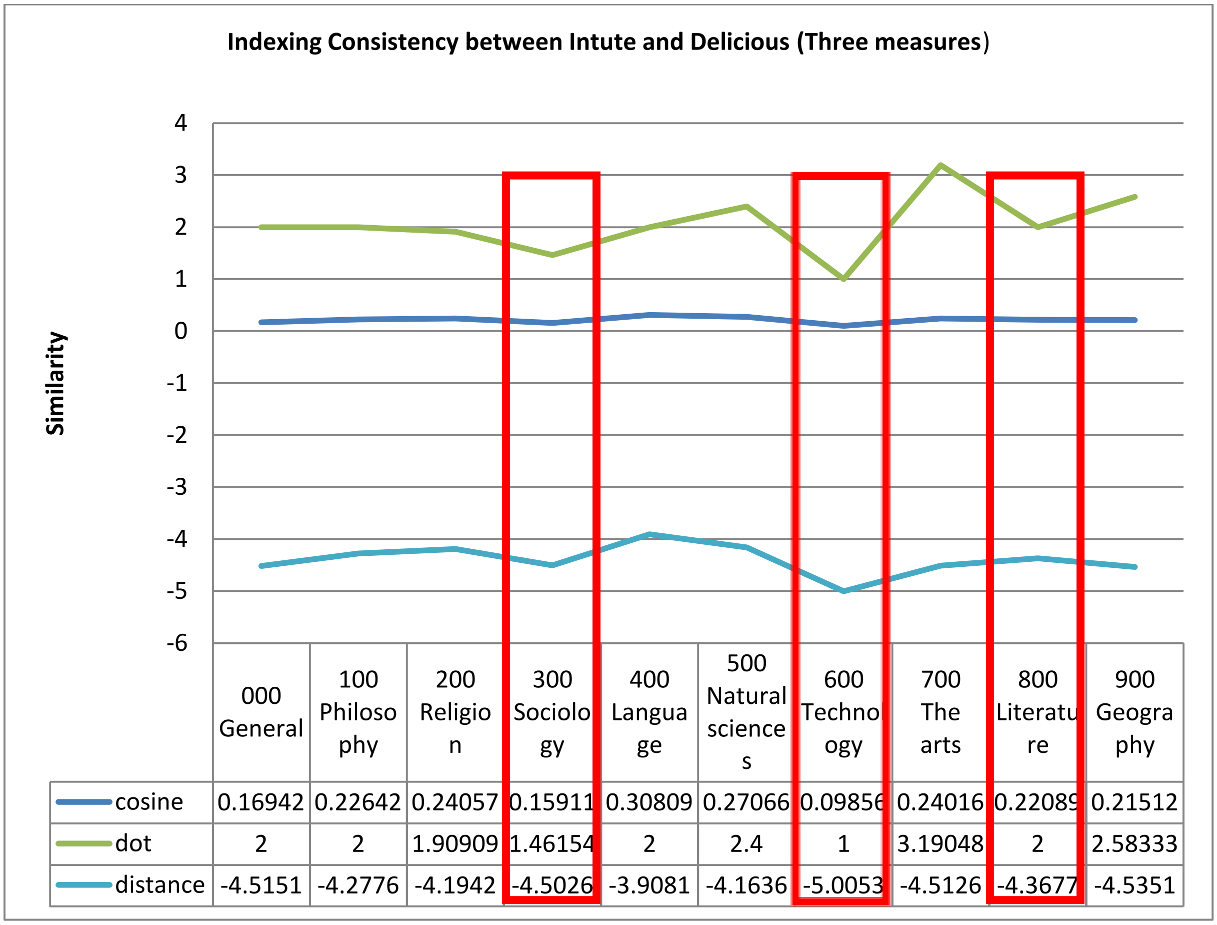
As discussed in Section *3.3.3.2 Tagging Exhaustivity*, higher exhaustivity leads to higher recall. Even though high recall does not necessarily imply high quality of indexing, it is interesting to note that regarding the subject categories with higher indexing similarity (Philosophy, Religion, Arts, and Literature), the tagging exhaustivity also reached relatively higher values (more than three tags per document) compared to other subject categories (Figure 45 and 46).

5.3 Relationship between Indexing Accuracy and Tag Specificity

5.3.1 Tag specificity

The term specificity implies the ability of the index terms to differentiate a document from other documents. If a term is frequently used for describing documents, the term is not a good differentiator between documents. To put it another way, if there are many documents where a tag is assigned, the tag is not a good differentiator, and on the contrary, if there are few documents where a tag is assigned, the tag is a good differentiator. On the other hand, Luhn (1958) and Salton et al. (1975b) suggested that mid-range terms are the best indicators of topicality, and that very frequent and very infrequent terms are weaker discriminators.

Thus, this research investigated the quality and accuracy of terms in various levels of specificity value. This research also examined term ranks calculated by the frequency of tags assigned to a document. To analyze tag specificity, we investigated three subject categories (Social Sciences, Literature, and Technology), which showed relatively low indexing similarity between tags and controlled keywords (see Figure 47).



	000 General	100 Philosophy	200 Religion	300 Sociology	400 Language	500 Natural sciences	600 Technoogy	700 The arts	800 Literature	900 Geography
cosine	0.16942	0.22642	0.24057	0.15911	0.30809	0.27066	0.09856	0.24016	0.22089	0.21512
dot	2	2	1.90909	1.46154	2	2.4	1	3.19048	2	2.58333
distance	-4.5151	-4.2776	-4.1942	-4.5026	-3.9081	-4.1636	-5.0053	-4.5126	-4.3677	-4.5351

Figure 47. Indexing consistency between Intute and Delicious (three measures)

As we discussed in Section 4.5 *Consistency between tagging in Delicious and professional indexing in Intute*, low similarity in Sociology and Literature between Delicious tags and Intute controlled keywords revealed that tags included additional access points with many newly-coined terms such as ebook, online, web, web 2.0, e-guides, e-learning and cyberspace which reflect more accurate descriptions of the web documents (see Table 22). Therefore, it is worthwhile to see if those tags are really powerful to differentiate each document among other documents, and if those tags represent the high degree of quality.

On the other hand, we investigated the Technology subject which also showed relatively low indexing similarity between tags and keywords (Figure 44). As discussed in Section 4.5, Technology showing low consistency indicated that there were different levels of indexing between Intute indexers and Delicious taggers. For example, regarding the document 610 *Medical sciences, medicine*, Intute keywords tend to be broader terms, i.e., “disease” and “patient education,” but Delicious tags consist of terms in various semantic relationships, e.g., broader terms or narrower terms. As shown in Table 29, tags in the Technology subject include “health”, “medical”, “medicine”, “drugs”, “healthcare” etc. In the Library of Congress Subject Headings (LCSH), a term “health” is represented as “narrower terms” of the term “medicine”. The term “healthcare” does not exist in the LCSH, but an alternative term “medical care” is represented as a narrower term of the term “public health” which is a narrower term of “health”.

Thus, in order to provide an in-depth analysis of tags in this discussion, the specificity values of those tags were analyzed. The analysis of tag specificity was conducted on documents showing the lowest indexing similarity between Delicious tagging and Intute indexing in three subject areas, i.e., Social sciences, Literature, and Technology.

Table 29. Subject areas showing lowest consistency between Intute and Delicious

	Document	Title	Intute keywords	Delicious tags
1	**Social Sciences (301 Sociology: general resources)**	Sociological Tour Through Cyberspace, www.trinity.edu/~mkearl/index.html	death, euthanasia, families, homicide, mass media, time	sociology, links, resources, research, culture, web, science, cyberspace, technology, web2.0, writing, social, internet, politics, reference, statistics
2	**Social Sciences (370 Education)**	Excellence Gateway, http://excellence.qia.org.uk/	numeracy, learning, key skills, literacy	resources, education, e-learning, qia, teaching, learning, learning_resource, agency, quality, materials, jobs, e-guides, curriculum
3	**Literature (808.8 Literature: general collections)**	Google Book Search, http://books.google.com/	writers, authors, books, search engines	books, google, search , e-books, reference, library, research, tools, literature, search engine, web2.0, education, reading, resources, online, web, database
4	**Literature (820 English, Scottish and Irish literature)**	Cambridge History of English and American Literature, http://www.bartleby.com/cambridge/	literature, poetry, fiction, drama, Renaissance, Restoration, English, American, poets, poems, Anglo Saxon, plays, writings, encyclopedias, history	literature, history, reference, encyclopedia, e-books, books, humanities, research, language, reading, criticism, academic, writing, resources, information, English literature
5	**Technology (610 Medical sciences, medicine)**	MedicineNet, http://www.medicinenet.com/script/main/hp.asp	disease, patient education	health, medical, medicine, reference, drugs, information, education, news, research, healthcare, dictionary, science, search, resources, doctors, diseases, biology
6	**Technology (630 Agriculture and related technologies)**	AgNIC: Agriculture Network Information Center, http://www.agnic.org/	agricultural_sciences, agriculture, agricultural_education, information_center	agriculture, research, food, information, statistics, environment, plants, farming, libraries, international, database, agnic, science, associations, produce, portal, horticulture

5.3.2 Latent Semantic Analysis (LSA) on the conceptual semantics of tags

In addition to the value of tag specificity, this research investigated the accuracy of tags at various levels of specificity as good index terms. This dissertation research examined whether the tags describe documents accurately, i.e., how well those tags represent important concepts related to the documents.

As discussed in Chapter 4, regarding three subject areas (Social science, Literature, and Technology), the factors affecting low similarity of indexing between professionals' keywords and users' tags can be summarized by: (1) tags included additional access points with newly-coined terminology, and (2) there were different levels of indexing between Intute indexers and Delicious taggers, e.g., various semantic relationships such as broader terms or narrower terms.

The focal point in this research is not to criticize the quality of professionals' keywords but to point out the lack of additional access points or complementary terms in controlled vocabularies which are used by professionals. Since the keywords provided by professionals are regarded as accurate terms describing topics of documents, it is worthwhile to see whether there are semantic relations between tags and professionals' keywords for the documents which are described by both tags and keywords. If tags are conceptually similar to professionals' keywords, those tags are also regarded as key terms or good descriptors in describing the document.

Accordingly, Latent Semantic Analysis (LSA) was conducted to investigate to what extent tags are conceptually related to professionals' keywords. High values of latent semantics between tags and professionals' keywords would demonstrate that those tags can be considered to be good index terms. BUBL and Intute indexers' descriptions of each document were also reviewed in order to provide complementary analysis of tagging accuracy and quality.

LSA is concerned with concepts instead of words themselves and does not need an exact match for terms. LSA is a technique in natural language processing, and it analyzes relationships between documents and the terms they contain and word semantics (Deerwester et al., 1990). Based on the mathematical model using statistical computations applied to a large corpus of text, LSA extracts and represents the contextual-usage meaning of words (Landauer and Dumais, 1997). LSA uses a term-document matrix to identify the co-occurrence of terms within a set of documents by correlating semantically related terms that are "latent" in a collection of text. Terms are represented as a vector of document scores. Semantic similarity between terms is measured as the extent to which two terms are applied in the same documents. Instead of using the original document space, in LSA, each term is represented by a 300-dimensional vector of scores on orthogonal factors. Since LSA computes the cosine value between two vectors, the highest value in LSA computation is one. The basic idea of LSA is that if two vectors or terms tend to occur in similar documents, the terms are similar. Table 30 shows the examples of LSA cosine values between two vectors. It shows that the semantic similarity (0.74) between two terms, which are "library" and "book", is higher than the semantic similarity (0.02) between "library" and "beach".

Table 30. Examples of LSA values between two vectors

Vector 1	Vector 2	LSA cosine values
library	book	0.74
library	beach	0.02
library	information	0.30
library	skirt	0.11
library	catalog	0.68

In this dissertation research, in order to compute semantic relatedness between tags and professionals' keywords in terms of a specific document, LSA was performed by using a web-based LSA tool, LSA@CU (http://lsa.colorado.edu/) with the semantic space "General Reading up to 1st year college (300 factors)" *TASA corpus* with one-to-many, term-to-term comparison. The *TASA corpus* contains approximately 10 million words and is a set of short English documents, extracted from novels, newspaper articles, and other sources. The corpus was collected by Touchstone Applied Science Associates, to develop The Educator's Word Frequency Guide (Turney and Littman, 2003).

LSA was conducted for six documents that showed the lowest similarity between Intute professionals' keywords and Delicious users' tags (see Table 29). Figures 48, 50, 52, 54, 56, and 58 show the distributions of the number of documents indexed with each of the top-ranked tags in the Delicious entry for a given document. In this dissertation research, the concept of tag specificity was redefined as "the number of documents described by one tag" (Spärck Jones, 1972; cited by Hassan-Montero and Herrero-Solana, 2006).

In Figures 48, 50, 52, 54, 56, and 58 it should be understood that tag specificity is higher when the number of documents indexed by that tag is smaller. The ordinal numbers inside parentheses indicate the rank associated with tag frequency for that document. Figures 49, 51, 53, 55, 57, and 59 represent the LSA values measured by the one-to-many, term-to-term comparison method of LSA@CU (http://lsa.colorado.edu/) tool. That is, the bar length represents the LSA cosine values between one Delicious tag and Intute professionals' keywords. For example, in Figure 49, the LSA cosine value was measured between a tag "social" and six Intute keywords, e.g., death, euthanasia, families, homicide, mass media, and time (Table 31), and the value was more than 0.25. The LSA cosine value between a tag "resources" and six Intute keywords was about 0.06. These values demonstrate that the tag "social" is conceptually or semantically more similar and related to Intute keywords.

- Document 1: *Sociological Tour Through Cyberspace*

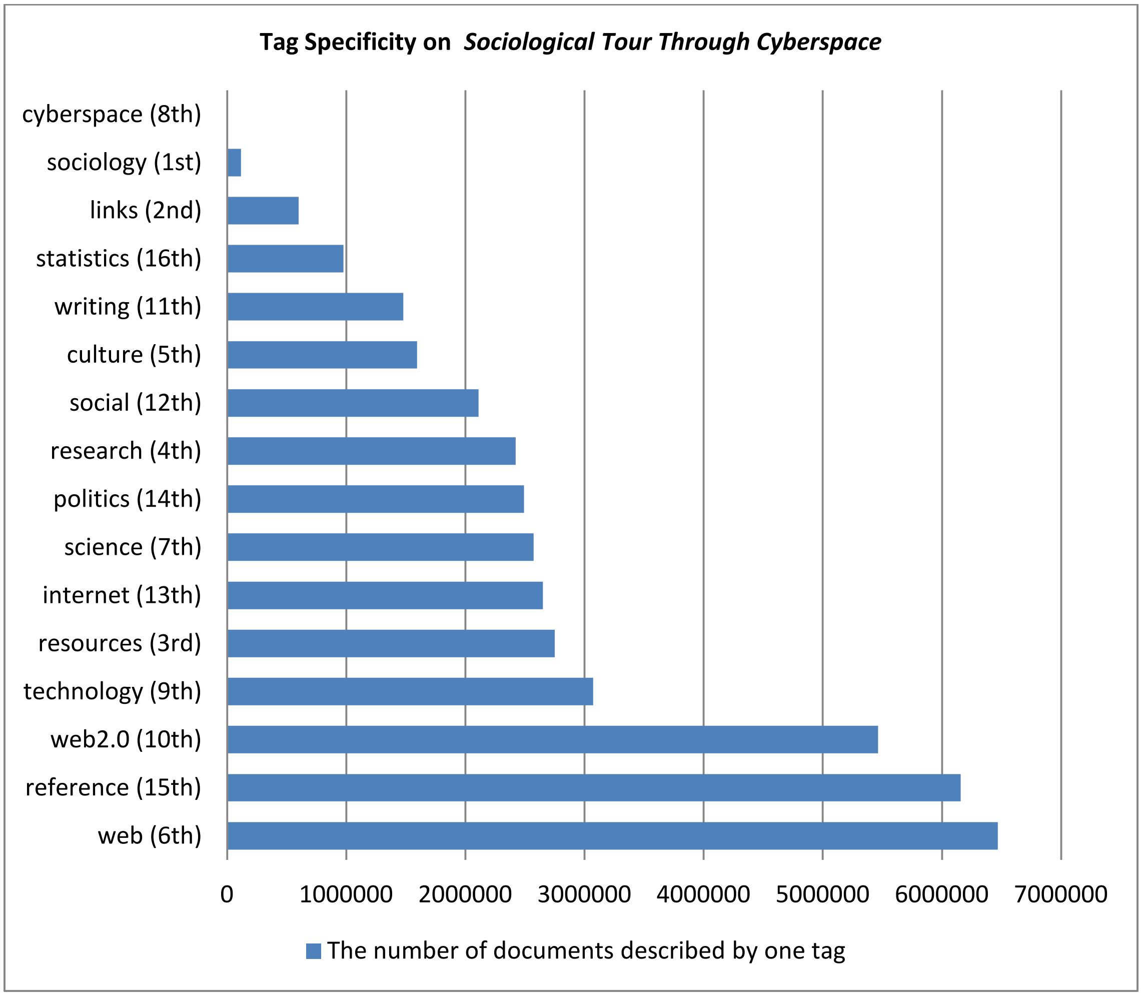
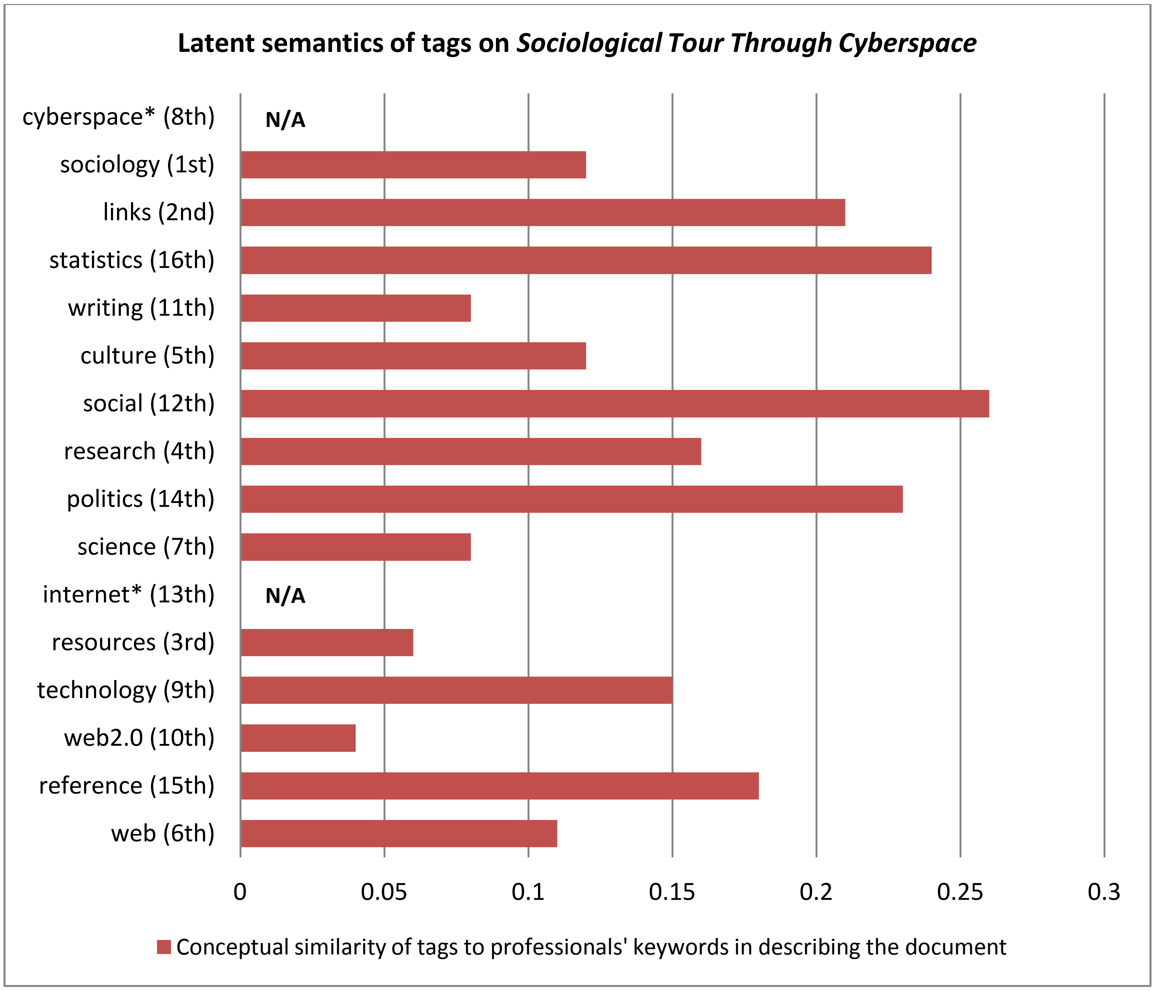


Figure 48. Specificity of tags assigned on the document "*Sociological Tour Through Cyberspace*"



(Note: The terms with an asterisk (*) do not exist in the corpus)

Figure 49. Latent semantic relatedness of tags on professionals’ keywords (the document “*Sociological Tour Through Cyberspace*”)

Table 31. Term-to-term comparison on *Sociological Tour Through Cyberspace*

Document	Title	Intute keywords	Delicious tags
Social Sciences (301 Sociology: general resources)	Sociological Tour Through Cyberspace, www.trinity.edu/~mkearl/index.html	death, euthanasia, families, homicide, mass media, time	sociology, links, resources, research, culture, web, science, cyberspace, technology, web2.0, writing, social, internet, politics, reference, statistics

Indexers in BUBL and Intute describe this document as:

"*Sociological commentary, data analyses, occasional essays, theories and research data.*" (BUBL)

"*Links, commentary, essays and graphs on the sociology's of death and dying, time, mass media, race and ethnicity, and family. The tour was designed for undergraduate classes at Trinity University, San Antonio Texas. As well as links to specific resources, the site includes a couple of search engines*" (Intute)

According to the descriptions provided by both BUBL and Intute, it can be observed that among tags obtaining higher specificity values, the terms, "links" and "social" are good descriptors for the document. These two terms also represent higher LSA values.

- Document 2: *Excellence Gateway*

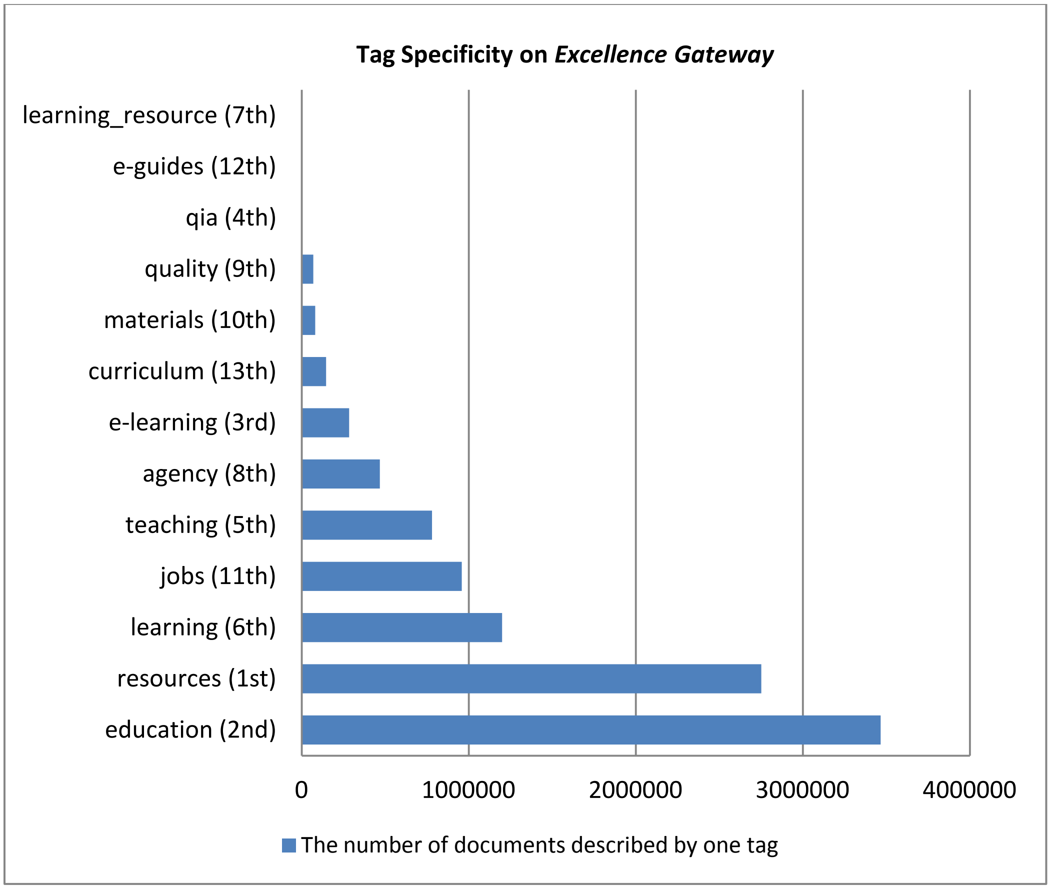
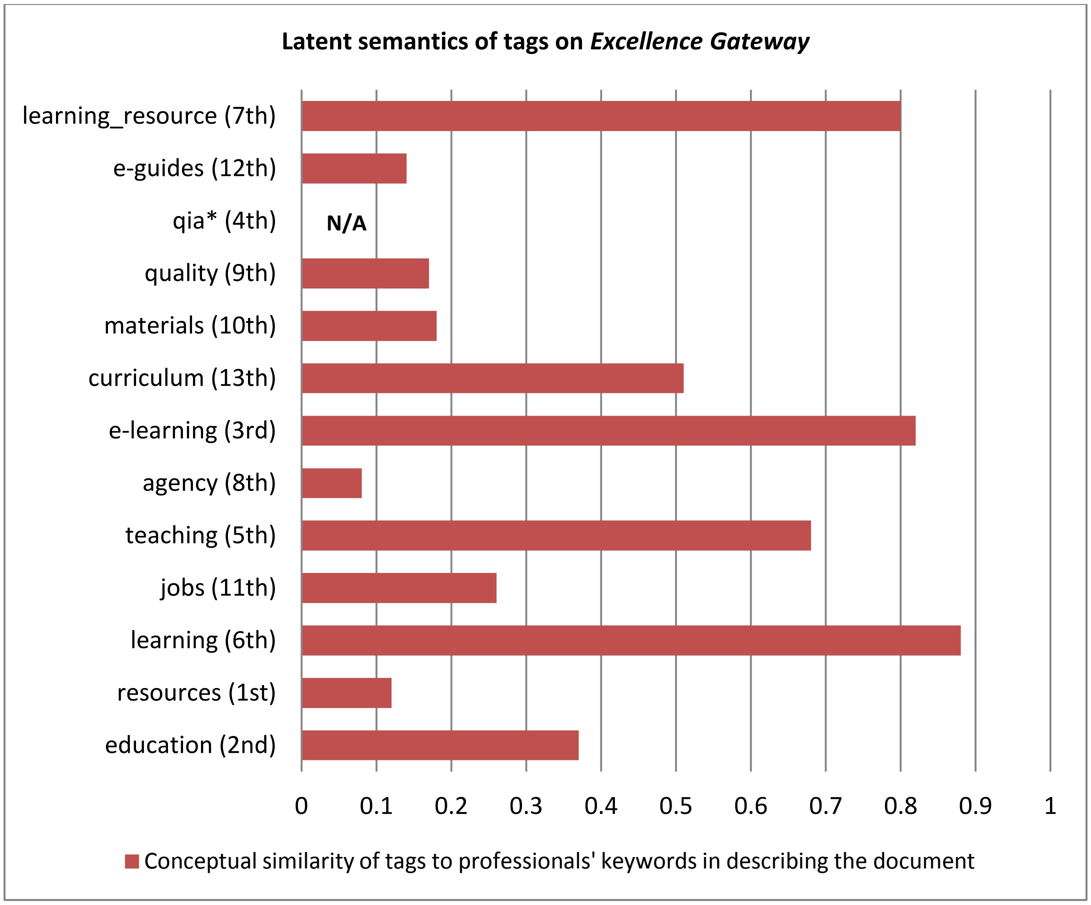


Figure 50. Specificity of tags assigned on the document "*Excellence Gateway*"



(Note: The terms with an asterisk (*) do not exist in the corpus)

Figure 51. Latent semantic relatedness of tags on professionals' keywords (the document "Excellence Gateway")

Table 32. Term-to-term comparison on *Excellence Gateway*

Document	Title	Intute keywords	Delicious tags
Social Sciences (370 Education)	Excellence Gateway, http://excellence.qia.org.uk/	numeracy, learning, key skills, literacy	resources, education, e-learning, qia, teaching, learning, learning resource, agency, quality, materials, jobs, e-guides, curriculum

Indexers in BUBL and Intute describe this document as:

"*Portal for the learning and skills sector, which brings together 50 education websites and up to date resources in one location*" (BUBL)

"*The Excellence Gateway provides access to education information and resources from the UK for education and training practitioners working at all levels in the learning and skills sector. It is funded by the DCSF through the Quality Improvement Agency, who are responsible for enhancing performance in the learning and skills sector. Users can search for news and event information or information resources collated from a number of learning and skills related sites*" (Intute)

The results of LSA showed that the tag "learning resource" in the highest specificity was very conceptually related to Intute keywords (LSA value>0.8) (see Figure 51). Furthermore, BUBL characterizes this document as a "portal for the learning and skills sector" and Intute explains that it is funded through the Quality Improvement Agency (QIA). Accordingly tags such as "e-guides" and "e-learning" which are relatively specific would also be good index terms which exactly describe the document. The tag "qia" also did not exist in the corpus, but is an acronym for the name of the agency, "Quality Improvement Agency".

- Document 3: *Google Book Search*

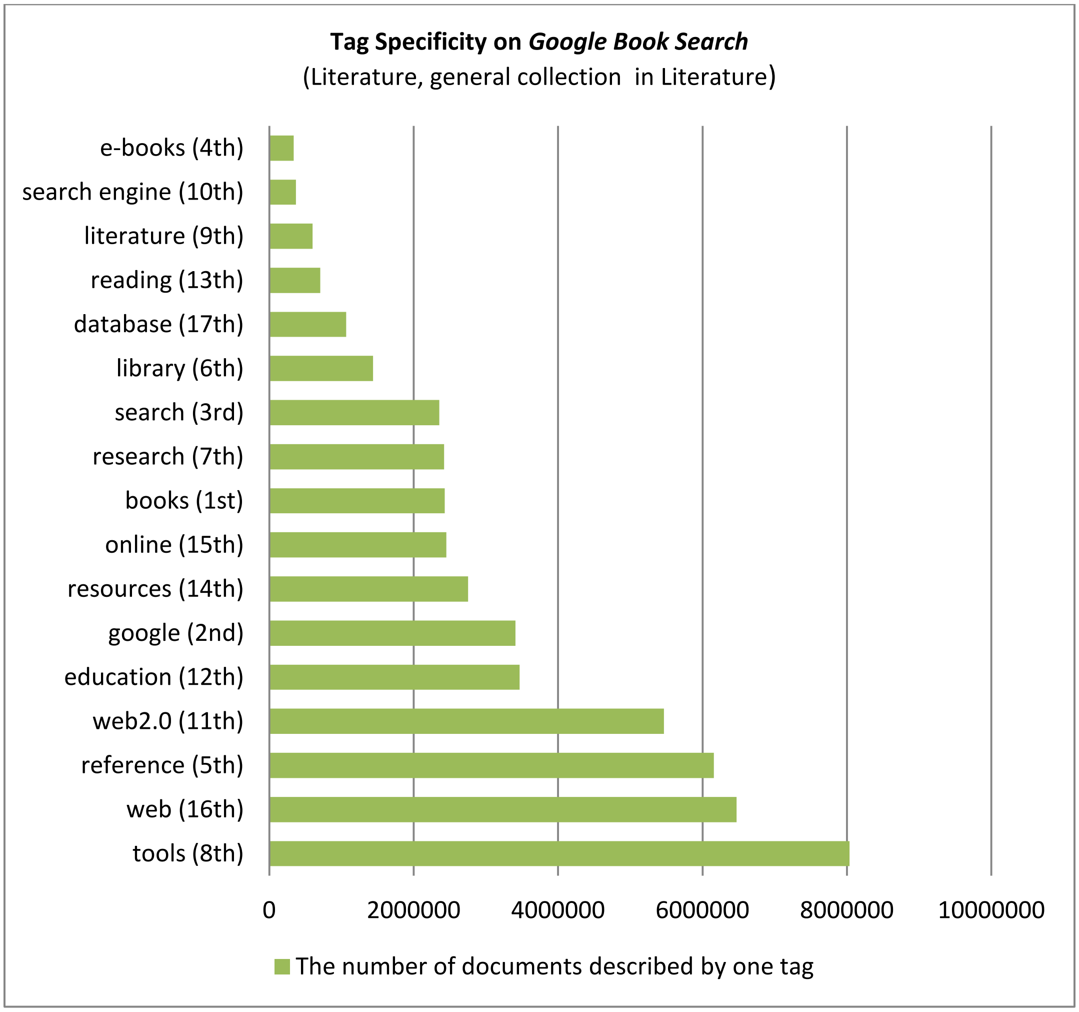


Figure 52. Specificity of tags assigned on the document "*Google Book Search*"

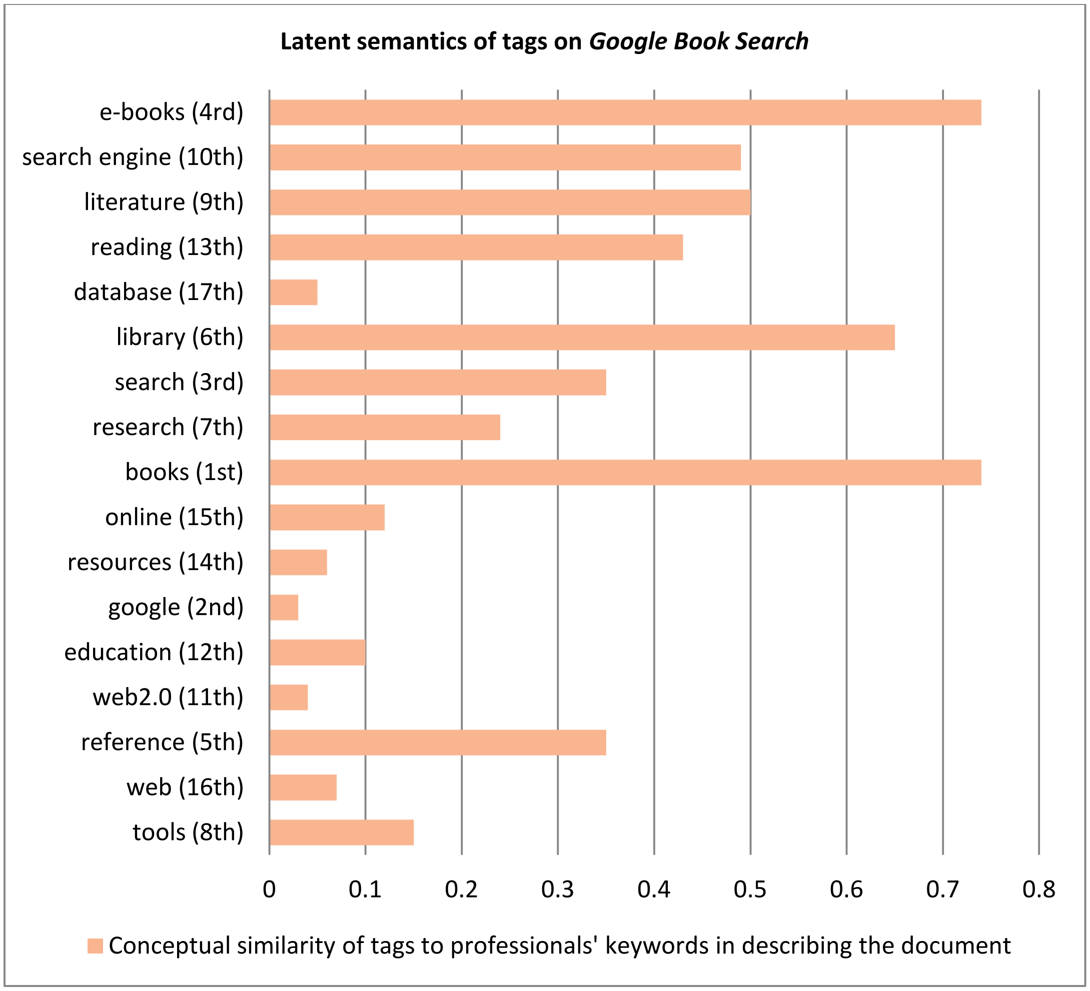


Figure 53. Latent semantic relatedness of tags on professionals' keywords (the document "Google Book Search")

Table 33. Term-to-term comparison on *Google Book Search*

Document	Title	Intute keywords	Delicious tags
Literature 808.8 Literature: general collections	Google Book Search, http://books.google.com/	writers, authors, books, search engines	books, google, search , e-books, reference, library, research, tools, literature, search engine, web2.0, education, reading, resources, online, web, database

The results of LSA showed that tags with relatively higher specificity value were more conceptually related to Intute keywords (Figure 53). The examples of those tags were "ebooks", "search engine", and "literature". BUBL and Intute's descriptions of this document show that these tags clearly express the topics of the document:

"*Google Book Search helps users search within and discover books using digital page-scans presented in a simple ebook format*" (BUBL)
"*Google Book Search is a specialist online search service from Google, providing free online access to selections from a large and rich collection of books*" (Intute)

As seen above in descriptions by BUBL and Intute, especially, the tag, "ebooks (4th rank)", as relatively new terminology, can be regarded as a good index term.

- Document 4: *Cambridge History of English and American Literature*

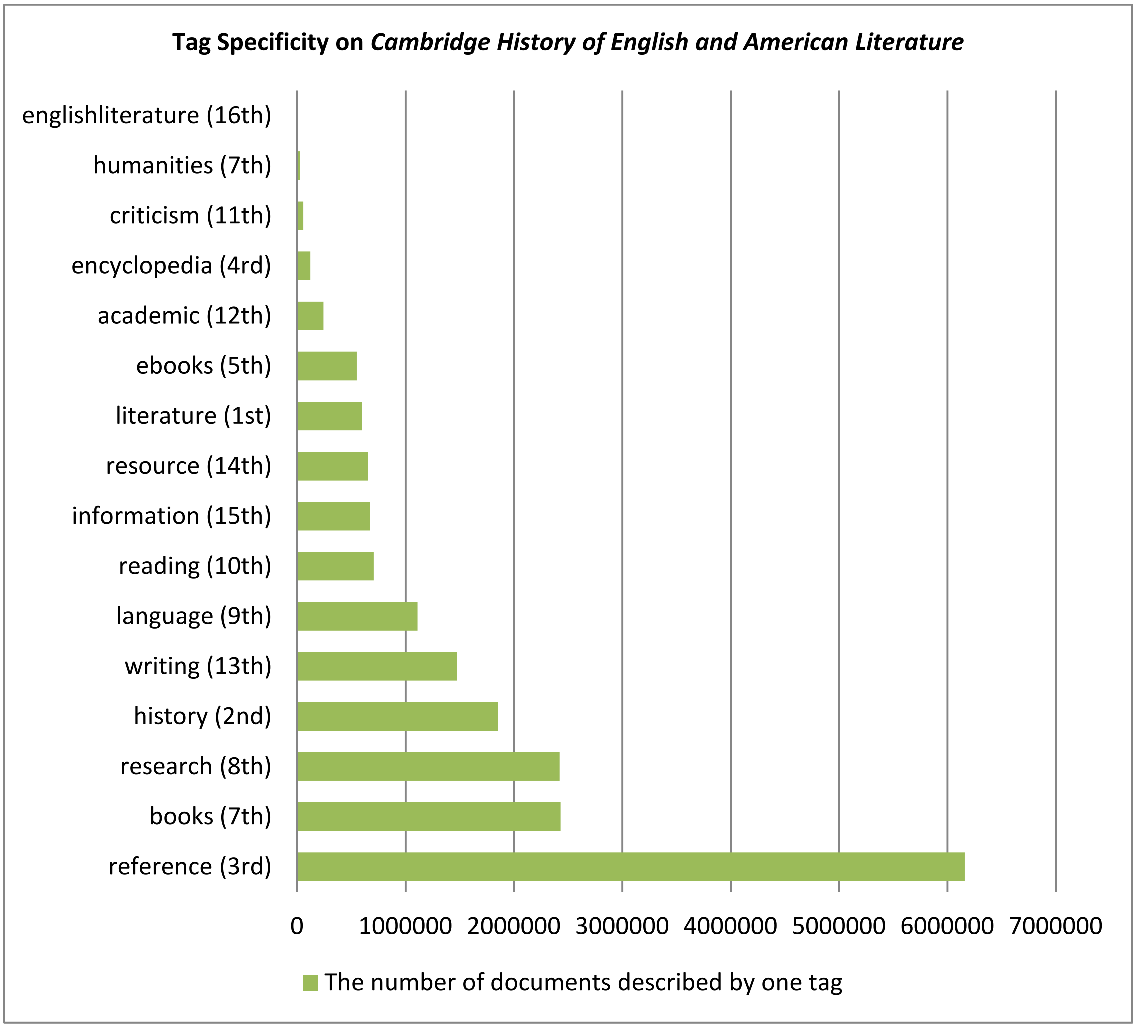


Figure 54. Specificity of tags assigned on the document "*Cambridge History of English and American Literature*"

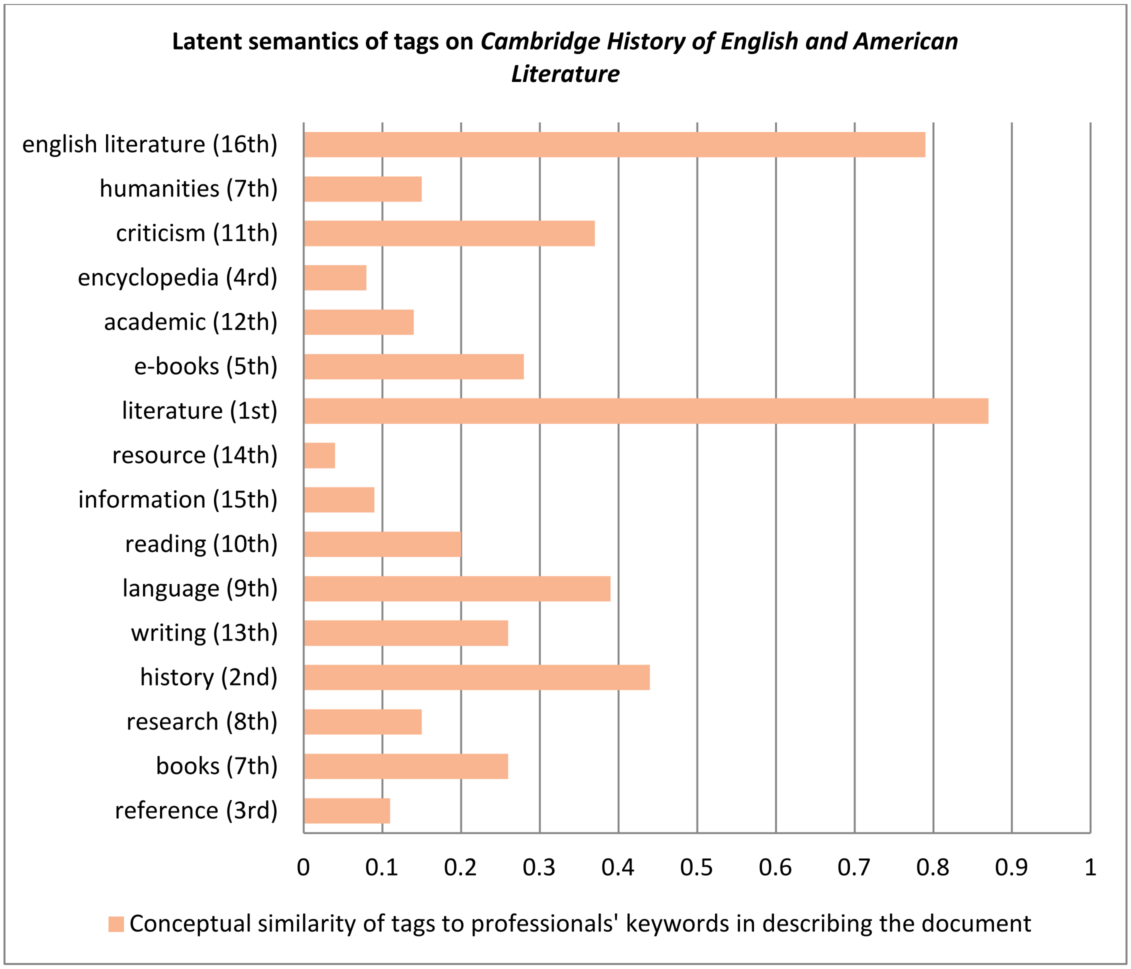


Figure 55. Latent semantic relatedness of tags on professionals' keywords
(the document "Cambridge History of English and American Literature")

Table 34. Term-to-term comparison on *Cambridge History of English and American Literature*

Document	Title	Intute keywords	Delicious tags
Literature 820 English, Scottish and Irish literature	Cambridge History of English and American Literature, http://www.bartleby.com/cambridge/	literature, poetry, fiction, drama, Renaissance, Restoration, English, American, poets, poems, Anglo Saxon, plays, writings, encyclopedias, history	literature, history, reference, encyclopedia, e-books, books, humanities, research, language, reading, criticism, academic, writing, resources, information, English literature

Indexers in BUBL and Intute describe this document as:

"*Encyclopedia tracing the history of literary movements in 18 volumes*" (BUBL)

"*The Cambridge History of English and American Literature is the online full-text of this work, originally printed between 1907 and 1921*" (Intute)

BUBL and Intute describe the document as “encyclopedia” “the online full-text” about English literature from the Middle Ages to the 20th century. It can be recognized that the tag, “ebooks” is a new term as well as taggers’ preferred term to “online full-text” (Figure 54). Also, its high rank (5th) shows many social taggers adopted the term to explain the document. Also, the tag “englishliterature” in the highest specificity value is regarded as a good index term by showing high semantic relatedness to professionals’ keywords.

- Document 5: *MedicineNet*

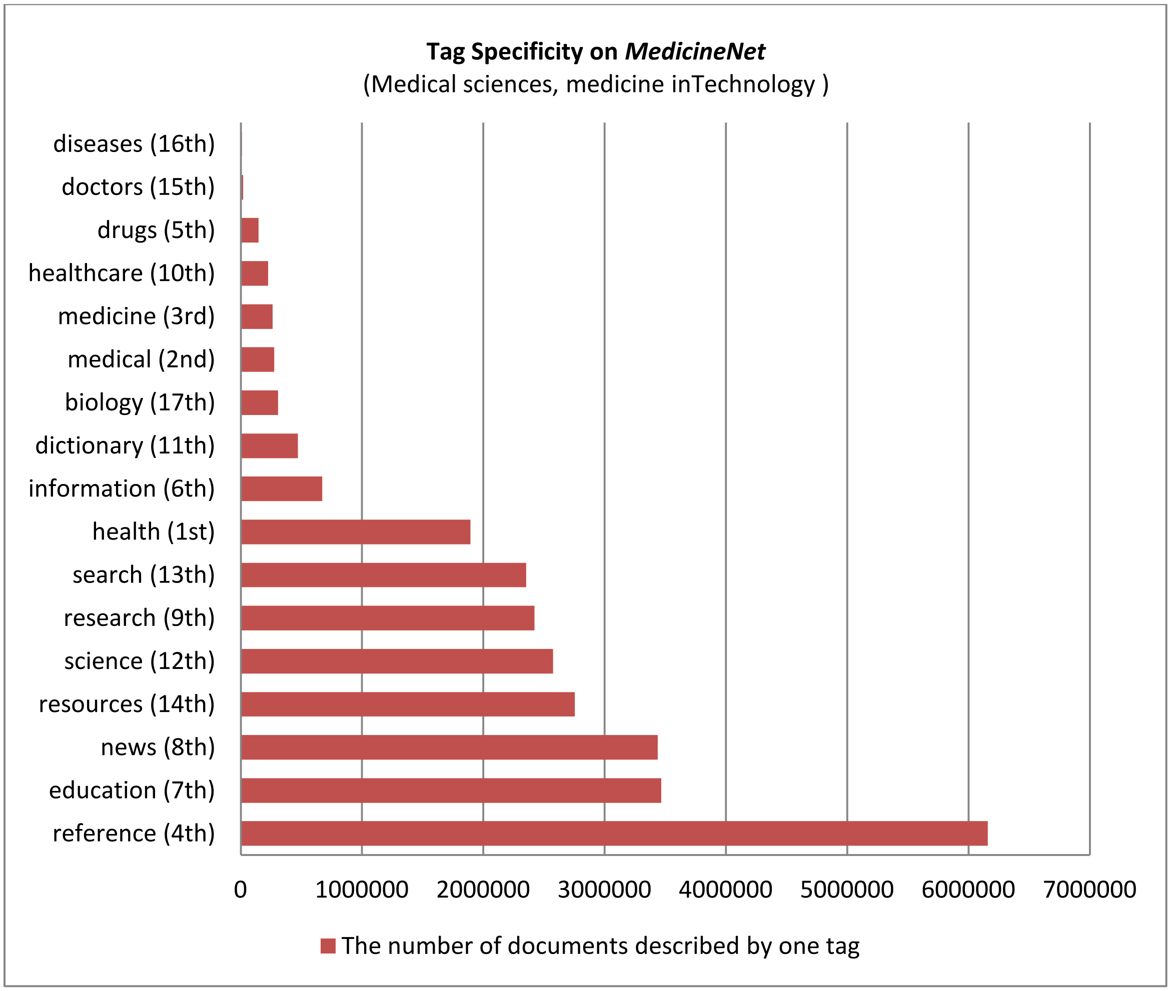
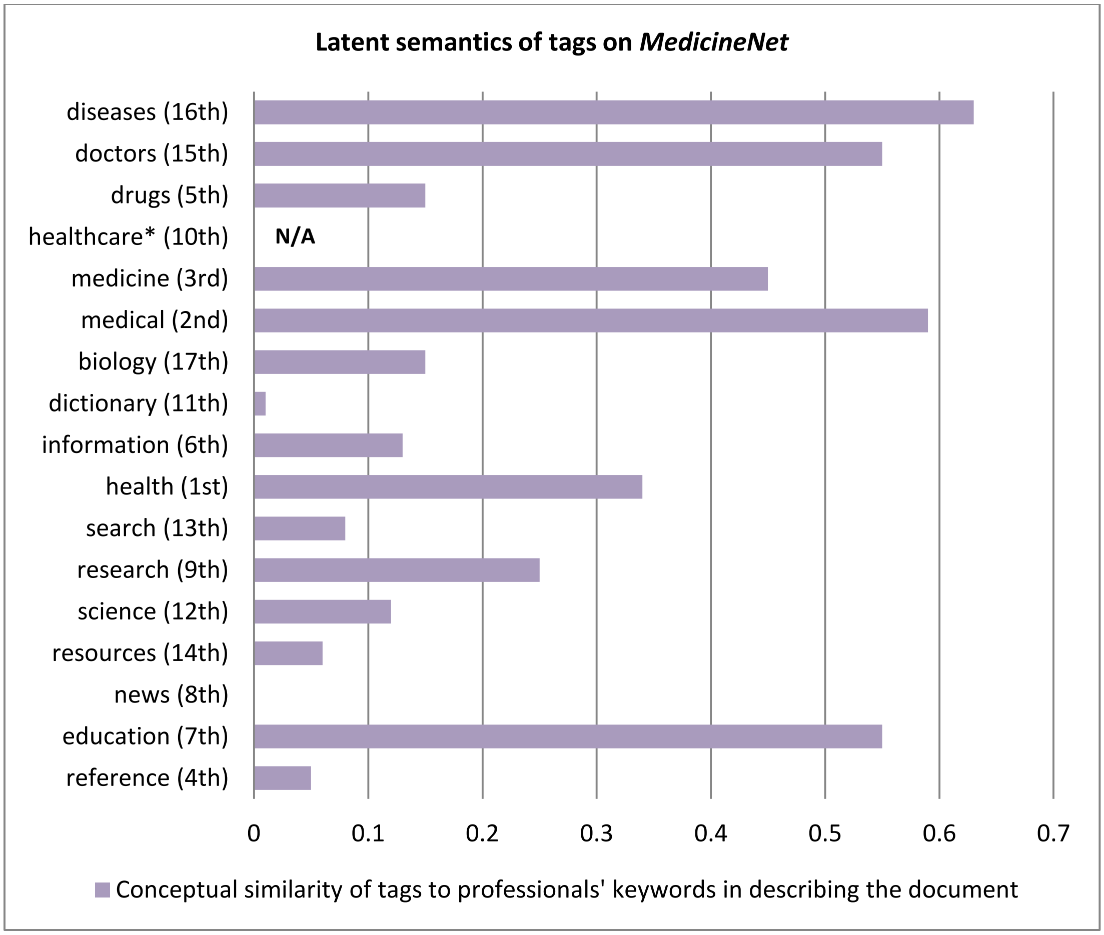


Figure 56. Specificity of tags assigned on the document “*MedicineNet*”



(Note: The term with an asterisk (*) does not exist in the corpus)

Figure 57. Latent semantic relatedness of tags on professionals' keywords (the document "*MedicineNet*")

Table 35. Term-to-term comparison on *MedicineNet*

Document	Title	Intute keywords	Delicious tags
Technology 610 Medical sciences, medicine	MedicineNet, http://www.medicinenet.com/script/main/hp.asp	disease, patient education	health, medical, medicine, reference, drugs, information, education, news, research, healthcare, dictionary, science, search, resources, doctors, diseases, biology

Indexers in BUBL and Intute describe this document as:

"*Articles providing health information including news, and details of diseases and treatments*" (BUBL)

"*aiming to provide medical information to the public. There is extensive information, divided into the features of Diseases and Conditions, Signs and Symptoms, Procedures and Tests, Medications, and a MedTerms Medical Dictionary*" (Intute)

BUBL and Intute's descriptions about this document include "health information like diseases", and "medical information". The highly ranked tags like "medical" (2nd) and "medicine" (3rd) as topical terms showed higher semantic relatedness to professionals' keywords than other tags (Figure 57). The term "healthcare" did not exist in the corpus, but it can be observed as a good index term which describes the document well. The term "diseases", which was described by BUBL and Intute, was the tag with the highest specificity value and semantic relatedness.

- Document 6: *AgNIC: Agriculture Network Information Center*

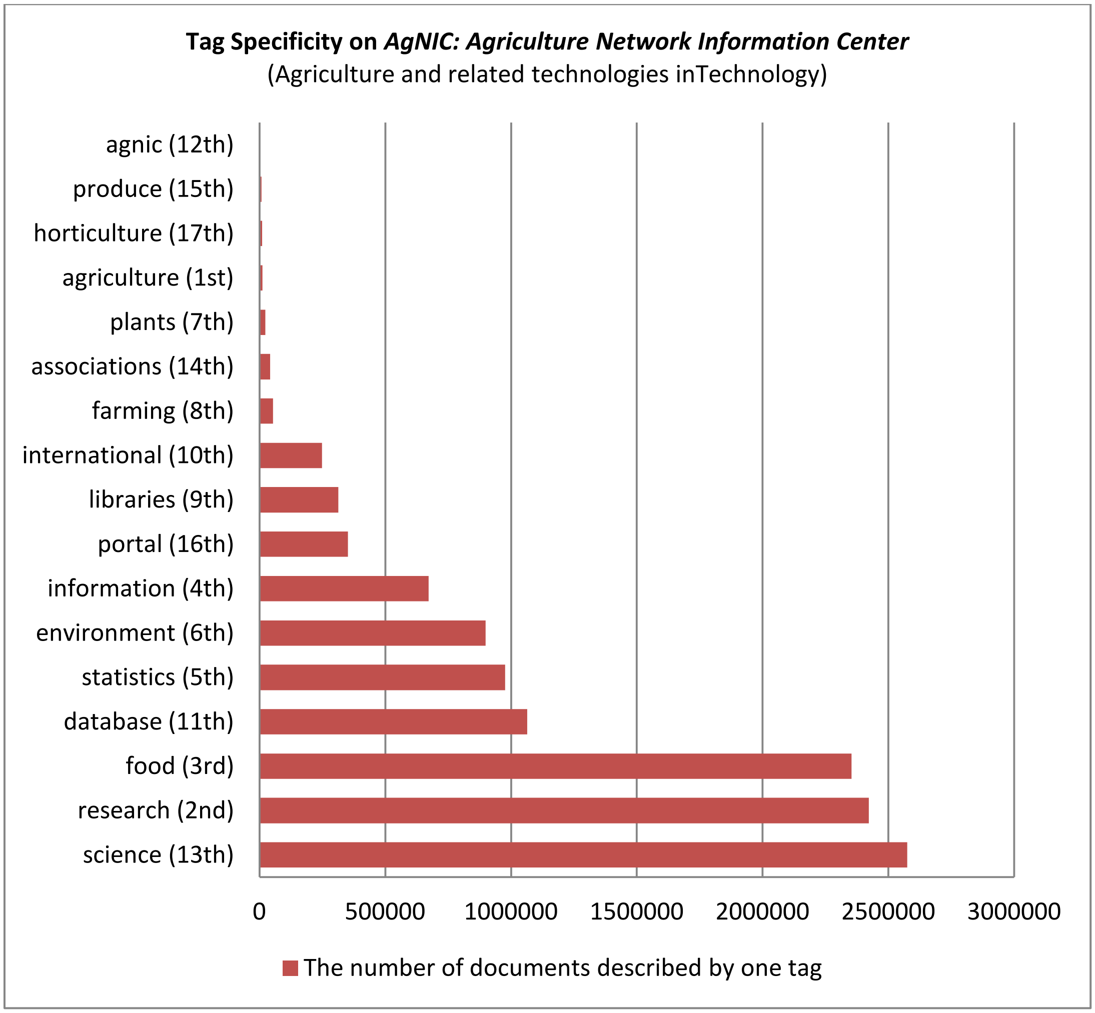
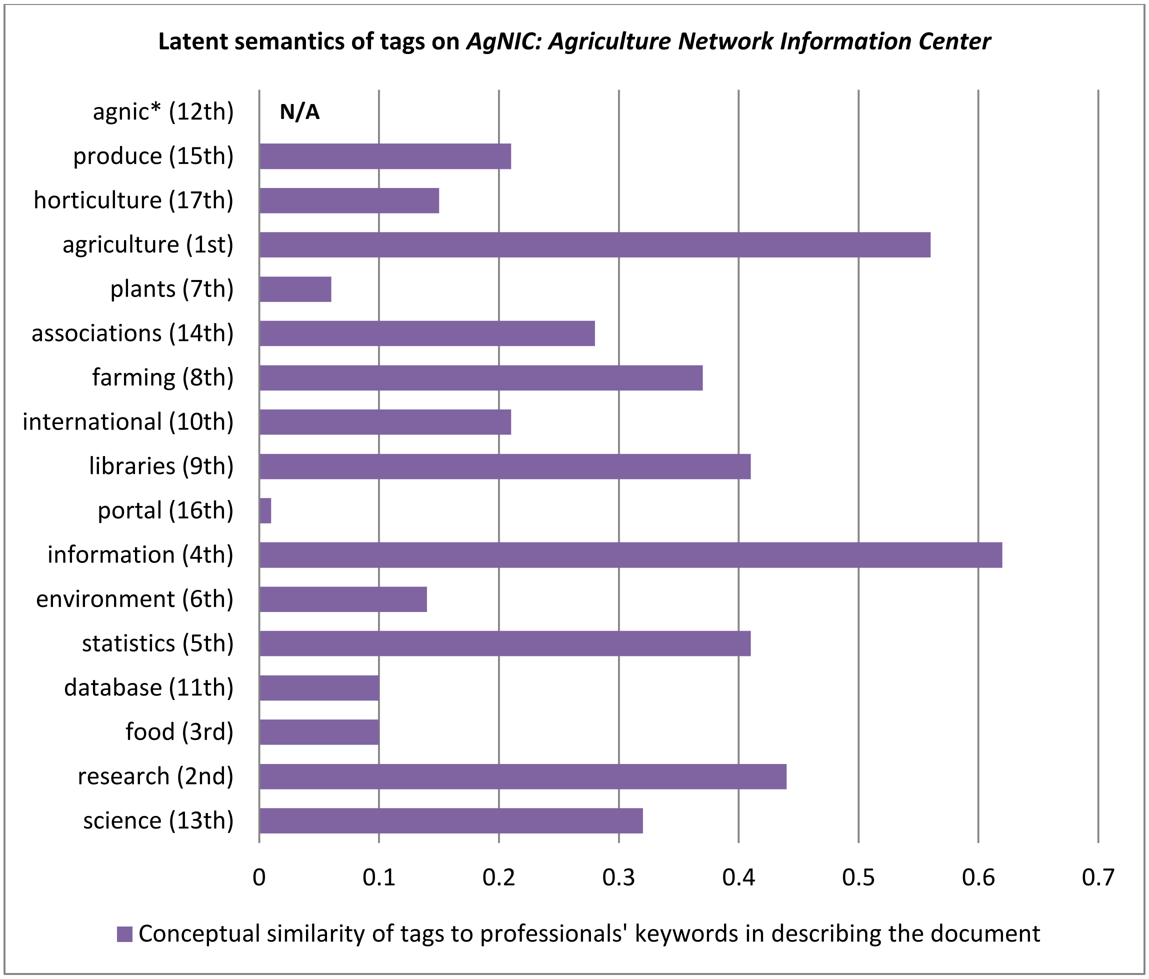


Figure 58. Specificity of tags assigned on the document "*AgNIC: Agriculture Network Information Center*"



(Note: The term with an asterisk (*) does not exist in the corpus)

Figure 59. Latent semantic relatedness of tags on professionals' keywords (the document "AgNIC: Agriculture Network Information Center")

Table 36. Term-to-term comparison on *AgNIC: Agriculture Network Information Center*

Document	Title	Intute keywords	Delicious tags
Technology 630 Agriculture and related technologies	AgNIC: Agriculture Network Information Center, http://www.agnic.org/	agricultural_sciences, agriculture, agricultural_education, information_center	agriculture, research, food, information, statistics, environment, plants, farming, libraries, international, database, agnic, science, associations, produce, portal, horticulture

The document is described as:

"*A distributed network that provides access to agriculture related information and subject area guides, such as economics, animal science, food science, plant science, forestry, and natural resources. An online reference service is also included*" (BUBL)

"*The Agriculture Network Information Center (AgNIC) is a guide to quality agricultural information on the Internet as selected by the National Agricultural Library, Land-Grant Universities, and other institutions*" (Intute)

The tag "agriculture" is a term with high specificity value as well as a term highly ranked (1st), which means that it is the tag that numerous taggers assigned to describe the document (Figure 58). It is also a term with high semantic relatedness (Figure 59).

5.3.3 Correlations of tag specificity and LSA values

To investigate whether there is a relationship between tag specificity and the latent semantics of tags to professionals' keywords, this research conducted a correlation analysis. Latent Semantic Analysis (LSA) of tags was analyzed to see how much tags are semantically related to professionals' keywords which are considered as good index terms for a document. Since the tag specificity was calculated based on the number of documents described by a tag, two variables for correlation analysis were (1) the number of documents described by a tag and (2) LSA values between tags and professionals' keywords which was described in Section 5.3.2 *Latent Semantic Analysis (LSA) on the Conceptual Semantics of Tags*.

There are several measures of correlation analysis. As a most common measure of correlation, Pearson product-moment correlation coefficient measures a linear relationship between two variables by explaining how one variable is linearly related to another in terms of the direction and degree. On the other hand, rank correlation measures relationships between different rankings on the same items, and there are two main measures: Spearman rank correlation coefficient and Kendall tau rank correlation coefficient (tau-a, tau-b, and tau-c)[5]. The Spearman rank correlation coefficient is similar to the Pearson product-moment correlation coefficient except it is calculated from the ranks of the data. Spearman rank correlation coefficient (Crichton, 1999) is a non-parametric measure, that is, it does not assume the nature of the relationship between variables, e.g., linear relationship or frequency distribution of the variables. Kendall tau rank correlation coefficient (Conover, 1980) is also a non-parametric rank-correlation. While the Spearman rank correlation coefficient measures a monotonic function between two variables, Kendall tau rank correlation coefficient represents a probability, that is, it measures the portion of ranks that match between two data sets (Conover, 1980).

The correlations range between plus and minus one. 0 is no correlation, 1 is perfect positive correlation, and -1 is perfect negative. It is considered to be a strong correlation if the correlation coefficient is greater than 0.8 and a weak correlation if the correlation coefficient is less than 0.5.

In this dissertation research, three correlation measures (Pearson correlation coefficient, Spearman rank correlation coefficient, and Kendall tau-b rank correlation coefficient) were used to test for relationships between two variables: the number of documents described by a tag and latent semantics of tags.

[5]. The differences among them are that tau-a does not make any adjustment for ties and is suitable for square tables (tables where the rows and columns are equal), tau-b makes adjustments for ties and is suitable for square tables, and tau-c is more suitable if the table is rectangular.

The null hypotheses and the alternative hypotheses for each measure were:

- Pearson correlation coefficient:

  H_0: r_{Prs} = 0 (there is no correlation between the variables)

  H_1: r_{Prs} < > 0 (variables are correlated)

- Spearman rank correlation coefficient:

  H_0: r_{Spm} = 0 (there is no correlation between the ranked pairs)

  H_1: r_{Spm} < > 0 (ranked pairs are correlated)

- Kendall tau-b correlation coefficient:

  H_0: $\tau_{Ken,b}$ = 0 (there is no correlation between the two variables)

  H_1: $\tau_{Ken,b}$ < > 0 (the two variables are correlated)

The results of correlation analysis on tag specificity and LSA values between tags and professionals' keywords showed that there *is* a correlation between two data sets or variables. The value of the Pearson correlation coefficient was statistically significant (0.002), at a significance level of 1% (Table 37). Thus, it rejected the null hypothesis that there is no correlation between the variables. Pearson correlation coefficient on two variables was -0.317, which means there is a weak correlation between the number of documents described by a tag and the latent semantics of tags.

In non-parametric correlation measures such as Spearman's rank correlation coefficient and Kendall-tau rank correlation coefficient, correlation coefficients were also statistically significant at the 0.01 level which rejected the null hypothesis that there is no correlation between the variables. In the Spearman rank correlation coefficient, it can be interpreted that there is a monotonic weak decreasing relationship, i.e., -.340 (r_{Spm}) between the number of documents described by a tag and the latent semantics of tags (Table 38). In addition, the value of the Kendall-tau correlation coefficient was -.235. ($\tau_{Ken,b}$), which implies decreasing agreement between rankings.

Table 37. Parametric correlations

		The number of documents described by a tag	**LSA**
The number of document described by a tag	Pearson Correlation	1	-.317**
	Sig. (2-tailed)		.002
	N	91	91
LSA	Pearson Correlation	-.317**	1
	Sig. (2-tailed)	.002	
	N*	91	91
**. Correlation is significant at the 0.01 level (2-tailed). *. N represents the number of tags used in LSA computation and excludes tags which did not exist in the LSA corpus			

Table 38. Nonparametric correlations

			The number of documents described by a tag	**LSA**
Spearman's rho	The number of documents described by one tag	Correlation Coefficient	1.000	-.340**
		Sig. (2-tailed)	.	.001
		N	91	91
	LSA	Correlation Coefficient	-.340**	1.000
		Sig. (2-tailed)	.001	.
		N	91	91
Kendall's tau_b	The number of documents described by one tag	Correlation Coefficient	1.000	-.235**
		Sig. (2-tailed)	.	.001
		N	91	91
	LSA	Correlation Coefficient	-.235**	1.000
		Sig. (2-tailed)	.001	.
		N	91	91
**. Correlation is significant at the 0.01 level (2-tailed).				

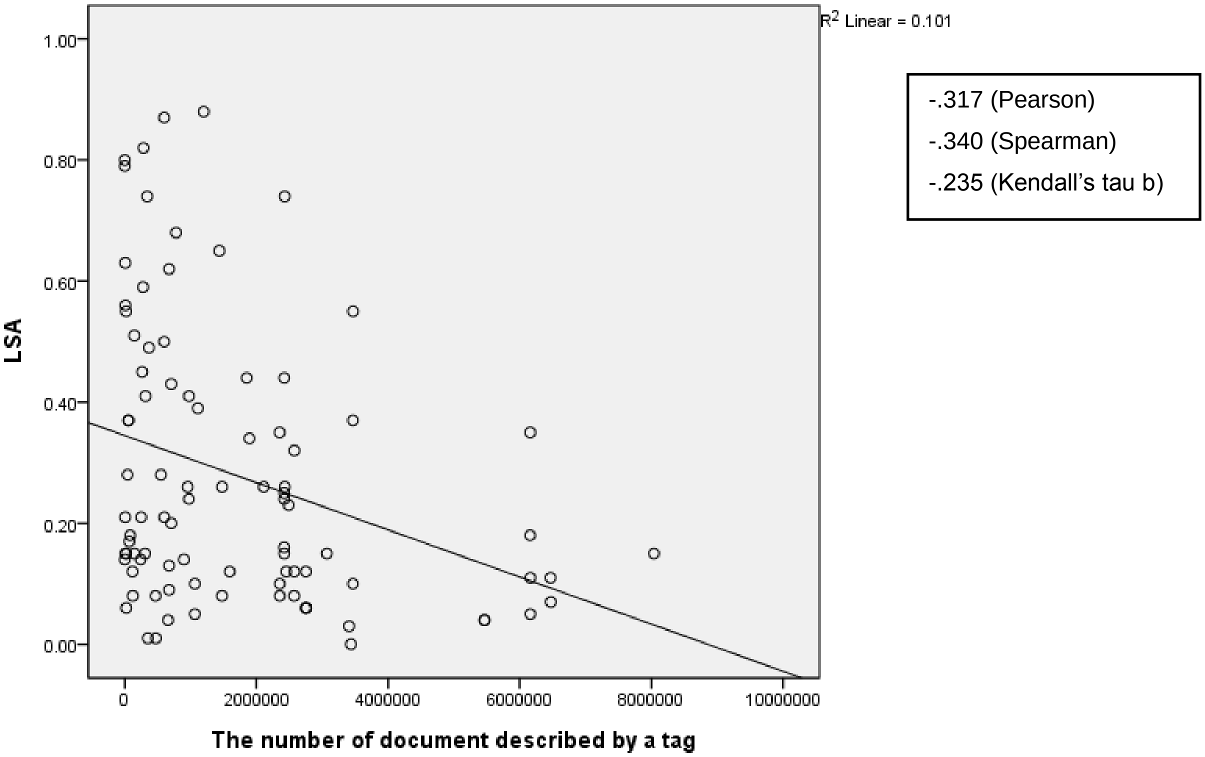


Figure 60. Plot of "the number of documents described by a tag" vs. LSA

Graphical representation of the number of documents described by a tag and latent semantics of tags is presented in Figure 60. In this dissertation research, the concept of tag specificity was redefined as "the number of documents described by one tag" (Spärck Jones, 1972; cited by Hassan-Montero and Herrero-Solana, 2006). That is, when the number of documents described by one tag is smaller, tag specificity is higher. Figure 60 shows a negative linear relationship between the number of documents described by a tag and the latent semantics of tags to professionals' keywords. Thus, it suggests that there is a *positive* linear relationship between tag specificity and the latent semantics of tags to professionals' keywords (Pearson). Furthermore, at least, there *is* an *increasing* relationship between tag specificity and LSA values as measured by the two non-parametric measures, Spearman and Kendall tau-b. Figure 60 also illustrates that low frequency tags do not consistently have high LSA values, while high frequency tags tend to have lower LSA values. This can be explained by examining the variations of LSA values. For example, in Figure 49, the values of conceptual similarity range between 0 and 0.3, while the values of conceptual similarity range between 0 and 0.9 in Figure 51 and Figure 55. Moreover, there is also a wide variation of the number of documents described by a tag on each document. For instance, the maximum number of documents described by a tag is respectively less than 4,000,000 in Figure 50 and Figure 58, while the maximum number of documents described by a tag is greater than 6,000,000 in Figure 52. These variations of tag specificity and LSA values affected the correlations and made the strength of relationship between two variables rather weak. In addition, comparing the pairs of figures (e.g., Figure 48 and 49), it is evident that there is inconsistency in the conceptual similarity of the low frequency tags--some have high values and others do not. That has contributed to the weak correlation.

5.4 Summary and Discussion

The analysis of tagging exhaustivity and tag specificity in relation to tagging effectiveness was conducted to ameliorate difficulties associated with limitations in the analysis of indexing consistency based on only the quantitative measures of vocabulary matching. The findings of this analysis demonstrated the potential value of user-generated tags as index terms. Even though high recall does not necessarily imply high quality of indexing, it is interesting to note that regarding the subject categories with higher indexing similarity (Philosophy, Religion, Arts, and Literature), the tagging exhaustivity also reached relatively higher values (more than three tags per document) compared to other subject categories.

As discussed in 5.3.1 *Tag specificity*, if a term is frequently used for describing documents, the term is not a good differentiator between documents. On the contrary, if there are few documents where a tag is assigned, the tag is a good differentiator. To better understand the properties of tags in Delicious, the top 20 tags associated with six documents were analyzed in detail. Bar graphs of tag frequency demonstrated a wide range, with each document having at least a few relatively low frequency tags.

To further investigate the quality of tags, a Latent Semantic Analysis (LSA) was conducted to determine to what extent tags are conceptually related to professionals' keywords as assigned by indexers in Intute. In this case professionals' keywords are assumed to be good index terms and thus presented as a standard for comparison with tags through LSA. Then to investigate whether there is a relationship between tag specificity and the latent semantics of tags, a correlation analysis was conducted. The results of the correlation analysis showed that tags of higher specificity tended to have a higher semantic relatedness to professionals' keywords. Tags with high specificity value are considered to be good differentiators, and professionals' keywords are regarded as accurate index terms. This leads to the conclusion that the term's power as a differentiator is related to its semantic relatedness to documents, with the caution that correlations between tag specificity and LSA values are limited to the top 20 ranked tags.

Other tags were not included in the LSA analysis because, as recently coined terms or acronyms, they did not exist in the LSA corpus. But they often were among the most specific tags associated with the document and thus also good differentiators.

CHAPTER 6: TAG ATTRIBUTES AND TAGGING BEHAVIOR

6.1 Overview

This chapter presents the results of the analyses of tag attributes based on the FRBR model. The results illustrate important tag attributes and tagging behaviors by subject. Research question 7 was answered:

> Research question 7) What are features and patterns of social tagging in describing a web document at Delicious? Do tags have other bibliographic attributes beyond describing subjects or topics of a document?

The content analysis on tag attributes was conducted on a total of 113 web documents regarding 11 attribute categories defined by FRBR (5 categories from Work entity and 6 categories from Expression entity). An intercoder reliability test between two coders on 29 web documents was conducted.

6.2 Results of the Intercoder Reliability Test

In order to improve research reliability and objectivity in the analysis of tag attributes, another coder was recruited and the intercoder reliability between two coders was calculated. The recruited coder was a PhD candidate in Library and Information Science. Two coders independently coded tags based on the coding instruction. A sample of coded web document is provided in Appendix G. Table 39 illustrates the crosstabulation of coded data by two coders. The intercoder reliability test was calculated by using the Holsti method, Scott's pi, Cohen's kappa and Krippendorf's alpha. The results of intercoder reliability test by subject areas are included in Appendix H.

Table 39. Crosstabulation of coded data

		CODER B												Total
		N/A	**WT**	**WF**	**WD**	**WI**	**WC**	**EF**	**ED**	**EL**	**ES**	**EU**	**ET**	
CODER A	**N/A**	255	2	3	0	2	4	5	0	0	0	0	0	**271**
	WT	6	41	0	0	0	2	0	0	0	0	0	0	**49**
	WF	5	1	54	0	0	0	7	0	0	0	0	0	**67**
	WD	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	WI	0	0	0	0	1	0	0	0	0	0	0	0	**1**
	WC	2	0	0	0	0	23	0	0	1	0	0	0	**26**
	EF	8	0	2	0	0	0	11	0	0	0	0	0	**21**
	ED	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	EL	1	1	0	0	0	0	0	0	5	0	0	0	**7**
	ES	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	EU	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	ET	0	0	0	0	0	0	0	0	0	0	0	0	**0**
Total		**277**	**45**	**59**	**0**	**3**	**29**	**23**	**0**	**6**	**0**	**0**	**0**	**442**

(Notes: N/A cells mean the cases where a tag was not determined as any categories such as WD, ED, ES, EU, and ET)

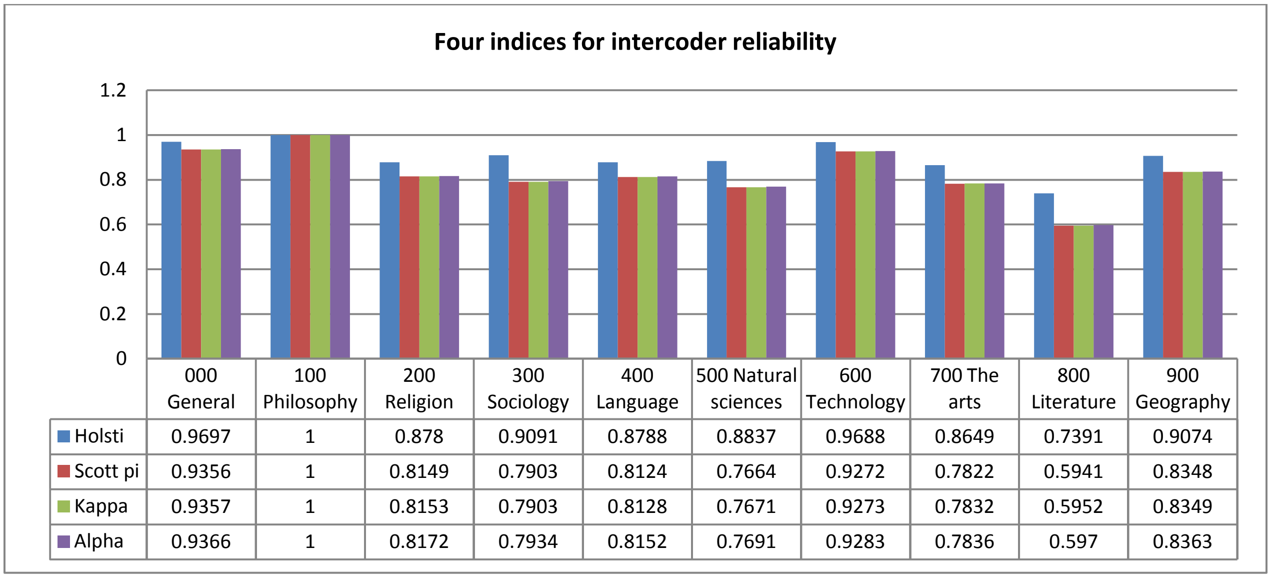
In terms of criteria for acceptability, index scales are analogous but it has been cautioned that different indices measure different things (Lombard et. al., 2004; Neuendorf, 2002). Therefore, a satisfactory level depends on the index used (Taylor and Watkinson, 2007).

Holsti (1969) suggests the agreement level of 85 % or more for the acceptable level. Banerjee et al. (1999) suggest that Cohen's kappa levels should exceed 0.75 for excellent agreement beyond chance, between 0.40-0.70 is fair to good agreement beyond chance, and <0.40 is poor agreement. Landis and Koch (1977) have provided a more detailed list of interpretation of kappa: 0.81 – 1.00 is almost perfect agreement**,** 0.61 – 0.80 is substantial agreement, 0.41 – 0.60 is moderate agreement, 0.21 – 0.40 is fair agreement, 0.0 – 0.20 is slight agreement and < 0 is poor agreement. For the case of Krippendorff's alpha, it has been suggested to exceed 0.70 for excellent agreement (Krippendorff, 2004; Taylor and Watkinson 2007). Therefore, in this dissertation research, in four indices, the results of the intercoder reliability test showed an excellent agreement as Table 40 shows:

Table 40. Results of intercoder reliability test using four indices

Measure of reliability	Value	Units
Holsti	.8824	442
Scott’s pi	.7963	
Cohen’s kappa	.7963	
Krippendorff's Alpha	.7965	

In order to investigate the degree of reliability among subject areas, the reliability test on each subject area was performed. The results of intercoder reliability test using four indices demonstrated that the Literature subject showed the lowest level of agreement among 10 different subject areas (Figure 61). Table 41 illustrates the crosstabulation of coded data by two coders on the Literature subject. More details on the results of intercoder reliability test by 10 subject areas are provided in Appendix H.



	000 General	100 Philosophy	200 Religion	300 Sociology	400 Language	500 Natural sciences	600 Technology	700 The arts	800 Literature	900 Geography
🟦 Holsti	0.9697	1	0.878	0.9091	0.8788	0.8837	0.9688	0.8649	0.7391	0.9074
🟥 Scott pi	0.9356	1	0.8149	0.7903	0.8124	0.7664	0.9272	0.7822	0.5941	0.8348
🟩 Kappa	0.9357	1	0.8153	0.7903	0.8128	0.7671	0.9273	0.7832	0.5952	0.8349
🟪 Alpha	0.9366	1	0.8172	0.7934	0.8152	0.7691	0.9283	0.7836	0.597	0.8363

Figure 61. The results by four indices for intercoder reliability

Table 41. Crosstabulation of coded data (Literature subject)

		CODER B												**Total**
		N/A	**WT**	**WF**	**WD**	**WI**	**WC**	**EF**	**ED**	**EL**	**ES**	**EU**	**ET**	
CODER A	**N/A**	31	1	1	0	0	1	3	0	0	0	0	0	**37**
	WT	2	7	0	0	0	0	0	0	0	0	0	0	**9**
	WF	2	0	8	0	0	0	3	0	0	0	0	0	**13**
	WD	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	WI	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	WC	0	0	0	0	0	3	0	0	0	0	0	0	**3**
	EF	3	0	0	0	0	0	2	0	0	0	0	0	**5**
	ED	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	EL	1	1	0	0	0	0	0	0	0	0	0	0	**2**
	ES	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	EU	0	0	0	0	0	0	0	0	0	0	0	0	**0**
	ET	0	0	0	0	0	0	0	0	0	0	0	0	**0**
Total		**39**	**9**	**9**	**0**	**0**	**4**	**8**	**0**	**0**	**0**	**0**	**0**	**69**

It was found that there was especially low agreement between two coders on two attribute categories, i.e., WF (Form of Work entity) and EF (Form of Expression entity). The examples of those tags were Books, Database, Magazine, Journal, and Encyclopedia. This disagreement on those attributes was caused by the fact that the documents, tagged with a term "Book", include the list of books or provide a feature of searching for books rather than books themselves (see Table 42).

Table 42. Web documents tagged with the term "book"

Document	Description
808.8 Google Book Search: http://books.google.com/	"helps users search within and discover books using digital page-scans presented in a simple ebook format" (provided by BUBL)
809 Literary history: http://literaryhistory.com/	"provides selected annotated links to critical articles on British and American literatures" (provided by Intute)

However, current definitions provided by FRBR do not explicitly distinguish these two attributes (i.e., WF and EF) about web documents. To make FRBR more applicable in practical aspect, FRBR should be able to describe digital heterogeneous media resources which are available in various formats and multi-dimensional structures.

As discussed, the results of the intercoder reliability test were very satisfactory with excellent agreement for all four indices, but it is very important to note that *reliability* and *validity* are different. Reliability is concerned with the consistency of the measurement while validity is related to the strengths of the results. Krippendorff (2008) asserts that validity is about truth and reliability relates to trust. He also argues that "reliability cannot guarantee validity". Thus, the results of the intercoder reliability test do not determine the validity of the conclusions on tag analysis, but instead, they contribute to enhancing confidence in reliability. In the following sections, the results on the analysis of tag attributes are discussed for the whole collection of documents.

6.3 Categories of Tag Attributes

During the process of content analysis on tag attributes, if a tag was determined to be a term related to subjects or topics describing documents, the tag was categorized into "Subject". Also, if a tag was identified as a term which cannot be categorized into any of the categories defined by FRBR, the tag was categorized into the "Others". Finally it was determined that the tags included in the "Others" would be assigned to sub-categories such as Feature, Utilization, and Institution etc, and the discussion of those tags will be provided later. The findings on the analysis of tag attributes are depicted as follows:

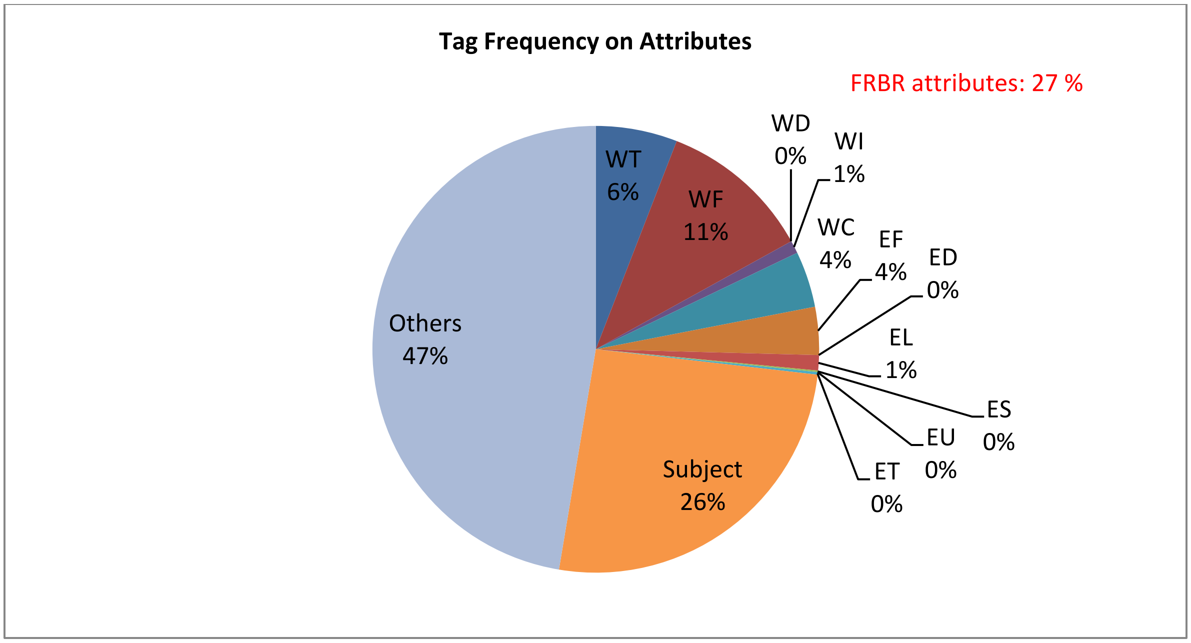


Figure 62. Tag frequency and attribute categories

Figure 62 illustrates that among tags assigned to the sampled documents, in the pie chart, 26 % of tags were subject-related terms, 27 % of tags were matched into the attributes of FRBR, and 47 % of tags were categorized into other attributes. This illustrates that many tags (about 74 %) include additional properties beyond subject or topic terms.

6.4 Tagging Behaviors

In order to investigate whether the attributes of tags could be described by the FRBR attributes, the matching process was conducted between tags and FRBR attributes. Tags were identified based on the attribute categories defined by FRBR as shown in Table 43. Table 43 excludes the WT (Title of work entity) category where tags consist of terms used in the title of the document.

Table 43. Identified tags and related FRBR attributes

Entities	Attributes	Identified tags
Work	Form of work (WF)	Reference, journal, research, magazine, news, paper, article, dictionary, archive, database, directory, book, essay, scripture, gov-doc, encyclopedia, glossary, tutorial
	Date of work (WD)	N/A
	intended audience (WI)	baby, doctor, engineer, artist, dealer, architect, author, writer, children, illustrator, poet, teacher
	context for the work (WC)	world, war, uk, primary source, 18c, India, usa, middleeast, federal, Boccaccio, Medieval, ancient
Expression	form (EF)	Music, ebook, texts, iconography, images, statistics, word, video, vocabulary, etext, bibtex, pictures, photos, multimedia, graphic, audio, sound, illustration, posters
	date (ED)	N/A
	language of expression (EL)	English, Hebrew, Greek,
	summarization of content (ES)	list
	use restrictions on the expression(EU)	N/A
	technique (graphic or projected image) (ET)	Graphic organizer, flash

Regarding the tags related to subject terms, in Language, Literature, and Geography subject, the number of subject-related tags was relatively low (Figure 63 and Figure 64).

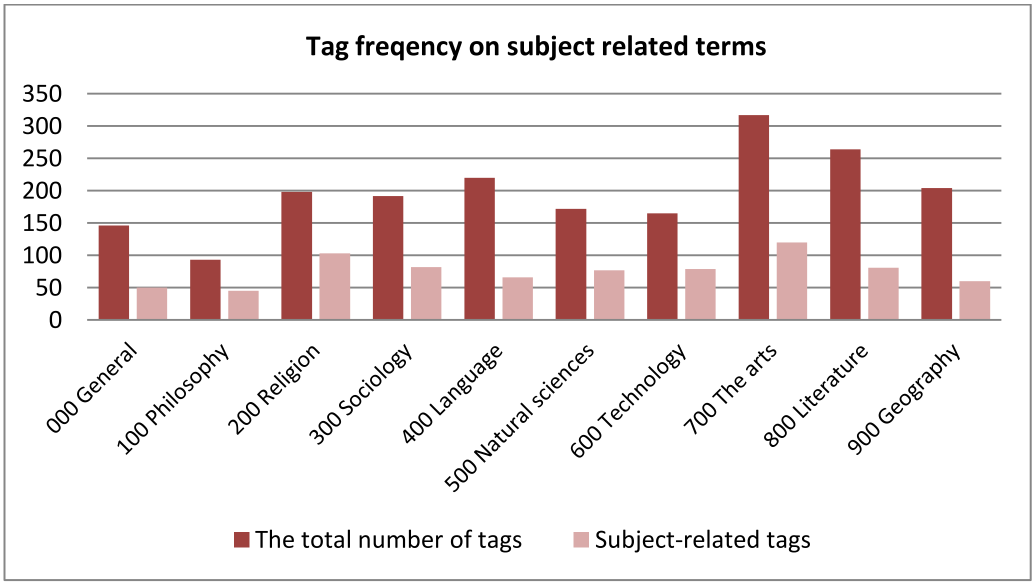


Figure 63. Tag frequency on subject related terms

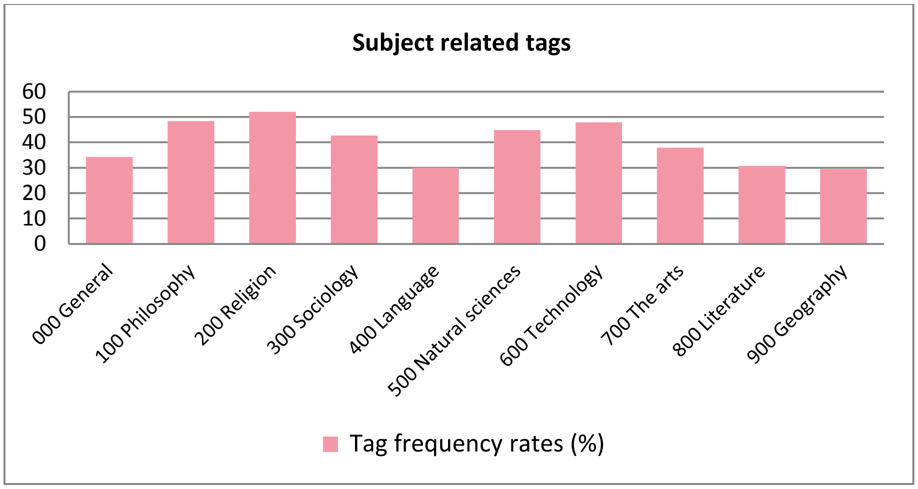


Figure 64. Tag frequency rates on subject related terms

Figures 65-67 below illustrate that in terms of web documents in those three subjects, taggers tend to focus more on other properties of documents rather than the subjects or topics of documents, that is, the Form of Work entity (WF) and Form of Expression entity (EF). Since the figures mainly show the comparison of subject-related tags and FRBR categorized tags, the "Others" category are not represented in those figures.

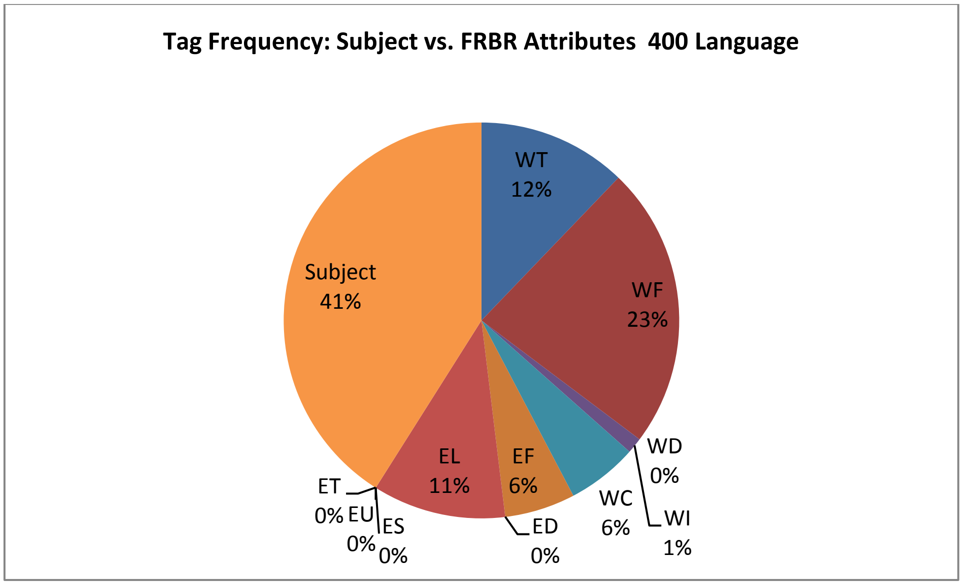


Figure 65. Tag frequency rates on Language subject

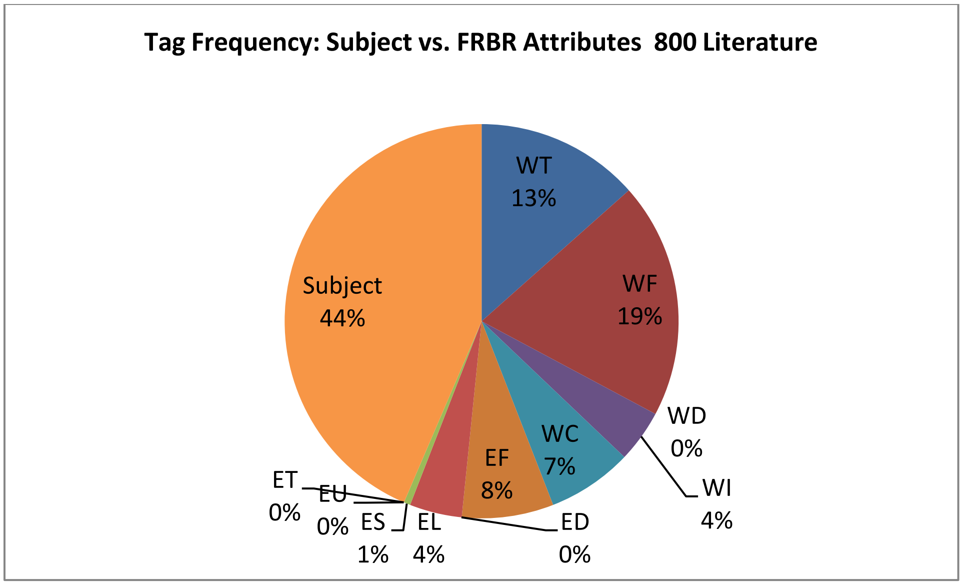


Figure 66. Tag frequency rates on Literature subject

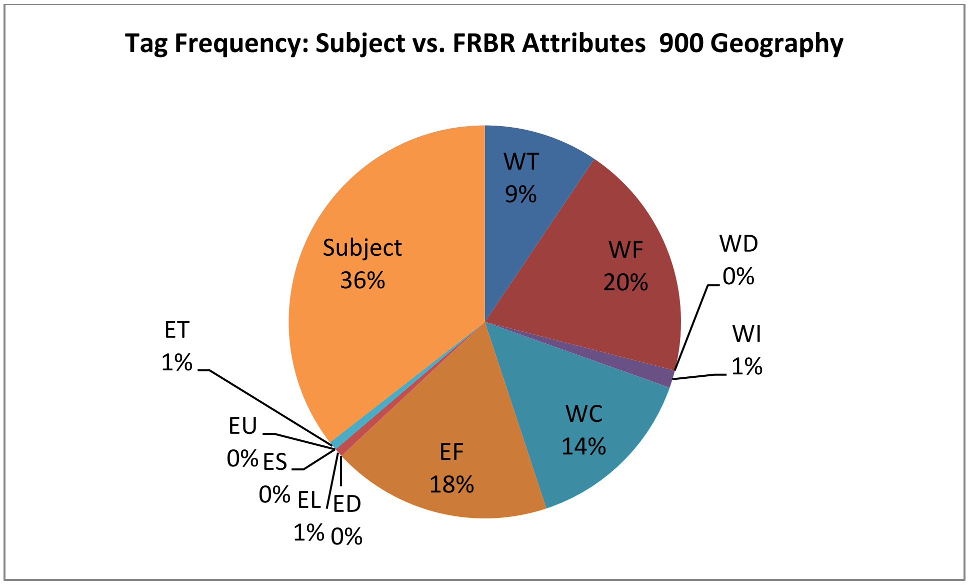


Figure 67. Tag frequency rates on Geography subject

A more in-depth analysis was conducted on the tendency of tagging in terms of 11 FRBR attribute categories. Figures 68 and 69 demonstrate that taggers tend to mainly assign tags on attributes related to WT (Title attribute of FRBR Work entity) and WF.

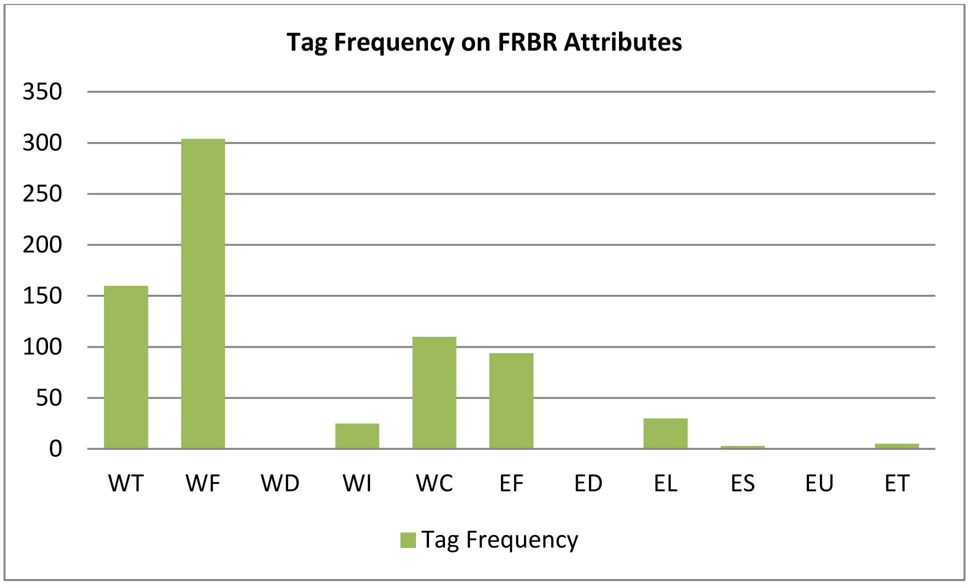


Figure 68. Tag frequency on FRBR attributes (bar graph)

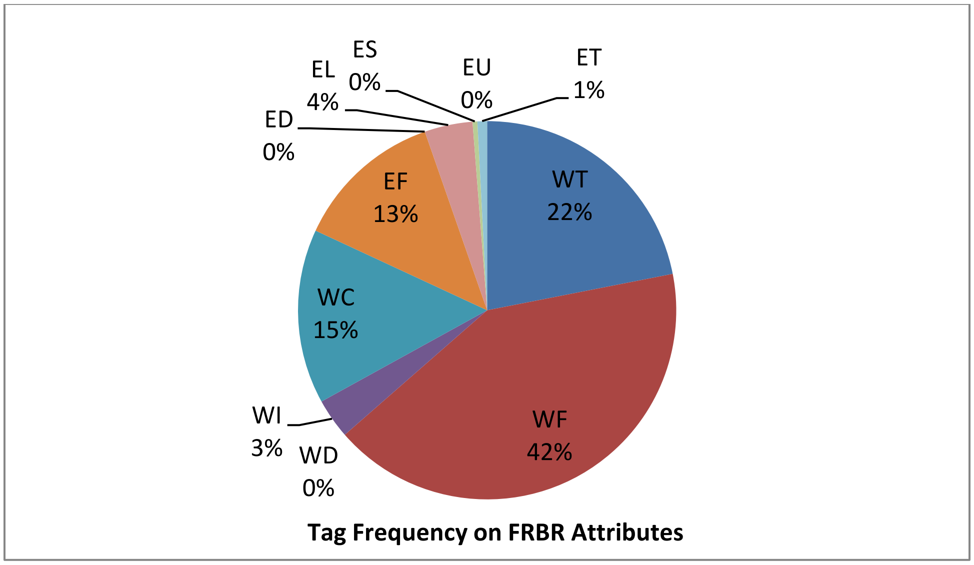


Figure 69. Tag frequency on FRBR attributes (pie chart)

In order to investigate the features and patterns of social tagging in assigning attributes matching those defined in FRBR, a thorough examination was conducted on tags categorized by FRBR attributes. Figure 70 and 71 show tag frequency on the categories defined by FRBR in terms of 10 different subject areas.

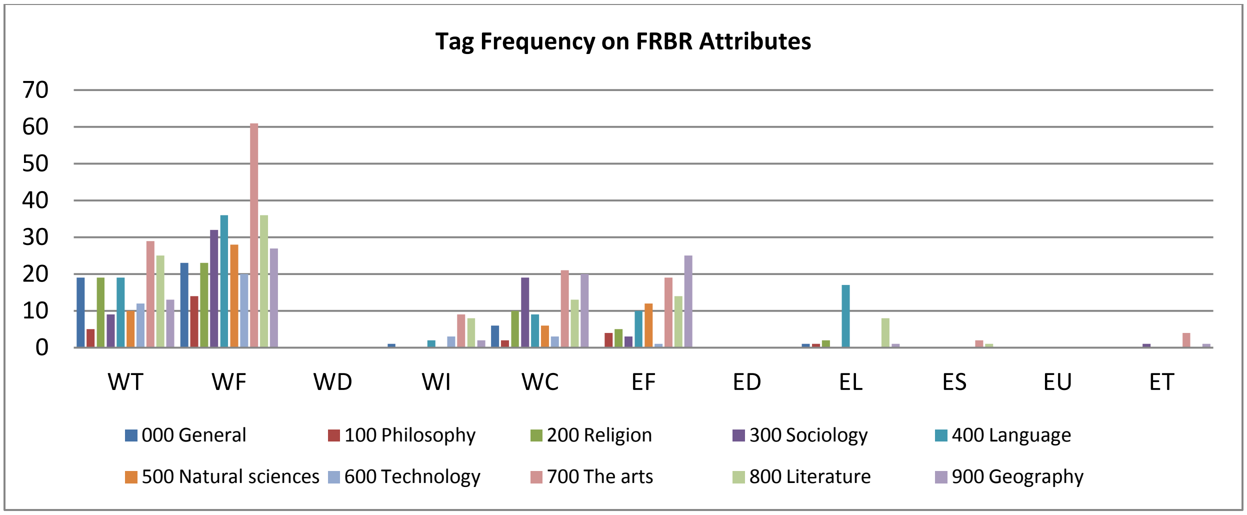
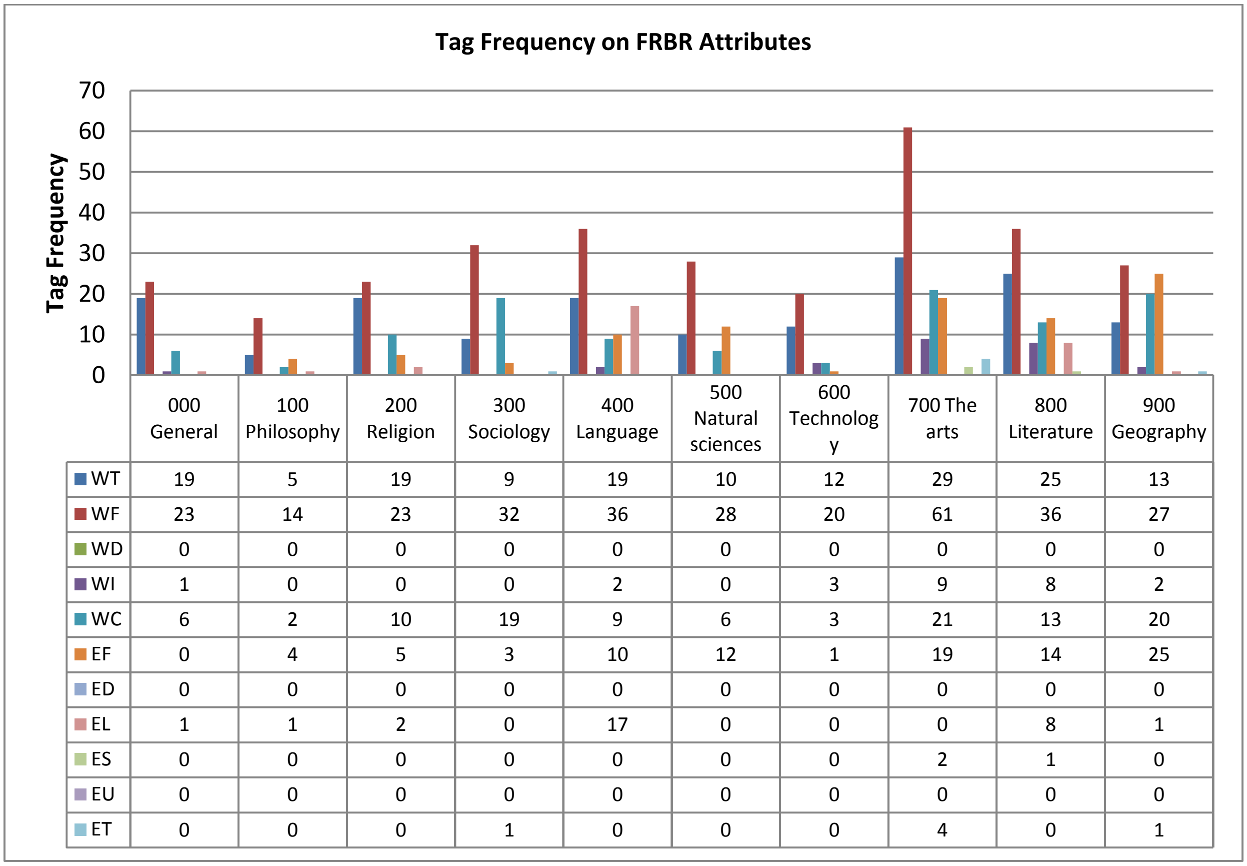


Figure 70. Tag frequency on FRBR attributes over all subjects (bar graph)



	000 General	100 Philosophy	200 Religion	300 Sociology	400 Language	500 Natural sciences	600 Technolog y	700 The arts	800 Literature	900 Geography
WT	19	5	19	9	19	10	12	29	25	13
WF	23	14	23	32	36	28	20	61	36	27
WD	0	0	0	0	0	0	0	0	0	0
WI	1	0	0	0	2	0	3	9	8	2
WC	6	2	10	19	9	6	3	21	13	20
EF	0	4	5	3	10	12	1	19	14	25
ED	0	0	0	0	0	0	0	0	0	0
EL	1	1	2	0	17	0	0	0	8	1
ES	0	0	0	0	0	0	0	2	1	0
EU	0	0	0	0	0	0	0	0	0	0
ET	0	0	0	1	0	0	0	4	0	1

Figure 71. Tag frequency on FRBR attributes over all subjects (bar graph II)

6.4.1 FRBR Intended Audience of Work entity (WI)

As shown in Figure 70 and 71, the tag frequency on FRBR attributes formed a different tendency depending on subject categories. For example, in three subject areas, Technology, Arts and Literature subjects, the tag frequency on FRBR WI (intended audience) attribute was relatively high (see Figure 72), which means that taggers tend to consider audience in these subject areas.

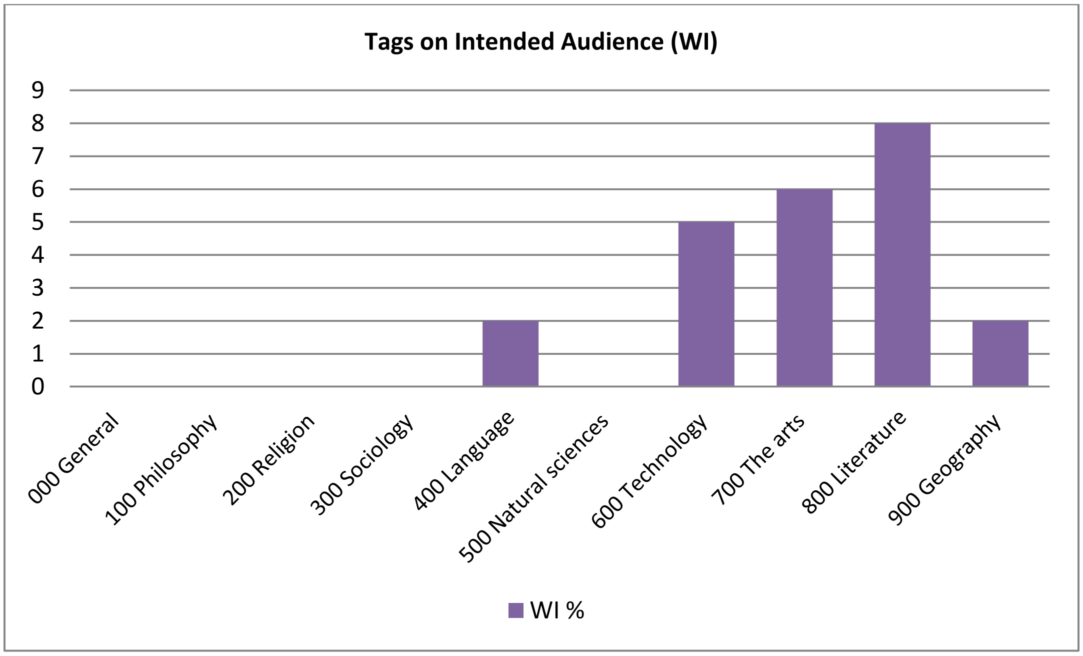


Figure 72. Tags on intended audience (WI)

In the Technology subject, the tags applied to the WI category were doctor, engineer etc. On the other hand, in the Art subject, the tags were artists, architects, and dealers etc. In the Literature subject, the tags were author, poet, children, and writer etc. It can be inferred that high frequency on the WI category in those subject areas reflects the characteristics of different user needs for metadata. For example, in Literature, many documents are intended for adults, so if a document is related to resources for children, taggers tend to specifically indicate it by assigning a tag, "children" as the intended audience.

6.4.2 FRBR Form of Expression entity (EF)

In terms of Natural sciences and Geography, the findings on tag frequency of the EF category showed relatively high proportions (respectively, 21% and 28%) in comparison with those of other subject categories (Figure 73).

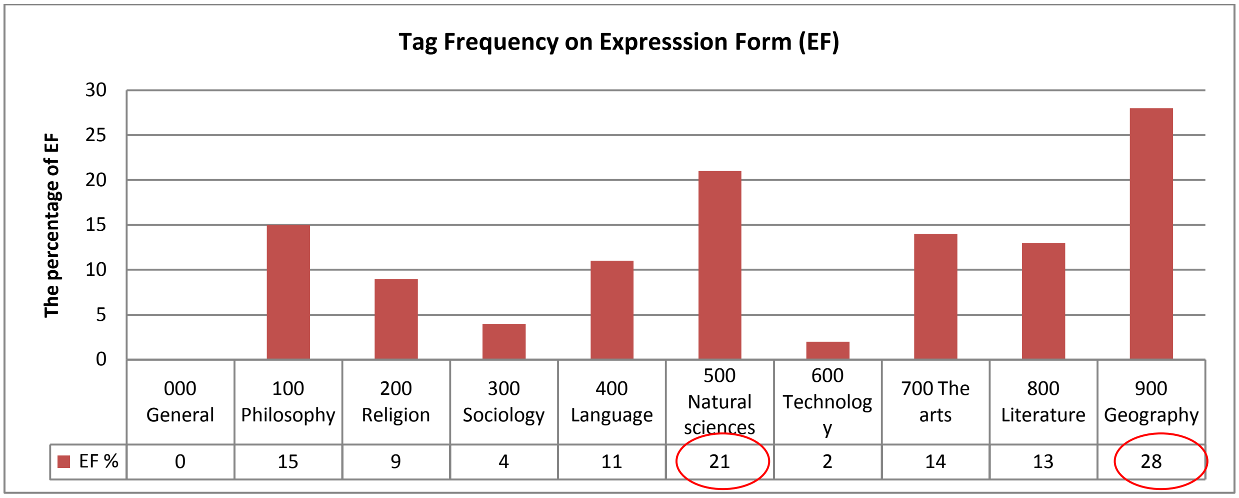


Figure 73. Tags on forms of expression (EF)

In both subject areas, the tags assigned to the EF category were image, video, picture, and photos etc. It implies that web documents in Natural sciences and Geography are mainly characterized by taggers with focus on specific forms.

6.4.3 Other tag attributes

Besides the categories mentioned above, tags having other types of attributes was 47 % (Figure 74).

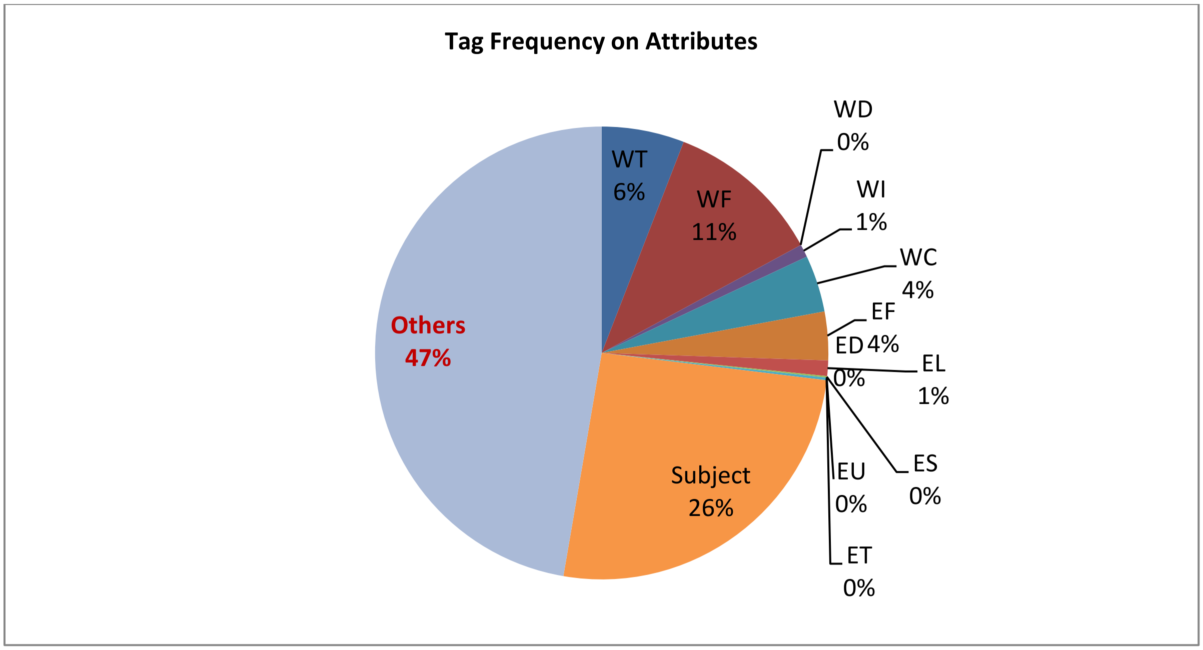


Figure 74. Other attributes of tag

Concerning the other attributes of tags which were not categorized into any attribute categories (FRBR attributes and subject categories), three sub categories were developed to sort out those tags, i.e., Feature, Utilization, and Institution. Also, if a tag cannot be assigned to any of the sub categories mentioned above, the tag was labeled as "Not Applicable" (Table 44).

Table 44. Tag categories for other attributes

Category	**Description**	**Tag**
Feature	Feature is a technical feature about web documents. It reflects the characteristics of web documents.	academic, library, conference, community, search, online, bookmarkbar, open_acess, web2.0, library2.0, homepage, networking, links, blog, tools, access, browse, portal, community, forum, public-domain, wiki
Utilization	Utilization is about the implied purpose of usage.	resources, education, information, learning, e-learning, writing, reading, study, teaching, job, career, tutorial
Institution	Institution	Association, organization, foundation
Not Applicable	cannot be determined as any categories above	imported, flickr

The tags in the Utilization sub category show rather subjective or personal properties. Tags such as resources, learning, teaching, job etc. indicate the user's intention for their subjective or personal purpose.

6.5 Summary and Discussion

In order to characterize the features and patterns of tags, the content analysis of tag attributes was performed based on attributes defined by the FRBR model. To enhance confidence in reliability of the coding, a coder other than the researcher was recruited and the coded results were compared to measure the intercoder reliability. To provide more credible and objective results on tag coding, this research employed four different indices for intercoder reliability: Holsti's method, Scott's pi (p), Cohen's kappa (κ) and Krippendorff's alpha (α). For all four indices, the results of the intercoder reliability test showed excellent agreement.

The findings identified the bibliographic attributes of tags beyond describing subjects or topics of a document. The findings also showed that tags have essential attributes matching those defined in FRBR. In terms of FRBR attributes, the results showed that taggers tend to mainly assign tags on attributes related to WT (Title attribute of FRBR Work entity) and WF (Form attribute of FRBR Work entity). Furthermore, regarding 10 different subject categories, the tag frequency on FRBR attributes showed different tendencies. For three subject areas, Technology, Arts and Literature subjects, tag frequency on the FRBR WI (intended audience) attribute was relatively high, which means that taggers tend to consider audience in these subject areas. In terms of Natural sciences and Geography, the tag frequency of EF (Form attribute of Expression entity) category showed relatively high proportion in comparison with those of other subject categories. This indicated that web documents in both those subject areas were characterized by taggers with a focus on specific forms. The other attributes of tags were sorted into three sub categories, Feature, Utilization, and Institution.

Furthermore, in terms of specific subject areas, taggers exhibited different tagging behaviors representing distinctive features and tendencies. These results have led to the conclusion that depending on subjects, there should be an increased awareness of diverse user needs in the process of metadata generation.

It should be noted that since the scope of data analysis focuses on tags describing web documents, in this research, consideration of the FRBR Manifestation entity and Item entity has been excluded. Given the characteristics of web documents in terms of "web publishing", a web document can be viewed as the "digital embodiment" of a print book or a print journal. In that case, FRBR definitions of manifestation also needed to be extended to identify different manifestations with the same content.

CHAPTER 7: CONCLUSIONS AND FUTURE RESEARCH

This dissertation research examined user-generated social tags in order to see whether they could be used to enhance access to web resources and provide additional access points beyond professionally-generated ones, and whether we could verify the usefulness of social tagging to obtain benefit from it. The main objective of this study was focused on examining the inter-indexer consistency of social tagging for systematically verifying its efficacy and quality. This final chapter provides the conclusions on research questions, the research contributions of this dissertation, and also discusses limitations and directions for future research.

7.1 Conclusions on Research Questions

To review, the following research questions have been central in this dissertation.

- Would social tagging be useful for subject indexing in organizing and providing access to the web? Could we *verify* the usefulness of social tagging to obtain benefit from it?
- How are web resources tagged or indexed at a social tagging site? What kinds of *benefits* could we obtain from tags?

The following specific research questions were answered when exploring the main focuses mentioned above.

1) How consistent is social tagging at Delicious regarding subject indexing of web resources? Is there a relationship between its indexing consistency and subject areas indexed?
2) How consistent is professional indexing between BUBL and Intute?

3) Are there various or alternative interpretations of the same web document between two professionally indexed subject gateways, BUBL and Intute?
4) How consistent is tagging/indexing between Delicious taggers and Intute professionals?
5) Would Delicious users' tags provide additional subject access points beyond index terms or keywords that Intute professionals provide?
6) What levels of tagging exhaustivity and tag specificity in Delicious characterize the indexing of web documents?
7) What are features and patterns of social tagging in describing a web document at Delicious? Do tags have other bibliographic attributes beyond describing subjects or topics of a document?

7.1.1 Conclusions on research question 1

RQ 1) How consistent is social tagging at Delicious regarding subject indexing of web resources? Is there a relationship between its indexing consistency and subject areas indexed?

Using the Information Retrieval (IR) Vector Space Model (VSM)-based indexing consistency measure on indexing consistency of taggers in Delicious, this research answered the first question. Chapter 4 reported that for three measures (Euclidean distance, dot product, and cosine), there was more consistent similarity among Delicious taggers over all subjects than that between two groups of professional indexers, BUBL and Intute. In the results of ANOVA, cosine p-value and dot product p-value suggest that there *is* a significant difference over all subjects in taggers at Delicious, but these p-values are more than p-values for professionals between the two groups BUBL and Intute. Thus indexing similarity among taggers is more consistent than indexing similarity between BUBL and Intute. In addition, in the results of the

Kruskal-Wallis test, cosine p-value and dot product p-value reject the null hypothesis that 10 different subject areas have the same distribution when the significance level is .05. Yet, these p-values are also more than p-values for professionals between the two groups BUBL and Intute, which means there are more similar patterns of indexing among taggers at Delicious than that between two professional groups. Furthermore, to investigate the distribution of tag data, that is, to see whether tagging consistency in Delicious is normally distributed or not, the test of normality was conducted. It was found that for the dot product measure, tagging consistency was not normally distributed while tagging consistency in the cosine similarity and Distance metrics measures derived from a normal distribution.

7.1.2 Conclusions on research question 2 and 3

RQ 2) How consistent is professional indexing between BUBL and Intute?

In chapter 4, the second question was answered by analyzing indexing consistency between two professional groups, BUBL and Intute. It was reported that there was inconsistency of indexing similarity over all subjects using three measures (Euclidean distance, dot product, and cosine). The findings of indexing consistency were also statistically tested by using inferential statistics. The analysis of ANOVA confirmed that between BUBL and Intute's professional indexers, there is inconsistency in the average indexing consistency among 10 different subject areas when α is 0.05. The p-values for all three measures (cosine, dot product, and Euclidean distance) were much less than the p-values in the results of ANOVA on tagging at Delicious, which means that indexing similarity between BUBL and Intute is less consistent than indexing similarity among taggers at Delicious. Also, the results of the Kruskal-Wallis test rejected the null hypothesis, which means that the distribution of similarity over all subjects is not the same between BUBL and Intute.

RQ 3) Are there various or alternative interpretations of the same web document between two professionally indexed subject gateways, BUBL and Intute?

Based on the results of indexing consistency between BUBL and Intute, the subject areas especially showing low similarity were investigated: Literature, Geography, Natural sciences, and Technology. It was reported that BUBL and Intute have different points of view on the same documents in those subject areas:

- *Literature*

For many documents in Literature, Intute categorized them into Languages and Area Studies while BUBL handles the Language category separately.

- *Geography*

In Intute, several documents in Geography were located in Modern Languages and Area Studies categories.

- *Natural sciences*

Intute categorized some documents into "Biological Sciences" while BUBL has "Biology" in the 570 Life sciences, biology category. BUBL also separates "Biology" from "Plants" in two different categories. However, it is important to note that two pairs of terms indexed by BUBL and Intute, "Plants" versus "Botany" and "Animals" versus "Zoology" are different terms but could be comparable. For example, "Botany" is defined as "the science of plants" while "Zoology" is the study of animals. In the Library of Congress Subject Headings, those two pairs of terms are defined as "Related Terms (RT)".

- *Technology*

In terms of Technology subject, Intute placed many documents on Technology into “Arts” or “Physical sciences”.

7.1.3 Conclusions on research question 4 and 5

RQ 4) How consistent is tagging/indexing between Delicious taggers and Intute professionals?

For all three measures, there was relatively high consistency concerning the Language subject but relatively low consistency regarding the Technology subject between Delicious tagging and Intute professional indexing. Low consistency in the Technology subject was due to different levels of indexing between Intute indexers and Delicious taggers. For example, regarding the document 610 *Medical sciences, medicine,* Intute keywords tend to be broader terms, i.e., “disease” and “patient education,” but Delicious tags consist of terms in various semantic relationships, e.g., broader terms or narrower terms. The tags on the document 610 *Medical sciences, medicine* include “health”, “medical”, “medicine”, “drugs”, “healthcare” etc. In the Library of Congress Subject Headings (LCSH), a term “health” is represented as “narrower terms” of that term “medicine”. The term “healthcare” does not exist in the LCSH, but an alternative term “medical care” is represented as a narrower term of the term “public health” which is a narrower term of “health”.

RQ 5) Would Delicious users' tags provide additional subject access points beyond index terms or keywords that Intute professionals provide?

In answer to question 5, the in-depth examination of tags and Intute keywords was performed. The results of indexing consistency on Delicious taggers and Intute professionals reported that there was low similarity in Sociology and Literature. It was revealed that tags included additional access points with many newly-coined terminology such as ebook, online, web, web 2.0, e-guides, e-learning and cyberspace which reflect more accurate descriptions on the web documents. On the other hand, in Natural Sciences, Delicious tags additionally supply users' preferred or up-to-date terms. Examples are "bioinformatics", "bioscience", "biotech" for the term "biotechnology", and "cheminformatics" for "chemistry".

7.1.4 Conclusions on research question 6

RQ 6) What levels of tagging exhaustivity and tag specificity in Delicious characterize the indexing of web documents?

The answer to question 5 reported that the Delicious tags provided additional access points, which resulted in low indexing consistency between Delicious tags and Intute keywords. Chapter 5 provided the results of the analysis on tagging exhaustivity and tag specificity in order to investigate if those tags as additional access points would be really powerful to differentiate each document among other documents, and if those tags would represent a high degree of quality.

Chapter 5 provided the answer to question 6 by associating tag exhaustivity with the degree of indexing consistency. The average number of tags per documents over all subjects ranges from 2.5 to 3.3 tags. Even though the high recall does not imply high quality of indexing, it was interesting to notice that regarding subject categories showing relatively higher indexing consistency, the tagging exhaustivity was relatively higher than for other subject categories.

To provide an in-depth analysis of tagging accuracy, BUBL and Intute indexers' descriptions were reviewed and compared with tags at various levels of specificity. This dissertation research employed the Latent Semantic Analysis (LSA) method to examine whether the tags describe documents accurately, i.e., how well those tags represent important concepts related to the documents. The term-to-term comparison (between professional keywords and user tags) was conducted concerning documents that showed the lowest similarity between Intute professionals' keywords and Delicious users' tags. To investigate whether there is a relationship between tag specificity and LSA values, this research conducted a correlation analysis. The results of correlation analysis indicated that there is a positive linear relationship between tag specificity and the latent semantics of tags to professionals' keywords. Although the values of the correlation coefficients showed a weak relationship between two variables, it is still important that there *is* a positive or increasing relationship between them. In other words, tags with high specificity value have more semantic similarity with professionals' keywords. Tags with high specificity value are considered to be good discriminators, and professionals' keywords are regarded as accurate index terms. This leads to the conclusion that the term's power as a discriminator is related to its semantic relatedness to documents.

7.1.5 Conclusions on research question 7

> **RQ 7**) What are features and patterns of social tagging in describing a web document at Delicious? Do tags have other bibliographic attributes beyond describing subjects or topics of a document?

In order to characterize the features and patterns of tags, the content analysis of tag attributes was performed. The process of identifying bibliographic attributes of tags was based on the FRBR model which defines attributes as "logical analysis of the data that are typically reflected in bibliographic records" (IFLA, 1998). To enhance confidence in reliability, a coder other than the researcher was recruited and the coded results were compared to measure the intercoder reliability. To provide more credible and objective results on tag coding, this research employed four different indices for intercoder reliability: Holsti's method, Scott's pi (p), Cohen's kappa (κ) and Krippendorff's alpha (α). The findings identified the bibliographic attributes of tags beyond describing subjects or topics of a document. The findings also showed that tags have essential attributes matching those defined in FRBR.

Chapter 6 reported that among tags assigned to the sampled documents, 26 % of tags were subject-related terms, 27 % of tags were matched into the attributes of FRBR, and 47 % of tags were categorized into other attributes. It implies that many tags (about 74 %) include additional properties beyond subject or topic terms. The results also showed that taggers tend to mainly assign tags on attributes related to WT (Title attribute of FRBR Work entity) and WF (Form attribute of FRBR Work entity). Furthermore, regarding 10 different subject categories, the tag frequency on FRBR attributes showed different tendencies. For three subject areas, Technology, Arts and Literature, tag frequency on the FRBR WI (intended audience) attribute was relatively high, which means that taggers tend to consider audience in these subject areas. In terms of Natural sciences and Geography, the finding on tag frequency of EF (Form attribute of Expression entity) category showed relatively high proportion in comparison with those of other subject categories. This suggests that web documents in both Natural sciences and Geography are characterized with a focus on specific forms by taggers. The other attributes of tags were sorted out into three sub categories, Feature, Utilization, and Institution. Furthermore, in terms of specific subject areas, taggers exhibited different tagging behaviors representing distinctive features and tendencies. These results have led to the conclusion that there should be an increased awareness of diverse user needs by subject in the process of metadata generation.

7.2 Contributions

This dissertation research has investigated social tagging in several ways in order to verify its quality and efficacy. The main contributions of this dissertation are as follows:

- **Combination of quantitative and qualitative approach to investigating social tags**

This study combined both quantitative (statistics) and qualitative (content analysis) approaches to vocabulary analysis of tags which provided a more complete examination of the quality of tags. Several researchers have discussed social tagging behavior and its usefulness for classification or retrieval. However, there was no systematic research which employed quantitative as well as qualitative analysis for the comparison of social tagging with professional indexing in terms of subject indexing of documents. The analysis of social tagging was divided into three phases: analysis of indexing consistency, analysis of tagging effectiveness, and analysis of tag attributes. Indexing consistency was assessed using the Inter-indexer Consistency Density formula with three different similarity measures (Euclidean distance, dot product, cosine).   An analysis of tagging effectiveness with tag specificity was conducted to ameliorate the drawbacks of consistency analysis based on only the quantitative measures of vocabulary matching.   Finally, the process of identifying bibliographic attributes of tags was based on the Functional Requirements for Bibliographic Records (FRBR) model.

- **Analyzing social tags using FRBR**

This dissertation conducted a FRBR-based qualitative analysis of tag attributes combined with quantitative approaches, which led to a clearer examination of the quality of tags. The finding implies that the tagging pattern was different regarding 10 different subjects and it leads to the practical implication that there should be a careful consideration of user needs when assigning metadata.

- **Demonstrating the potential of tags for web site indexing**

As noted in Chapter 2.2 *Organization of the Web*, both BUBL and Intute have lost the funding needed to sustain their operations and provide professional-level indexing of web sites. Through the detailed analysis of tag properties undertaken in this dissertation, we have a clearer understanding of the extent to which social tagging can be used to replace (and in some cases to improve upon) professional indexing. By treating the results of tagging as indexing and exploring indexing consistency, exhaustivity, specificity, semantic quality, and the applicability of FRBR categories, this dissertation research has provided a basis for further study of the strengths and limitations of tagging. This is particularly critical given the decline in support for professional indexing at the same time that web resources continue to proliferate and the need for guidance in their discovery and selection remains.

7.3 Limitations

This study limited the scope of sample web documents to the common document collection of BUBL and Intute, and only if a web document was listed at both locations were tags assigned to the web document at Delicious collected and analyzed. Thus, conclusions about properties of tags in Delicious were limited to web documents selected for inclusion in subject gateways and indexed by professional indexers. In addition analysis for tag specificity and content analysis of tag attributes focused on the top 20 ranked tags. A more thorough study of tagging behavior would encompass a larger number of assigned tags associated with each document.

7.4 Future Research

This study has identified a number of potentially fruitful directions for further study of social tags as indexing. Topics in need of further study include:

- **Analyzing personal or emotional tag attributes**

For the analysis of social tags, this dissertation developed the stoplist or a set of terms which was excluded for processing (see Appendix C). The stoplist included an explicit list of the terms that Sen et al. (2006) define as subjective and personal tags, since those types of tags are not meaningful for indexing subjects of documents.   However, in terms of categorizing tag types, subjective or emotional tags could also be crucial metadata describing important factors represented in the document. In Chapter 6, the “Others” category, resulted from the analysis of tag attributes using FRBR, included some tags showing subjective or personal properties. Those tags such as resources, learning, teaching, and job imply user’s intent to use documents for particular purposes. In future research, therefore, the examination of personal tags will be conducted. In addition, a survey or user study on tagging behavior would help to extend understanding of social tagging practices.

- **Developing user vocabulary in health information organization**

The findings on indexing consistency over all subjects revealed that there was relatively low consistency in Medical and Health Information subjects between taggers and professional indexers. The analysis of tagging effectiveness also demonstrated that tags were accurate expressions for the topics in that area. Information in health or medical areas is very critical and should be accessible to users without difficulty. However, the growing amount of health information on the web has increased concern about effective access to quality health information because terminology, currently used for organizing health information, is generated by professionals and is not familiar to users. Thus, my future work will develop a consumer health thesaurus reflecting user needs and user-preferred terms by investigating social tags assigned to web documents in the health domain.

- **Redefining FRBR attributes for the web environment**

The results found in this dissertation revealed that there was some disagreement between two coders on two FRBR attribute categories, i.e., WF (Form of Work entity) and EF (Form of Expression entity). The examples of those tags were Books, Database, Magazine, Journal, and Encyclopedia. This disagreement on those attributes was caused by the fact that the documents, tagged with a term “Book”, include the list of books or provide a feature of searching for books rather than books themselves. However, current definitions provided by FRBR do not explicitly distinguish these two attributes about web documents. To make FRBR more applicable, FRBR should be able to describe digital heterogeneous media resources which are available in various formats and multi-dimensional structures. Therefore, an important future direction for my research will involve expanding current FRBR definitions on entities and attributes for web documents in digital environments.

BIBLIOGRAPHY

Abbott, R. (2004). Subjectivity as a concern for information science: A Popperian perspective. *Journal of Information Science*, 30(2): 95-106.

Agresti, A. & Finlay, B. (1999). *Statistical Methods for the Social Sciences*. 3rd ed. Harlow: Pearson Education.

Bakeman, R. (2000). Behavioral observation and coding. In H. T. Reis & C. M. Judge (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 138-159). New York: Cambridge University Press.

Banerjee, M., Capozzoli, M., McSweeney, L. & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1): 3–23.

Bao, S., et al. (2007). Optimizing web search using social annotations. *Proceedings of the 16th international conference on World Wide Web*. Retrieved from http://www2007.org/papers/paper397.pdf

Bawden, D. & Robinson, L. (2002). Internet subject gateways revisited. *International Journal of Information Management*, 22(2): 157-162.

Boyd, D. (2005). Issues of culture in ethnoclassification/folksonomy. *Many-to-Many*. Retrieved from, http://www.corante.com/many/archives/2005/01/28/issues_of_culture_in_ethnoclassificationfolksonomy.php

Burton, P. & Mackie, M. (1999). The use and effectiveness of the eLib subject gateways: a preliminary investigation. *Program: electronic library & information systems*, 33(4): 327-337.

Caras , G.J. (1968). Indexing from abstracts of documents. *Journal of Chemical Documentation*, 8 (1): 20-22.

Chen, Xu. (2008). *Indexing consistency between online catalo*gues. Dissertation, Humboldt Universitat zu Berlin.

Choy, S.O. & Lui, A.K. (2006). Web information retrieval in collaborative tagging systems. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*.

Chun, S. & Jenkins, M. (2005). Cataloguing by Crowd; a proposal for the development of a community cataloguing tool to capture subject information for images (A Professional Forum). *Museums and the Web 2005*, Vancouver. Retrieved from http://www.archimuse.com/mw2005/abstracts/prg_280000899.html

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20: 37-46.

Conover W. J. (1980). *Practical Non-Parametric Statistics*. 2nd edn. New York: John Wiley and Sons.

Cooper, W. S. (1969). Is interindexer consistency a hobgoblin? *American Documentation*, 20(3): 268-278.

Craig, R. T. (1981). Generalization of Scott's Index of Intercoder Agreement. *Public Opinion Quarterly*, 45(2): 260-264.

Crichton, N.J. (1999). Information point: Spearman's rank correlation. *Journal of Clinical Nursing,* 8: 763.

Crockford, D. (2006). *The application/json Media Type for JavaScript Object Notation (JSON).* Retrieved from http://www.ietf.org/rfc/rfc4627.txt?number=4627

David, C., Giroux, L., Bertrand-Gastaldy S. & Lanteigne, D. (1995). Indexing as problem solving: A cognitive approach to consistency. *In Proceedings of the ASIS Annual Meeting,* Medford, NJ, pp. 49-55.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science,* 41: 391-407.

Dempsey, L. (2000). The subject gateway: experiences and issues based on the emergence of the Resource Discovery Network. *Online Information Review,* 24(8): 8-23.

DESIRE Consortium. 2000. Retrieved from http://www.desire.org/

Dewey, M. E. (1983). Coefficients of agreement. *British Journal of Psychiatry*, 143: 487-489.

Dubois, C. P. R. (1987). Free text vs. controlled vocabulary: A reassessment. *Online Review,* 11(4): 243-253.

Fidel, R. (1991). Searchers' selection of search keys: II. Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science*, 42(7): 501-514.

Furnas, G. W., Landauer, T. K., Gomez, L. M. & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM,* 30(11): 964-971.

Furner, J. & Tennis, J. T. (2006). *Advances in Classification Research, Volume 17: Proceedings of the 17th ASIS&T Classification Research Workshop,* Austin, Texas, USA.

Gold, J. (1996). Introducing a new service from BUBL [Libraries of Networked Knowledge]. *The Serials Librarian*, 30(2): 21-26.

Golder, S. & Huberman, B. A. (2005). The structure of collaborative tagging systems. Retrieved from http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf.

Golder, S. A. & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*. 32 (2): 198-208.

Golub, K. (2006). Using controlled vocabularies in automated subject classification of textual web pages, in the context of browsing. *IEEE TCDL Bulletin,* 2(2): 1-11. Retrieved from http://www.ieee-tcdl.org/Bulletin/v2n2/golub/golub.html

Hassan-Montero, Y. & Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. *Proceedings of Multidisciplinary Information Sciences and Technologies*, InSciT2006Merida, Spain: October, 2006.

Hayes, A. F. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1(1): 77-89.

Hayman, S. (2007). Folksonomies and tagging: New developments in social bookmarking. *Ark Group Conference: Developing and Improving Classification Schemes 27-29 June, Rydges World Square, Sydney.* p.18. Retrieved from http://www.educationau.edu.au/jahia/webdav/site/myjahiasite/shared/papers/arkhayman.pdf.

Heymann, P., Koutrika, G, & Garcia-Molina, H. (2008). Can Social Bookmarking Improve Web Search*? Proceedings of the 1st International Conference on Web Search and Data Mining*.

Hiom, D. (2006). Retrospective on the RDN. *Ariadne.* 47. Retrieved from http://www.ariadne.ac.uk/issue47/hiom/

Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.

Hooper, R.S. (1965). *Indexer consistency tests-Origin, measurements, results and utilization*. IBM, Bethesda, Md., IBM Corp., p. 3. (TR 95-56)

Horri, A. & Neshat, N. (2006). A study of subject indexing consistency between the National Library of Iran and humanities libraries in the area of Iranian studies. *Cataloging & Classification Quarterly,* 43(1): 67–76.

IFLA Study Group. (1998). *Functional requirements for bibliographic records: final report* München: K.G. Saur.

Joint Information Systems Committee. (JISC). http://www.jisc.ac.uk/

Joyce, A. M., Wickham, J., Cross, P. & Stephens, C. (2008). Intute integration. *Ariadne*, Issue 55, April 2008. Retrieved from http://www.ariadne.ac.uk/issue55/joyce-et-al/

Keen, E. M. & Digger, J. A. (1972). *Report of an Information Science Index Languages Test*. Aberystwyth College of Librarianship, Wales.

Kim, G. (2006). Relationship between index term specificity and relevance judgment. *Information Processing and Management*, 42(5): 1218-1229.

Kipp, M.E., & Campbell, D.G. (2010). Searching with Tags: Do Tags Help Users Find Things?, 2010. *Knowledge Organization*. 37(4): 239-255.

Kipp, M.E. (2011). User, author and professional Indexing in context: An exploration of tagging practices on CiteULike. *Canadian Journal of Information and Library Science*. 35(1): 17-48.

Knapp, S. D., Cohen, L. B. & Juedes, D. R. (1998). A natural language thesaurus for the humanities: The need for a database search aid. *The Library Quarterly*, 68(4): 406-430.

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

Kohonen, T. (1995). *Self-Organizing Maps*. Berlin: Springer-Verlag.

Krippendorff, K. (1978). Reliability of binary attribute data. *Biometrics*, 34: 142-144.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

Krippendorff, K. (1987). Association, agreement, and equity. *Quality and Quantity*, 21: 109-123.

Krippendorff, K. (2004) *Content Analysis: An introduction to Its Methodology*, 2nd edn. Beverly Hills CA: Sage, Chapter 11.

Krippendorff, K. (2008). Testing the reliability of content analysis data: what is involved and why. In K. Krippendorff and M.A. Bock (eds.). *The Content Analysis Reader* (pp. 350-357). Thousand Oaks, CA: Sage.

Lancaster, F. W. (1972). *Vocabulary Control for Information Retrieval*. Washington, D.C.: Information Resources Press.

Lancaster, F, W. (2003). *Indexing and Abstracting in Theory and Practice.* 3rd ed. Champaign, IL: University of Illinois.

Lancaster, F. W. & Smith, L. C. (1983). *Compatibility issues affecting information systems and services*. Paris: United Nations Educational, Scientific, and Cultural Organization.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review,* 104: 211-240.

Landbeck, C. (2007). Trouble in paradise: Conflict management and resolution in social classification environments. *Bulletin of the American Society for Information Science and Technology,* 34(1): 16-20.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159-74.

Leonard, L. E. (1977). "*Inter-indexer consistency studies, 1954-1975: a review of the literature and summary of study results*". Champaign, IL: University of Illinois, Graduate School of Library Science. Occasional Papers. No. 131, December 1977. p.51.

Lin, X., Beaudoin, J. E., Bui, Y. & Desai, K. (2006). Exploring characteristics of social classification. Advances in Classification Research, Volume 17; *Proceedings of the 17th ASIS&T Classification Research Workshop*, Austin, Texas, USA.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2005). *Practical resources for assessing and reporting intercoder reliability in content analysis research projects.* Retrieved from http://astro.temple.edu/~lombard/reliability/

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2): 159–165. The article is also included in *H. P. Luhn: Pioneer of Information Science, Selected Works.*

Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organization and resource discovery tool. *Library Review*, 55(5): 291-300.

Mai, J.-E. (2004a). Classification of the Web: challenges and inquiries. *Knowledge Organization*, 31(2): 92–97.

Mai, J.-E. (2004b). Classification in context: relativity, reality, and representation. *Knowledge Organization*, 31(1): 39-48.

Makani, J. & Spiteri, L.F. (2010). The dynamics of collaborative tagging: An analysis of tag vocabulary. *Journal of Information and Knowledge Management*, 9(2), pp.93-103

Maltby, A. (1975). *Sayers' Manual of Classification for Librarians*. 5th ed. London: Andre Deutsch.

Maron, M. E. (1977). On indexing, retrieval and the meaning of "about". *Journal of the American Society for Information Science,* 28: 38-43.

Medelyan, O. & Witten, I. H. (2006). Measuring inter-indexer consistency using a thesaurus. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2006, Chapel Hill, NC, USA, June 11 - 15, 2006 (pp. 296-297). New York: ACM.

Merholz, P. (2004). *Metadata for the Masses, adaptive path*. Retrieved from http://www.adaptivepath.com/ideas/e000361

Merriam-Webster's online dictionary. Retrieved from http://www.merriam-webster.com/spanish/compras

Metzler, D. & Croft, W.B. (2006). Beyond Bags of Words: Modeling Implicit User Preferences in Information Retrieval. *Proceedings of AAAI'06.*

Morville, P. (2005). *Ambient Findability: What we find changes who we become*. Cambridge: O'Reilly.

Muddamalle, M. R. (1998). Natural language versus controlled vocabulary in information retrieval: A case study in soil mechanics. *Journal of the American Society for Information Science,* 49(10): 881-887.

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, CA: Sage.

Nicholson, D. et al. (2001). *HILT: High Level Thesaurus Project: Final Report*. Retrieved from http://hilt.cdlr.strath.ac.uk/Reports/Documents/HILTfinalreport.doc

Noruzi, A. (2006). Folksonomies: (Un) Controlled vocabulary? *Knowledge Organization,* 33(4): 199-203.

Nowick, E. A. & Mering, M. (2003). Comparisons between Internet users' free-text queries and controlled vocabularies: a case study in water quality. *Technical Services Quarterly,* 21(2): 15-32.

Olson, H. A. & Boll, J. J. (2001). *Subject analysis in online catalogs*. 2nd ed. Englewood, Colorado: Libraries Unlimited.

Olson, H. & Wolfram, D. (2006), Indexing Consistency and its Implications for Information Architecture: A Pilot Study. *IA Summit.*

Paek, T. & Chandrasekar, R. (2005). Windows as a second language: an overview of the jargon project. *Proceedings of the First International Conference on Augmented Cognition.* Retrieved from http://research.microsoft.com/en-us/um/people/timpaek/papers/augcog2005.pdf

Pearson, A. V. & Hartley, H. O. (1972). *Biometrica Tables for Statisticians, Vol 2*, Cambridge, England: Cambridge University Press.

Peterson, E. (2006). Beneath the Metadata: Some philosophical problems with folksonomy. *D-Lib Magazine*, 12(11). Retrieved from http://www.dlib.org/dlib/november06/peterson/11peterson.html

Quintarelli, E. (2005). Folksonomies: power to the people. *Proceedings of the 1st International Society for Knowledge Organization* (Italy) (ISKOI), UniMIB Meeting, June 24, Milan, Italy, ISKOI, Italy. Retrieved from http://www.iskoi.org/doc/folksonomies.htm

Reade, M. & Romaniuk, B. (2005). *Acronyms, initialisms & abbreviations dictionary: a guide to acronyms, abbreviations, contractions, alphabetic symbols, and similar condensed appellations*. 35th. ed. Detroit : Thomson/Gale.

Renear, A. H. & Choi, Y. (2006). Modeling Our Understanding, Understanding Our Models: The Case of Inheritance in FRBR. In *Proceedings of the Annual Meeting of the American Society for Information Science.* November 3-8, 2006, in Austin, TX

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management*, 17: 69-76.

Salton, G., Wong, A. & Yang, C.S. (1975a). A vector space model for automatic indexing. *Communications of the ACM*, 18(11): 613-620.

Salton G., Yang, C. S. & Yu, C. T. (1975b). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1): 33-44.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 17: 321-325.

Sen, S. et al. (2006). Tagging, communites, vocabulary, evolution. *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. Retrieved from http://www.grouplens.org/papers/pdf/sen-cscw2006.pdf

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4): 591-611.

Shirky, C. (2005a), *Ontology is Overrated: Categories, Links and Tag*s, Shirky.com, New York, USA. Retrieved from http://shirky.com/writings/ontology_overrated.html

Shirky, C. (2005b), *Semi-Structured Meta-data has a Posse: A response to Gene Smith, You're It!* A Blog on Tagging. Retrieved from http://tagsonomy.com/index.php/semi-structured-meta-data-has-a-posse-aresponse- to-gene-smith/

Slamecka, V. & Jacoby, J. J. (1963). Effect of indexing aids on the reliability of indexers. *Final technical note*. Bethesda, MD, Documentation Inc. RADC-TDR-63-116.

Smith, G. (2004). *Folksonomy: social classification*. Atomiq / information Architecture [blog]. Retrieved from http://atomiq.org/archives/2004/08/folksonomy_social_classification.html.

Smith, T. (2007). *Cataloging and you: Measuring the efficacy of a folksonomy for subject analysis*. In Lussky, Joan, Eds. Proceedings of the 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, Wisconsin. Retrieved from http://dlist.sir.arizona.edu/2061

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28: 111-121.

Spiteri, L. F. (2005). *Controlled Vocabularies and Folksonomies*. Presentation at Canadian Metadata Forum, Ottawa, ON, September 27, 2005. p. 23. Retrieved from http://www.collectionscanada.ca/obj/014005/f2/014005-05209-e-e.pdf.

Spiteri, L.F. (2007). The structure and form of folksonomy tags: The road to the public library catalog. *Information Technology and Libraries*. 26(3): 13-25.

Strutz. D. N. (2004). *Communal Categorization: The Folksonomy*. INFO622: Content Representation.

Taylor, J. & Watkinson, D. (2007). Indexing reliability for condition survey data. *The Conservator*, 30: 49-61.

Tennis, Joseph T. (2006). Social Tagging and the Next Steps for Indexing. In Furner, Jonathan and Tennis, Joseph T., Eds. *Proceedings 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*, Austin, Texas.

Tonta, Y. (1991). A study of indexing consistency between Library of Congress and British Library catalogers. *Library Resources & Technical Services*, 35(2): 177-185.

Trant, J. (2006). Social Classification and Folksonomy in Art Museums: early data from the steve.museum tagger prototype. Advances in Classification Research, 17. p.19 *Proceedings of the 17th ASIS&T Classification Research Workshop*, Austin, TX.

Trant, J. (2009). Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information*, 10(1). Retrieved from http://journals.tdl.org/jodi/article/viewDownloadInterstitial/269/278

Turney, P. D. & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21 (4): 315–346.

University of Kent (2009). *Library Services Subject Guides*. Retrieved from http://www.kent.ac.uk/library/subjects/healthinfo/subjgate.html

van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd ed. London: Butterworths. Online edition 1999. Retrieved from http://www.dcs.gla.ac.uk/~iain/keith/

Vander Wal, T. (2005a). Folksonomy Definition and Wikipedia. *Off the Top*. Retrieved from http://www.vanderwal.net/random/entrysel.php?blog=1750

Vander Wal, T. (2005b). *Explaining and Showing Broad and Narrow Folksonomies*. Retrieved from http://www.personalinfocloud.com/2005/02/explaining_and_.html

Vander Wal, T. (2007). *Folksonomy Coinage and Definition*. Retrieved from http://www.vanderwal.net/folksonomy.html

Voss, J. (2007). Tagging, Folksonomy & Co - Renaissance of Manual Indexing? *Proceedings of the International Symposium of Information Science*, p. 234-254. Retrieved from http://arxiv.org/PS_cache/cs/pdf/0701/0701072v2.pdf .

Weber, J. (2006). *Folksonomy and controlled vocabulary in LibraryThing*. Unpublished Final Project, University of Pittsburgh.

Weinberger, D. (2006). *Beneath the Metadata* - a reply. Joho the Blog [blog] Retrieved from http://www.hyperorg.com/blogger/mtarchive/beneath_the_metadata_a_reply.html

Wikipedia: The free encyclopedia. (2009). FL: Wikimedia Foundation, Inc. Retrieved from http://www.wikipedia.org

Wimmer, R., & Dominick, J. (1987). *Mass Media Research: An Introduction.* 2nd Edition. Belmont, CA: Wadsworth Publishing Company.

Wolfram, D., & Olson, H.A. (2007). A Method for Comparing Large Scale Inter-indexer Consistency Using IR Modeling. *Proceedings of the 35th Annual Conference of the Canadian Association for Information Science*.

Wong , S. K. M., Ziarko, W. & Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval, *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, p.18-25, June 05-07, 1985, Montreal, Quebec, Canada

Yanbe, Y., Jatowt, A., Nakamura, S., & Tanaka, K. (2006). Can social bookmarking enhance search in the web? *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries.*

Zhou, X., Zhang, X. & Hu, X. (2007). *The Dragon toolkit developer guide*. Data Mining and Bioinformatics Laboratory, Drexel University. Retrieved from http://www.dragontoolkit.org/tutorial.pdf, 2007.

Zunde, P. & Dexter, M.E. (1969). Indexing consistency and quality. *American Documentation,* 20(3): 259-267.

APPENDIX A. LIST OF THE SELECTED WEB DOCUMENTS

Table 45. List of the selected web documents

Subject	Title and URL
001 Knowledge, humanities and research	Institute for Psychohistory: http://www.psychohistory.com/
002 The book	Book Arts Web: http://www.philobiblon.com/
003 Systems	N/A
004-006 Computing and the Internet / 004 computer_science	Communications of the ACM: http://www.acm.org/pubs/cacm
010 Bibliography	N/A
020 Library and information sciences	ASIS: American Society for Information Science and Technology: http://www.asis.org/
030 General encyclopaedic works	Encyclopaedia Britannica: http://www.britannica.com
060 General organisations and museology	ICOM: International Council of Museums: http://icom.museum/
070 News media, journalism, publishing/ 070.5 Publishers and publishing	Amazon.com: www.amazon.com
080 General collections	Index to Theses: http://www.theses.com/
090 Manuscripts, rare books, other rare printed materials	N/A
100 Philosophy, general resources	Philosophy: philosophy.eserver.org/
100 Philosophy, departments	N/A
100 Philosophy, journals	NOESIS: Philosophical Research Online, http://noesis.evansville.edu/
100 Philosophy, societies	N/A
107 Philosophy education	N/A
110 Metaphysics	N/A
120 Epistemology, causation, humankind	N/A
130 Paranormal phenomena	N/A
140 Specific philosophical schools and viewpoints	Karl Popper Web: http://www.eeng.dcu.ie/~tkpw/
150 Psychology, general resources	CogPrints: http://cogprints.org/
160 Logic	N/A
170 Ethics (moral philosophy)	Ethics Updates: http://ethics.sandiego.edu/
180 Ancient, medieval, oriental philosophy	N/A
190 Modern Western philosophy	Eighteenth Century Philosophy Resources, http://andromeda.rutgers.edu/~jlynch/18th/
200 Religion, general resources	Virtual Religion Index, http://virtualreligion.net/vri/
200 Religion, education and research	N/A
210 Philosophy and theory of religion	Philosophy of Religion Info, http://www.philosophyofreligion.info/
220 Bible, general resources	BibleGateway.com
230 Christianity, general resources	New Testament Gateway, http://www.ntgateway.com/

Table 45 (cont.)

230 Christianity, journals	N/A
234.161 Baptism	N/A
238 Christian creeds and catechisms	N/A
242 Christian writings	Christian Classics Ethereal Library, http://www.ccel.org/
246 Christian art	Religion and the Founding of the American Republic , http://www.loc.gov/exhibits/religion/religion.html
252 Texts of sermons	N/A
268 Christian education	N/A
270 History of Christianity	Christian Catacombs of Rome, http://www.catacombe.roma.it/
280 Christian denominations	N/A
294.3 Buddhism	Access to Insight, Readings in Theravada Buddhism: http://www.accesstoinsight.org/
294.4 Jainism	N/A
294.5 Hinduism	Bhagavad Gita, http://www.bhagavad-gita.org/
294.6 Sikhism	N/A
295 Zoroastrianism	Avesta Zoroastrian Archives, http://www.avesta.org/
296 Judaism	Judaism and Jewish Resources, http://www.shamash.org/trb/judaism.html
297 Islam	Islam Online, http://www.islamonline.com/
297.93 Baha'i Faith	N/A
299.514 Taoism	N/A
299 Other religions	N/A
300 social_sciences, general resources	Online Dictionary of the social sciences, http://bitbucket.icaap.org/
301 Sociology, general resources	Sociological Tour Through Cyberspace, www.trinity.edu/~mkearl/index.html
310 International statistics	IDB Population Pyramids: International Data Base (IDB) - Pyramids , http://www.census.gov/ipc/www/idb/pyramids.html
320 Political science, general resources	Glossary of Political Economy Terms, http://www.auburn.edu/~johnspm/gloss/
330 Economics, general resources	History of Economic Thought, http://cepa.newschool.edu/het/
340 Law, general resources	WashLaw Web, http://www.washlaw.edu/
350 Public administration, general resources	Jane's Information Group, http://www.janes.com/
355 Military science, general resources	DOD Dictionary of Military Terms, http://www.dtic.mil/doctrine/jel/doddict/
360 Social problems and services, associations	Centre for Policy Studies, http://www.cps.org.uk/
370 Education	Excellence Gateway, http://excellence.qia.org.uk/
380 Commerce, communications, transportation	globalEDGE International Business Resource Desk, http://globaledge.msu.edu/
390 Customs, etiquette, folklore	Costumer's Manifesto, Ethnic Dress, http://www.costumes.org
400 Language, general resources	iLoveLanguages, http://www.ilovelanguages.com/
401 Language, philosophy and theory	N/A
403 Dictionaries and encyclopaedias	Online Etymology Dictionary, http://www.etymonline.com/
407 Language education and research	CILT: Centre for Information on Language Teaching and Research, http://www.cilt.org.uk/
409 Languages spoken around the world	N/A
410 Linguistics	N/A

Table 45 (cont.)

411 Writing systems	N/A
414 Phonology and phonetics	International Phonetics Association, http://www.langsci.ucl.ac.uk/ipa/
415 Grammar	N/A
419 Sign language	American Sign Language Browser, http://commtechlab.msu.edu/Sites/aslweb/browser.htm
420 English language /423 English dictionaries	AskOxford.com, http://www.askoxford.com/ : www.askoxford.com/?view=uk
430 German language, general resources	Goethe Institutes, www.goethe.de/
439.31 Dutch language	N/A
439.7 Swedish language	N/A
439.82 Norwegian language	N/A
440 French language	ARTFL Project: French English Dictionary, http://humanities.uchicago.edu/orgs/ARTFL/
450 Italian language	N/A
459 Romanian language	N/A
460 Spanish language	N/A
469 Portuguese language	N/A
470 Latin language	N/A
480 Greek language	N/A
490 Other languages/ 495.1 Chinese language	On-line Chinese Tools, http://www.mandarintools.com/
500 Natural sciences, national centres	National Science Foundation, http://www.nsf.gov/
510 Mathematics, general resources	MathSciNet, http://www.ams.org/mathscinet/
520 Astronomy, general resources	Astronomy Picture of the Day, http://antwrp.gsfc.nasa.gov/apod/astropix.html
530 Physics, general resources	Albert Einstein Online, http://www.westegg.com/einstein/
540 Chemistry	Linux4Chemistry, http://www.redbrick.dcu.ie/~noel/linux4chemistry/
550 Earth sciences	GeoGuide, http://www.geo-guide.de/
560 Palaeontology, general resources	Museum of Paleontology, www.ucmp.berkeley.edu/
570 Life sciences, biology	BBSRC: Biotechnology and Biological_sciences Research Council, http://www.bbsrc.ac.uk/
580 Plants 580 Plants, general resources	Botanical Society of America Online Image Collection, http://images.botany.org/
590 Animals, general resources	Animal Diversity Web, http://animaldiversity.ummz.umich.edu/site/
600 Technology, general resources	EurekAlert, http://www.eurekalert.org/
610 Medical sciences, medicine	MedicineNet, http://www.medicinenet.com/script/main/hp.asp
620 Engineering, education and research	EDINA, http://edina.ac.uk/index.shtml
630 Agriculture and related technologies	AgNIC: Agriculture Network Information Center, http://www.agnic.org/
640 Home economics and family living	N/A
650 Management and auxiliary services	American Institute of Certified Public Accountants, http://www.aicpa.org/
660 Chemical_engineering	American Institute of Chemical Engineers, http://www.aiche.org/
670 Manufacturing	Bad Human Factors Designs, http://www.baddesigns.com/
680 Manufacture for specific uses	N/A
690 Buildings	Advanced Building Technologies, http://www.advancedbuildings.org/
700 The arts, general resources	Art deadlines list, http://artdeadlineslist.com/

Table 45 (cont.)

700 Fine and decorative arts, general resources	Arts in Context, http://www.artincontext.org/
700 Fine and decorative arts, artists	Gustav Klimt, http://www.expo-klimt.com/
701 Fine and decorative arts, philosophy	Aesthetics Online, http://www.aesthetics-online.org/
703 Fine and decorative arts, dictionaries and encyclopaedias	Artcyclopedia, http://www.artcyclopedia.com/
705 Fine and decorative arts, journals and magazines	Ceramics today, http://www.ceramicstoday.com/
706 Fine and decorative arts, organisations	The Arts & Crafts Society, http://www.arts-crafts.com/
707 Fine and decorative arts, art schools	Royal College of Art, http://www.rca.ac.uk/
708 Art galleries and museums in the UK	Warhol, http://www.warhol.org/
708 Art galleries and museums in the US	The Metropolitan Museum of Art, http://www.metmuseum.org/
708 Art galleries and museums worldwide	N/A
709 History of art	Futurism, http://www.unknown.nu/futurism/
710 Civic and landscape art	RUDI : Resource for Urban Design Information, http://www.rudi.net/
720 Architecture	American Institute of Architects, http://www.aia.org/
730 Plastic arts, sculpture	The First European Portal on Public Art, http://www.art-public.com/
740 Drawing and decorative arts	Computer Animation: From the Studio to the Home PC, http://animation.about.com/
750 Painting and paintings	Art Crimes: The Writing on the Wall, http://www.graffiti.org/
760 Graphic arts, printmaking and prints	Powers of Persuasion, http://www.archives.gov/exhibits/powers_of_persuasion/powers_of_persuasion_home.html
770 Photography and photographs	FotoFest , http://www.fotofest.org/index.htm
780 Music	Harmony Central, http://www.harmony-central.com/
790 Recreational and performing_arts, art	OSCAR.com - 81st Annual Academy Awards - Homepage, www.oscar.com/
796 Sport and outdoor activities	International Rugby Board - Home : http://www.irb.com
800 Literature, general resources	Literary Traveler: http://www.literarytraveler.com/
808.8 Literature, general collections	Google Book Search: http://books.google.com/
808 General rhetoric	Handbook of Rhetorical Devices : http://www.virtualsalt.com/rhetoric.htm
808.02 Authorship, writing and editorial techniques	Psychology with Style: http://www.uwsp.edu/psych/apa4b.htm
808.83 Fiction	N/A
808.838 Science fiction	Ultimate Science Fiction Web Guide http://www.magicdragon.com/UltimateSF/SF-Index.html
808.88 Quotations	A Dictionary of Scientific Quotations : http://naturalscience.com/dsqhome.html
808.899 Children's literature	CHILDE project - Children's Historical Literature Disseminated throughout Europe, http://www.bookchilde.org/
809.89 Women writers	Victorian women writers project, http://www.indiana.edu/~letrs/vwwp/index.html
809 Literary study and criticism	Literary history : http://literaryhistory.com/
809 Poetry, general resources	Favorite Poem Project, http://www.favoritepoem.org
810 American literature in English	N/A
810 Canadian literature	N/A
820 English, Scottish and Irish literature	Cambridge History of English and American Literature, http://www.bartleby.com/cambridge/
830 German literature	19th Century German Stories, http://www.fln.vcu.edu/menu.html
840 French literature	N/A

Table 45 (cont.)

850 Italian literature	Decameron Web, http://www.brown.edu/Departments/Italian_Studies/dweb/index.php
860 Spanish and Portuguese literature	N/A
870 Latin literature	N/A
880 Classical Greek literature	The Internet classics archives, http://classics.mit.edu/
890 Literature in other languages	Modern Haiku, http://www.modernhaiku.org/
900 History, general resources	Historical Text Archive: http://historicaltextarchive.com/
900 History, departments	N/A
900 History, journals	N/A
909 World history	Center for Jewish history : http://www.cjh.org/
910 Geography and travel	CountryWatch.com: http://www.countrywatch.com/
920 Biography	Exploring Leonardo: http://www.mos.org/sln/Leonardo/
929 Genealogy, names, insignia	FamilySearch Internet Genealogy Service: http://www.familysearch.org/
930 History of ancient world	English Heritage: http://www.english-heritage.org.uk
940 History of Europe	World War I Document Archive : http://wwi.lib.byu.edu/index.php/Main_Page
941-2 History of the British Isles	SCRAN: Scottish Cultural Resource Access Network: http://www.scran.ac.uk/
960 History of Africa	Story of Africa: http://www.bbc.co.uk/worldservice/africa/features/storyofafrica/
970 History of North America	Images Canada: http://www.imagescanada.ca/index-e.html
980 History of South America	Latin American Network Information Center, http://lanic.utexas.edu/
990 History of other parts of the world	Picture Australia: http://www.pictureaustralia.org/

APPENDIX B. INTER-INDEXER CONSISTENCY COMPUTATION

B.1 Overview of inter-indexer consistency calculator

A program is written for data acquisition, data pre-processing and the Inter-indexer Consistency Density calculation. This performs some of the most computationally intensive tasks and enables a study on large data sets. A block diagram of the program is depicted in Figure 20 (see Section 3.3.2).

B.2 Input Data

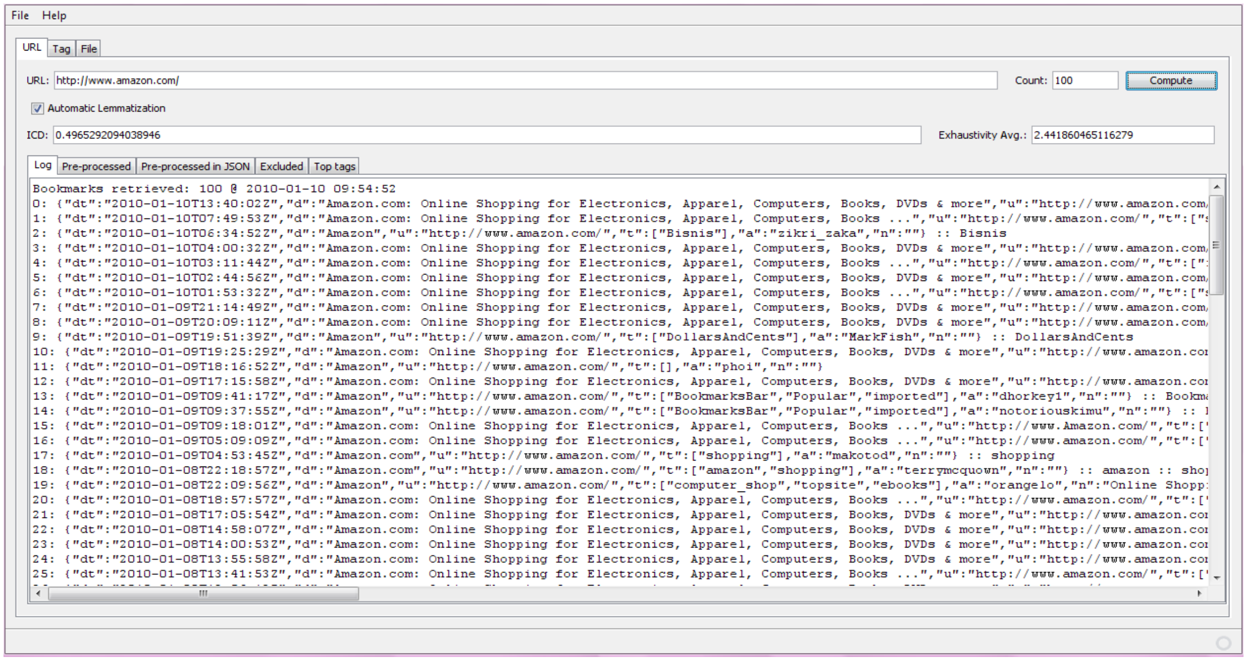
Input data will be a pair of URL and tags assigned to it in a JSON format [6]. JSON can represent name/value pairs in smaller size than Extensible Markup Language (XML) and many open source JSON parsers are available. A web document is associated with timestamp (dt), description (d), tags (t), author (a), and notes (n). A sample bookmark in JSON format is as follows:

```
{"dt":"2009-10-16T20:08:05Z","d":"Amazon.com: Online Shopping for
Electronics, Apparel, Computers, Books, DVDs &
more","u":"http://www.amazon.com/","t":["online","shopping","store","dvd","mu
sic","shop","books","entertainment"],"a":"cjpa318","n":""}
```

The collection of tagging data at Delicious was fully automated, since Delicious provides HTTP-based APIs (Application Programming Interface) for access to web documents and tagging data. The figure below shows collected tags through the Delicious API. On the other hand, for index

[6] JavaScript Object Notation (JSON) (Crockford, 2006)

terms from BUBL and Intute, JSON input was manually constructed.



URL: http://www.amazon.com/

Count: 100

☑ Automatic Lemmatization

ICD: 0.4965292094038946

Exhaustivity Avg.: 2.441860465116279

```
Bookmarks retrieved: 100 @ 2010-01-10 09:54:52
0: {"dt":"2010-01-10T13:40:02Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.com
1: {"dt":"2010-01-10T07:49:53Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books ...","u":"http://www.amazon.com/","t":["
2: {"dt":"2010-01-10T06:34:52Z","d":"Amazon","u":"http://www.amazon.com/","t":["Bisnis"],"a":"zikri_zaka","n":""} :: Bisnis
3: {"dt":"2010-01-10T04:00:32Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.com
4: {"dt":"2010-01-10T03:11:44Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books ...","u":"http://www.amazon.com/","t":["
5: {"dt":"2010-01-10T02:44:56Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.com
6: {"dt":"2010-01-10T01:53:32Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books ...","u":"http://www.amazon.com/","t":["
7: {"dt":"2010-01-09T21:14:49Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.com
8: {"dt":"2010-01-09T20:09:11Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.com
9: {"dt":"2010-01-09T19:51:39Z","d":"Amazon","u":"http://www.amazon.com/","t":["DollarsAndCents"],"a":"MarkFish","n":""} :: DollarsAndCents
10: {"dt":"2010-01-09T19:25:29Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.co
11: {"dt":"2010-01-09T18:16:52Z","d":"Amazon","u":"http://www.amazon.com/","t":[],"a":"phoi","n":""}
12: {"dt":"2010-01-09T17:15:58Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.co
13: {"dt":"2010-01-09T09:41:17Z","d":"Amazon","u":"http://www.amazon.com/","t":["BookmarksBar","Popular","imported"],"a":"dhorkey1","n":""} :: Bookma
14: {"dt":"2010-01-09T09:37:55Z","d":"Amazon","u":"http://www.amazon.com/","t":["BookmarksBar","Popular","imported"],"a":"notoriouskimu","n":""} :: 
15: {"dt":"2010-01-09T09:18:01Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books ...","u":"http://www.Amazon.com/","t":['
16: {"dt":"2010-01-09T05:09:09Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books ...","u":"http://www.amazon.com/","t":['
17: {"dt":"2010-01-09T04:53:45Z","d":"Amazon.com","u":"http://www.amazon.com/","t":["shopping"],"a":"makotod","n":""} :: shopping
18: {"dt":"2010-01-08T22:18:57Z","d":"Amazon.com","u":"http://www.amazon.com/","t":["amazon","shopping"],"a":"terrymcquown","n":""} :: amazon :: sho
19: {"dt":"2010-01-08T22:09:56Z","d":"Amazon","u":"http://www.amazon.com/","t":["computer_shop","topsite","ebooks"],"a":"orangelo","n":"Online Shopp
20: {"dt":"2010-01-08T18:57:57Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books ...","u":"http://www.amazon.com/","t":['
21: {"dt":"2010-01-08T17:05:54Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.co
22: {"dt":"2010-01-08T14:58:07Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.co
23: {"dt":"2010-01-08T14:00:53Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.co
24: {"dt":"2010-01-08T13:55:58Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more","u":"http://www.amazon.co
25: {"dt":"2010-01-08T13:41:53Z","d":"Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books ...","u":"http://www.amazon.com/","t":['
```

Figure 75. A screenshot of collected tags in the program

B.3 Phases of tag processing

Processing tagging data in the program is conducted through three main phases: (1) pre-processing tags, (2) constructing the Indexer/Tagger space, and (3) computing inter-indexer consistency density. Pre-processing is only necessary for uncontrolled tagging data from Delicious.

B.3.1 Pre-processing tags

After automatically extracting most recently assigned tagging data from up to 100 taggers, the preliminary processing on the tags will be performed. This was based on the rules of exact match between terms which is described in section 3.3.1. Additionally, a web document without any tag was excluded.

B.3.2 Constructing Indexing Space

For each valid web document, an indexer-tag pair is constructed as a vector and all the resulting vectors constitute an indexing space.

B.3.3 Computing data

The program calculates the indexing centroid by computing distances between centroid and an individual vector, and finally generates the value of Inter-Indexer (Tagger) Consistency Density.

APPENDIX C. STOPLIST

For the comparison of tags and professionals' indexing terms, this research developed a stoplist or a list of terms which can be excluded for processing (Table 46). The stoplist included an explicit list of the terms that Sen et al. (2006) define as subjective and personal tags, since those types of tags are not meaningful for indexing subjects of documents.

Table 46. Stoplist

affordable
awesome
babyas
bad
base
befolkning
best_of_the_web
bestoftheweb
bookmarksbar
bourse
by
ccstuff
cdweb
check
collectibles
convenient
cool
download
fact
favorite
for_student
free
free.to.everyone
funny
good
good_info
good_information
good_practice
gooddesign
good-design
goodinfo
goodpractice
grad
grad_school
gradschool
guide
help
how_things_work
howto
humor
interesting
informative
personal
popular
portal
post_graduate
postgraduate
prekindergarten
pre-k-kindergarten
professional
professional_resource
read_later
recommend
recommended_site
recommendedsite
ref_source
search
self-help
sharing
staring_site
startingsite
student
stumbleupon
stumbleuponfavorite
tip
to.read
toread
to_be_better_tagged
toblog
tocatalog
todescribe
toread
useful
useful_link
useful_stuff
usefulstuff
vital_record
vitalrecord
worth
wow

APPENDIX D. CODING SCHEME FOR TAG ATTRIBUTES DURING CONTENT ANALYSIS

D.1 List of FRBR attributes to apply

Table 47. List of FRBR attributes to apply

Entities	Logical attributes	Description
Work	title of the work (WT)	The title of the *work* is the word, phrase, or group of characters naming the *work*. There may be one or more titles associated with a *work*.
	form of work (WF)	The form of *work* is the class to which the *work* belongs (e.g., novel, play, poem, essay, biography, symphony, concerto, sonata, map, drawing, painting, photograph, etc.).
	date of the work (WD)	The date of the *work* is the date (normally the year) the *work* was originally created. The date may be a single date or a range of dates. In the absence of an ascertainable date of creation, the date of the *work* may be associated with the date of its first publication or release. .
	intended audience (WI)	The intended audience of the *work* is the class of user for which the work is intended, as defined by age group (e.g., children, young adults, adults, etc.), educational level (e.g., primary, secondary, etc.), or other categorization.
	context for the work (WC)	Context is the historical, social, intellectual, artistic, or other context within which the *work* was originally conceived (e.g., the 17th century restoration of the monarchy in England, the aesthetic movement of the late 19th century, etc.).
Expression	form (EF)	The form of *expression* is the means by which the *work* is realized (e.g., through alpha-numeric notation, musical notation, spoken word, musical sound, cartographic image, photographic image, sculpture, dance, mime, etc.).
	date (ED)	The date of *expression* is the date the *expression* was created (e.g., the date the particular text of a *work* was written or revised, the date a song was performed, etc.). The date may be a single date or a range of dates. In the absence of an ascertainable date of *expression*, the date of the *expression* may be associated with the date of its publication or release.
	language of expression (EL)	The language of the *expression* is the language in which the *work* is expressed. The language of the *expression* may comprise a number of languages, each pertaining to an individual component of the *expression*.
	summarization of content (ES)	A summarization of the content of an *expression* is an abstract, summary, synopsis, etc., or a list of chapter headings, songs, parts, etc. included in the *expression*.
	use restrictions on the expression(EU)	Use restrictions are restrictions on access to and use of an *expression*. Use restrictions may be based in copyright, or they may extend beyond the protections guaranteed in law to the owner of the copyright.
	technique (graphic or projected image) (ET)	Technique is the method used to create a graphic image (e.g., engraving, etc.) or to realize motion in a projected image (e.g., animation, live action, computer generation, 3D, etc.).

D.2 Coding Instruction

If you determine that a tag can be associated with a specific category of FRBR attributes, enter a number “1” in the cell. If you determine that a tag cannot be associated with any categories of FRBR attributes, leave the cell blank, and you can put your comments in the “N/A” cell, if possible. For instance, if you determine that a tag can be regarded as a “subject term”, enter an “S” in the N/A cell. Otherwise, describe it, if possible, or just put a question mark “?”.

Table 48. Coding sheet for tag attributes

Subject	**Title**	**Tags**	**Work**					**Expression**						**N/A***
			WT	WF	WD	WI	WC	EF	ED	EL	ES	EU	ET	
001	Institute for Psycho-history: http://www.psychohistory.com/	psychology												
		history												
		politics												
		psychohistory												
		science												
		culture												
		reference												
		world												
		war												
		abuse												
		theory												
		academic												
		sociology												
		parenting												

*: Not Applicable

APPENDIX E. LIST OF THE CODED WEB DOCUMENTS FOR INTERCODER RELIABILITY TEST

Table 49. List of the coded web documents for intercoder reliability test

	Subject	Title and URL
1	001 Knowledge, humanities and research	Institute for Psychohistory: http://www.psychohistory.com/
2	002 The book	Book Arts Web: http://www.philobiblon.com/
3	100 Philosophy, general resources	Philosophy: philosophy.eserver.org/
4	170 Ethics (moral philosophy)	Ethics Updates: http://ethics.sandiego.edu/
5	230 Christianity, general resources	New Testament Gateway, http://www.ntgateway.com/
6	246 Christian art	Religion and the Founding of the American Republic , http://www.loc.gov/exhibits/religion/religion.html
7	270 History of Christianity	Christian Catacombs of Rome, http://www.catacombe.roma.it/
8	300 social_sciences, general resources	Online Dictionary of the social sciences, http://bitbucket.icaap.org/
9	370 Education	Excellence Gateway, http://excellence.qia.org.uk/
10	390 Customs, etiquette, folklore	Costumer's Manifesto, Ethnic Dress, http://www.costumes.org
11	400 Language, general resources	iLoveLanguages, http://www.ilovelanguages.com/
12	414 Phonology and phonetics	International Phonetics Association, http://www.langsci.ucl.ac.uk/ipa/
13	540 Chemistry	Linux4Chemistry, http://www.redbrick.dcu.ie/~noel/linux4chemistry/
14	550 Earth sciences	GeoGuide, http://www.geo-guide.de/
15	570 Life sciences, biology	BBSRC: Biotechnology and Biological_sciences Research Council, http://www.bbsrc.ac.uk/
16	600 Technology, general resources	EurekAlert, http://www.eurekalert.org/
17	650 Management and auxiliary services	American Institute of Certified Public Accountants, http://www.aicpa.org/
18	701 Fine and decorative arts, philosophy	Aesthetics Online, http://www.aesthetics-online.org/
19	706 Fine and decorative arts, organisations	The Arts & Crafts Society, http://www.arts-crafts.com/
20	750 Painting and paintings	Art Crimes: The Writing on the Wall, http://www.graffiti.org/
21	760 Graphic arts, printmaking and prints	Powers of Persuasion, http://www.archives.gov/exhibits/powers_of_persuasion/powers_of_persuasion_home.html
22	790 Recreational and performing_arts, art	OSCAR.com - 81st Annual Academy Awards - Homepage, www.oscar.com/
23	808.8 Literature, general collections	Google Book Search: http://books.google.com/
24	809 Literary study and criticism	Literary history : http://literaryhistory.com/
25	820 English, Scottish and Irish literature	Cambridge History of English and American Literature, http://www.bartleby.com/cambridge/
26	890 Literature in other languages	Modern Haiku, http://www.modernhaiku.org/
27	910 Geography and travel	CountryWatch.com: http://www.countrywatch.com/
28	930 History of ancient world	English Heritage: http://www.english-heritage.org.uk
29	980 History of South America	Latin American Network Information Center, http://lanic.utexas.edu/

APPENDIX F. NORMALITY Q-Q PLOTS OF TAGGING CONSISTENCY

F.1 Normality Q-Q Plots in the cosine similarity measure (10 subjects)

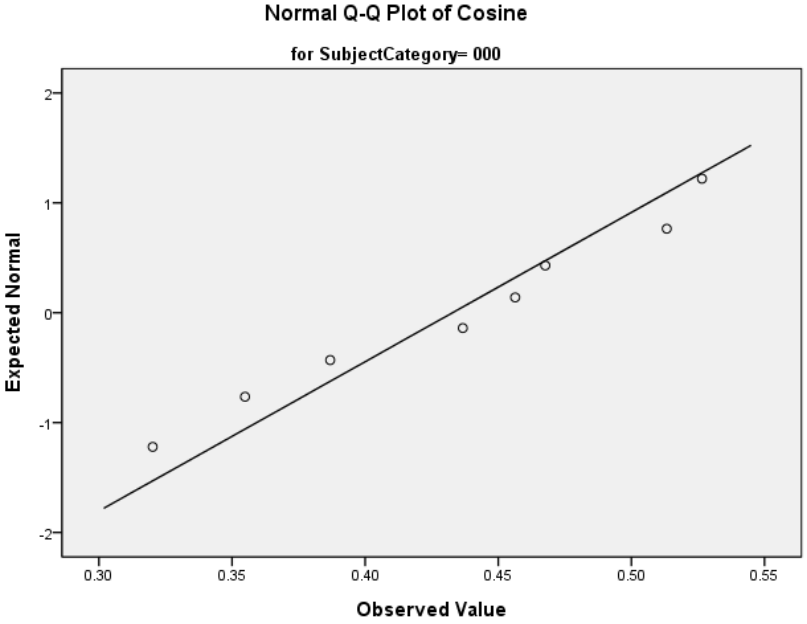


Figure 76. Normality Q-Q Plots in the cosine similarity measure (000 subject)

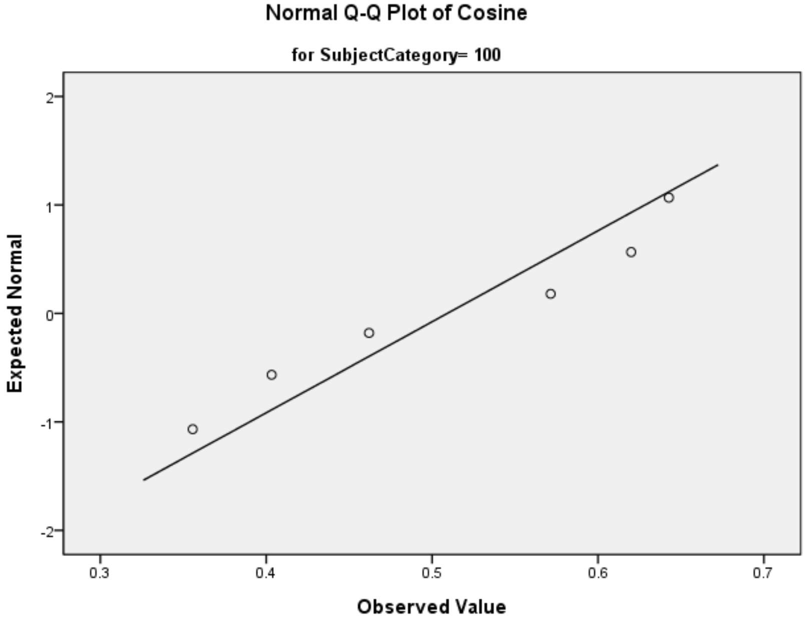


Figure 77. Normality Q-Q Plots in the cosine similarity measure (100 subject)

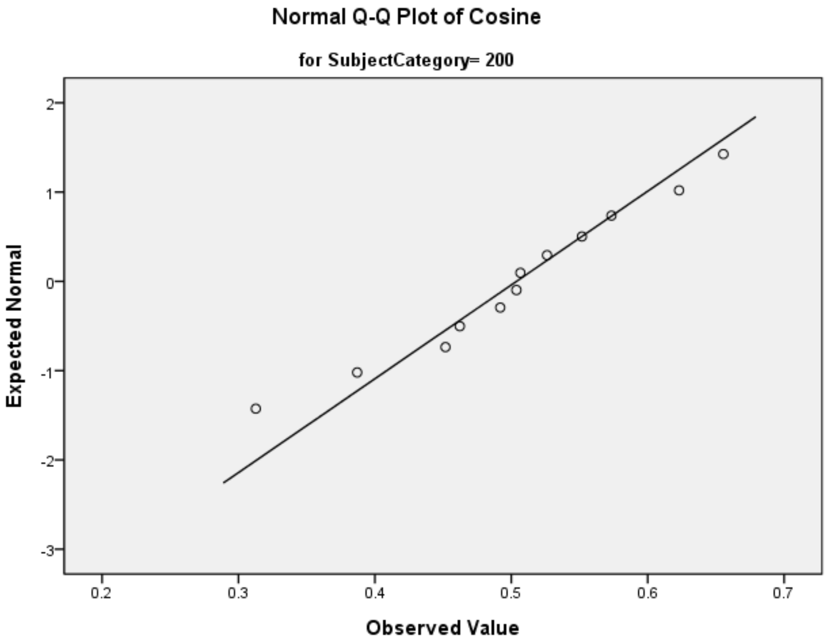


Figure 78. Normality Q-Q Plots in the cosine similarity measure (200 subject)

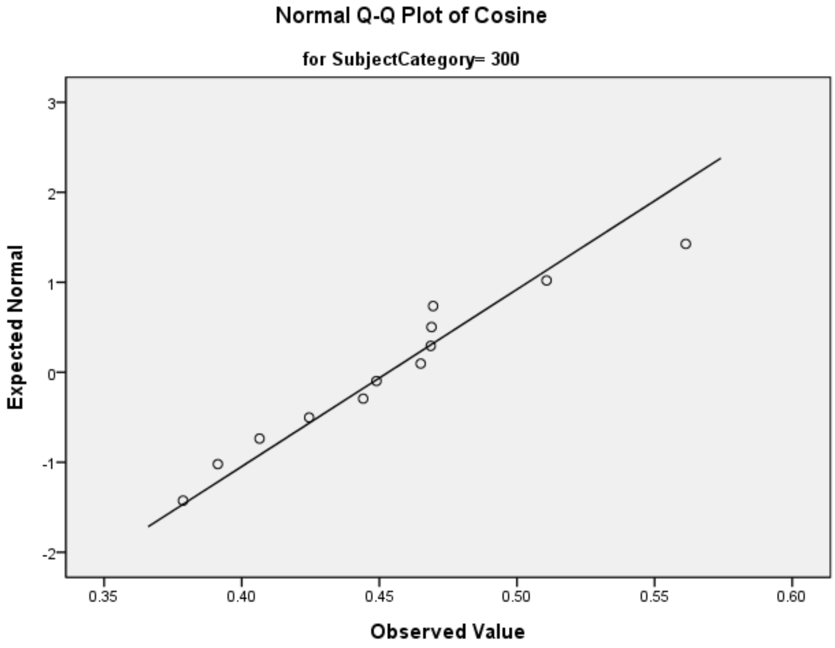


Figure 79. Normality Q-Q Plots in the cosine similarity measure (300 subject)

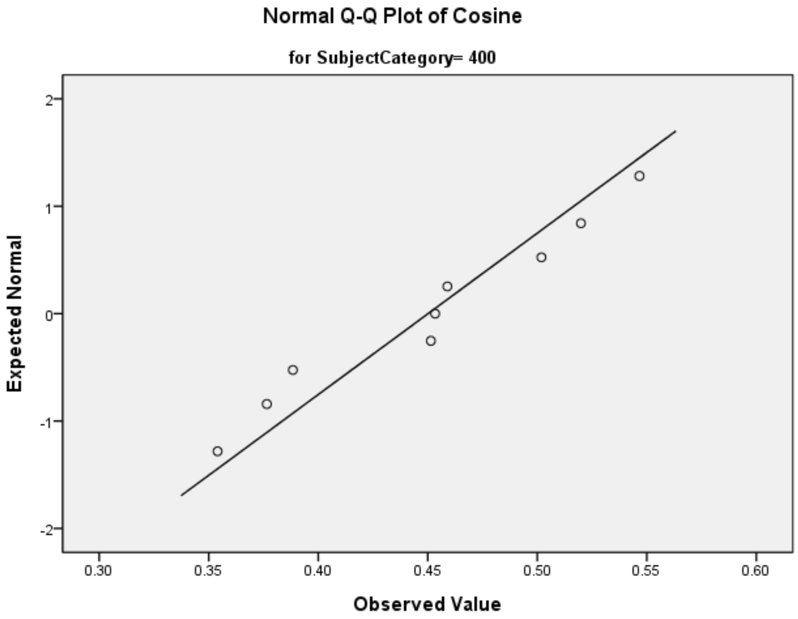


Figure 80. Normality Q-Q Plots in the cosine similarity measure (400 subject)

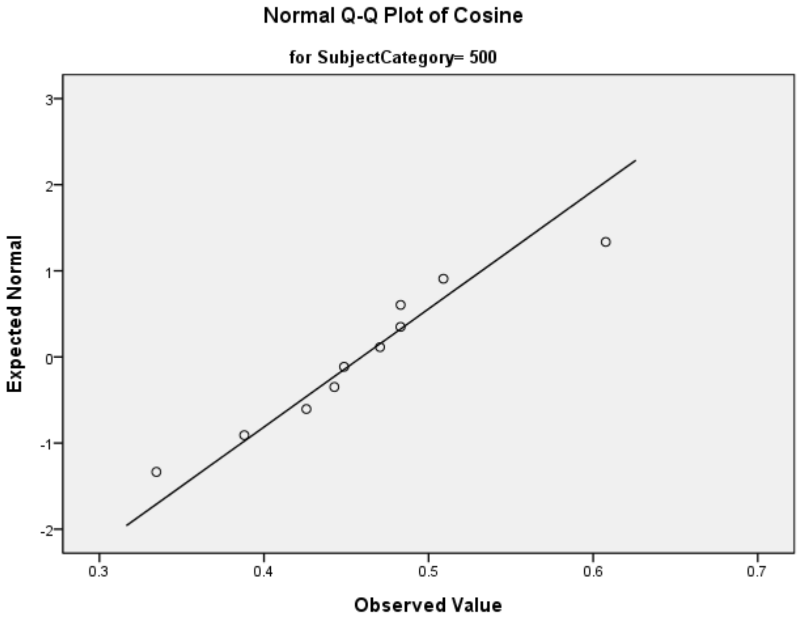


Figure 81. Normality Q-Q Plots in the cosine similarity measure (500 subject)

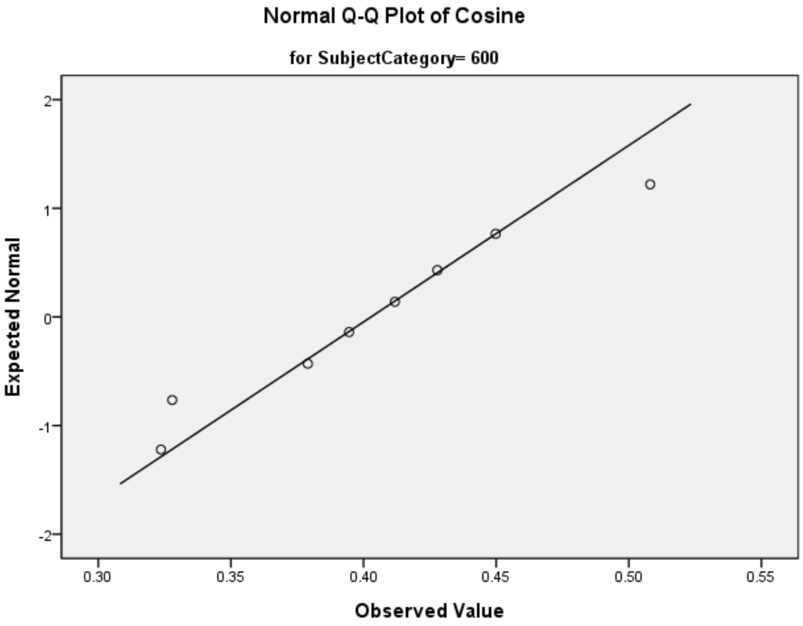


Figure 82. Normality Q-Q Plots in the cosine similarity measure (600 subject)

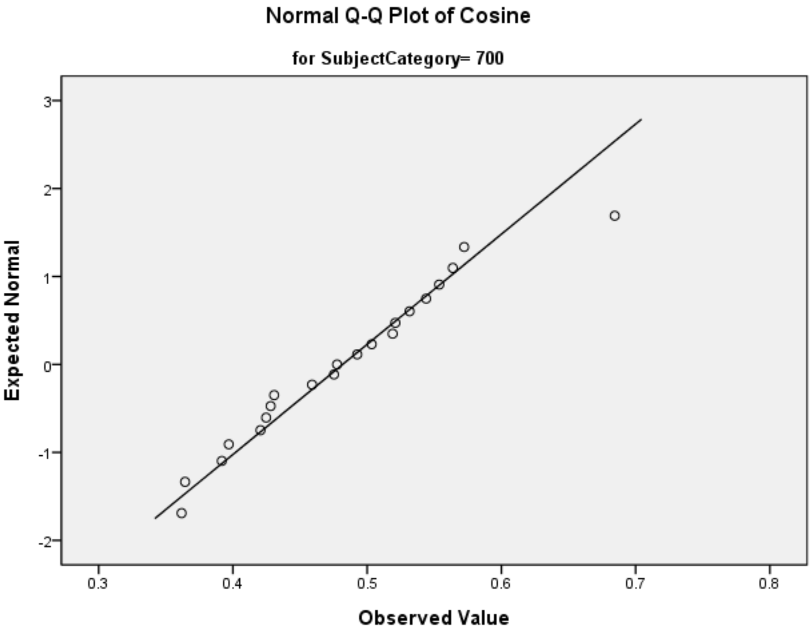


Figure 83. Normality Q-Q Plots in the cosine similarity measure (700 subject)

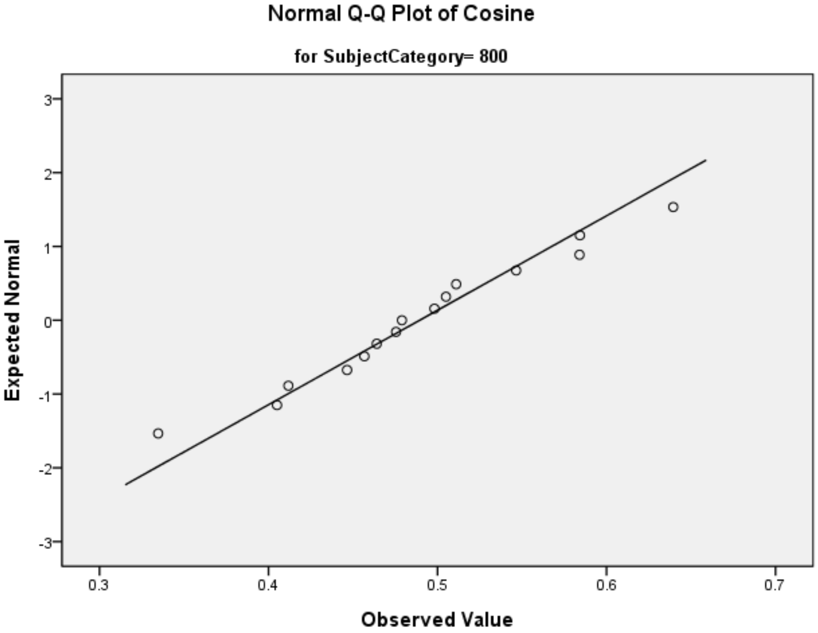


Figure 84. Normality Q-Q Plots in the cosine similarity measure (800 subject)

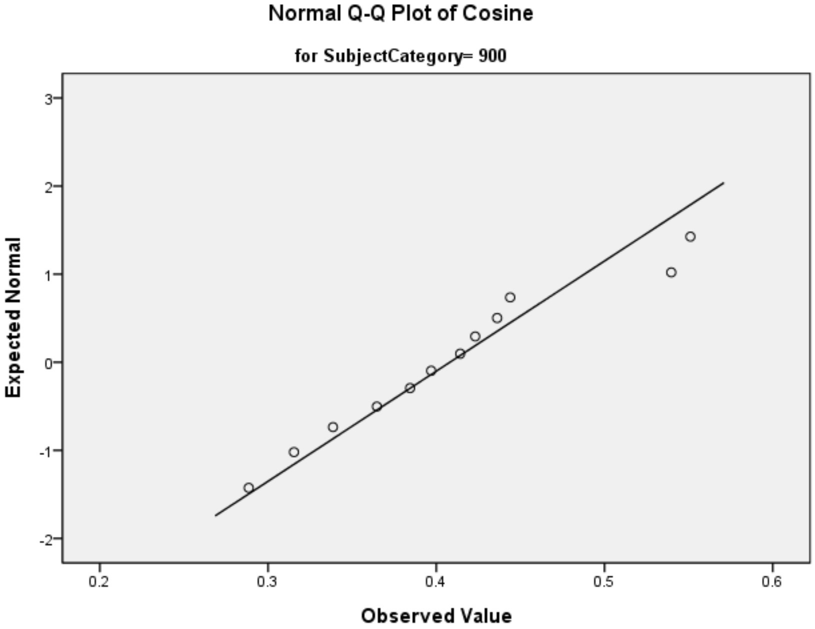


Figure 85. Normality Q-Q Plots in the cosine similarity measure (900 subject)

F.2 Normality Q-Q Plots in the Dot product similarity measure (10 subjects)

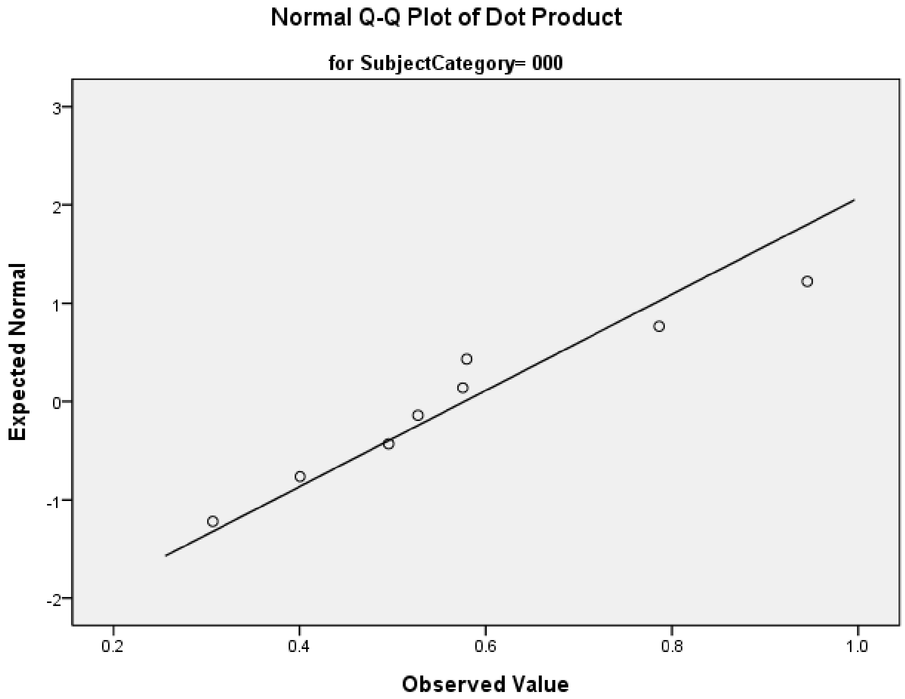


Figure 86. Normality Q-Q Plots in the dot product similarity measure (000 subject)

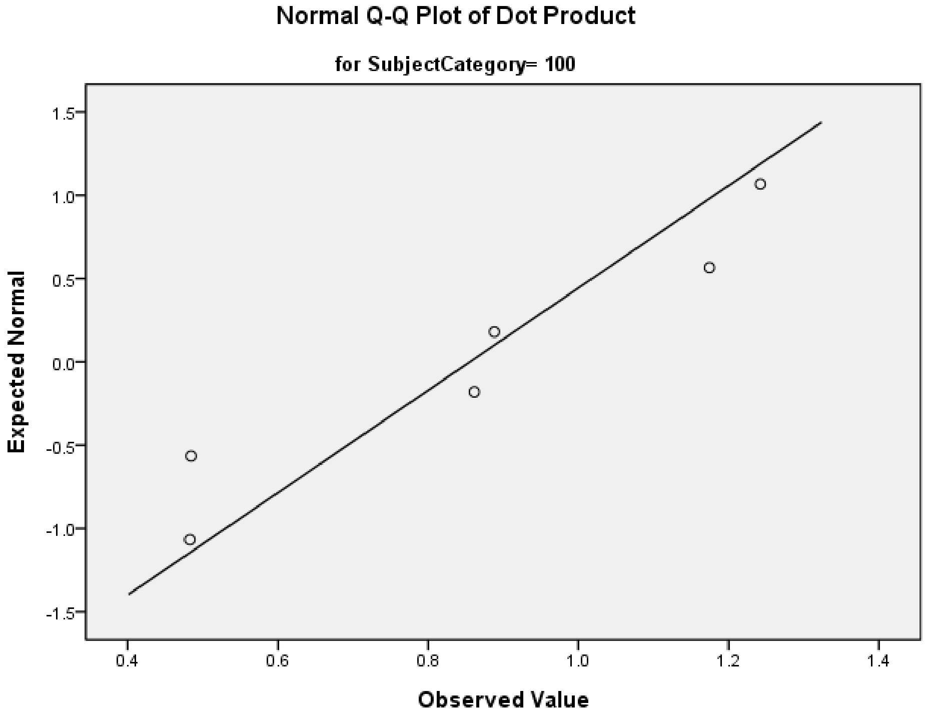


Figure 87. Normality Q-Q Plots in the dot product similarity measure (100 subject)

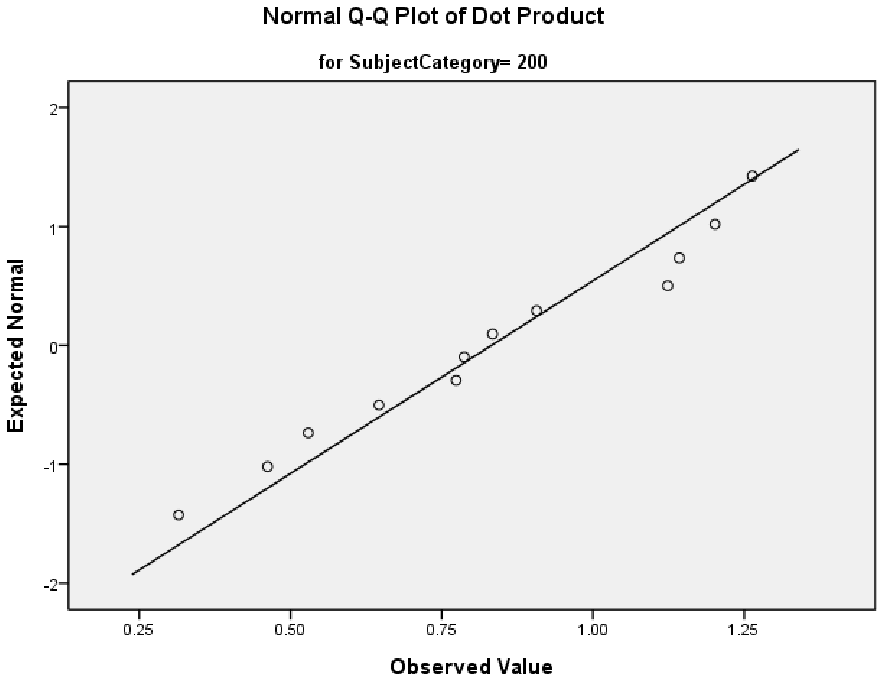


Figure 88. Normality Q-Q Plots in the dot product similarity measure (200 subject)

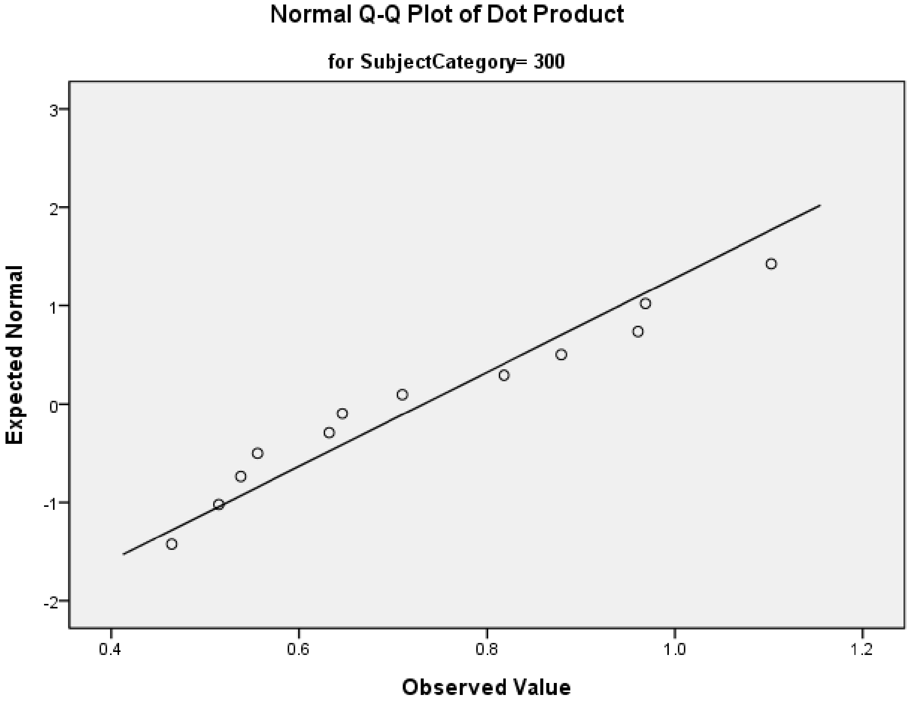


Figure 89. Normality Q-Q Plots in the dot product similarity measure (300 subject)

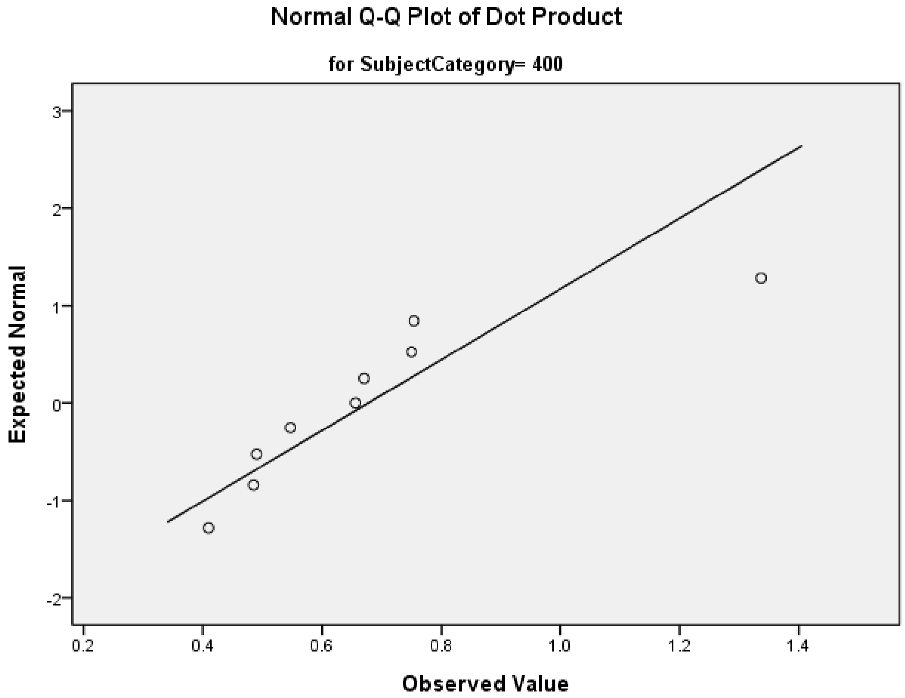


Figure 90. Normality Q-Q Plots in the dot product similarity measure (400 subject)

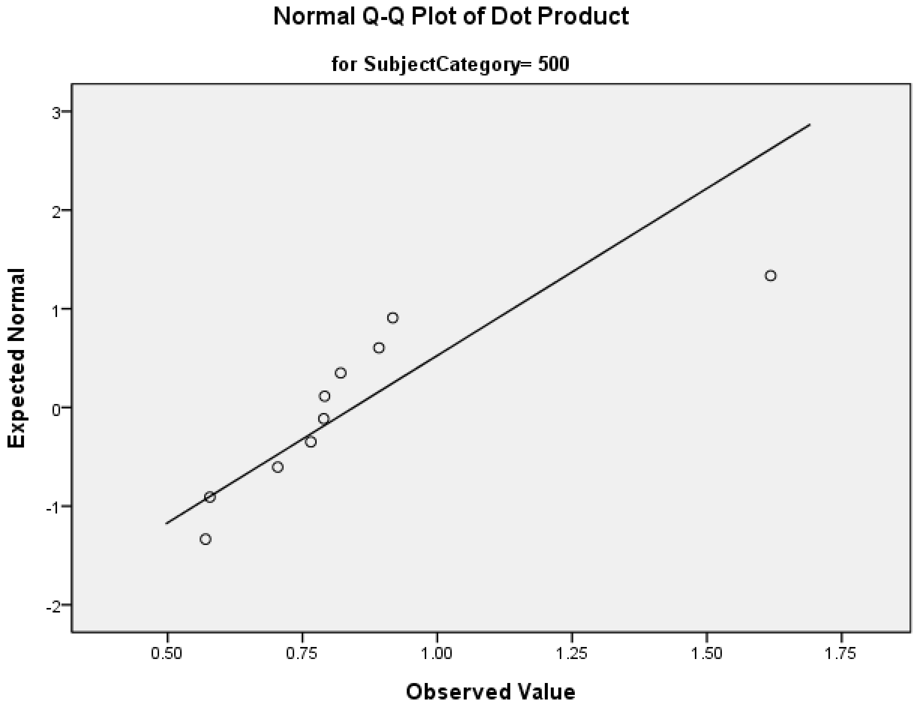


Figure 91. Normality Q-Q Plots in the dot product similarity measure (500 subject)

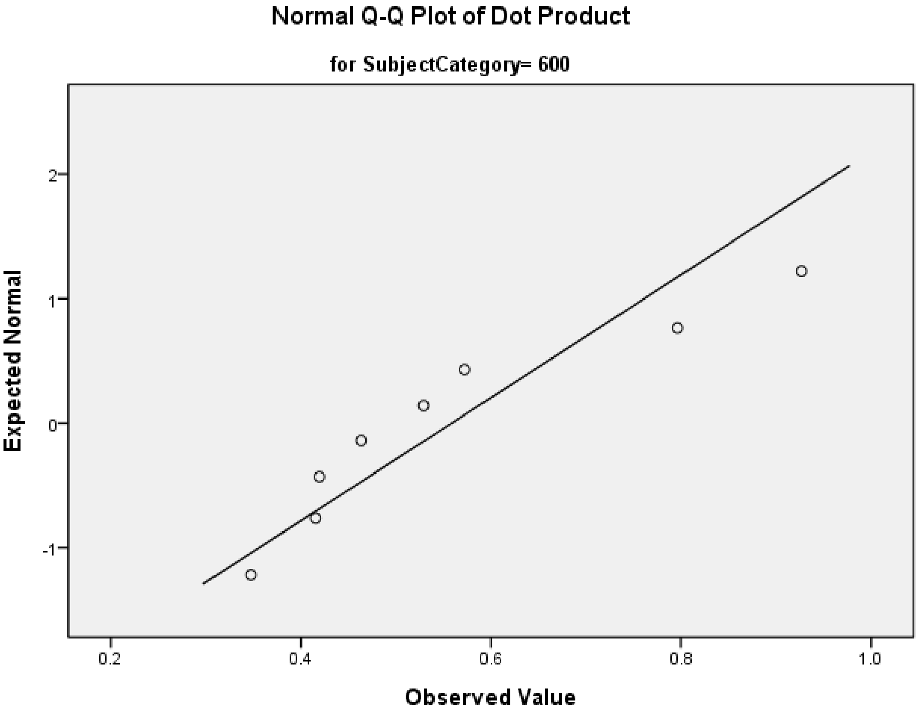


Figure 92. Normality Q-Q Plots in the dot product similarity measure (600 subject)

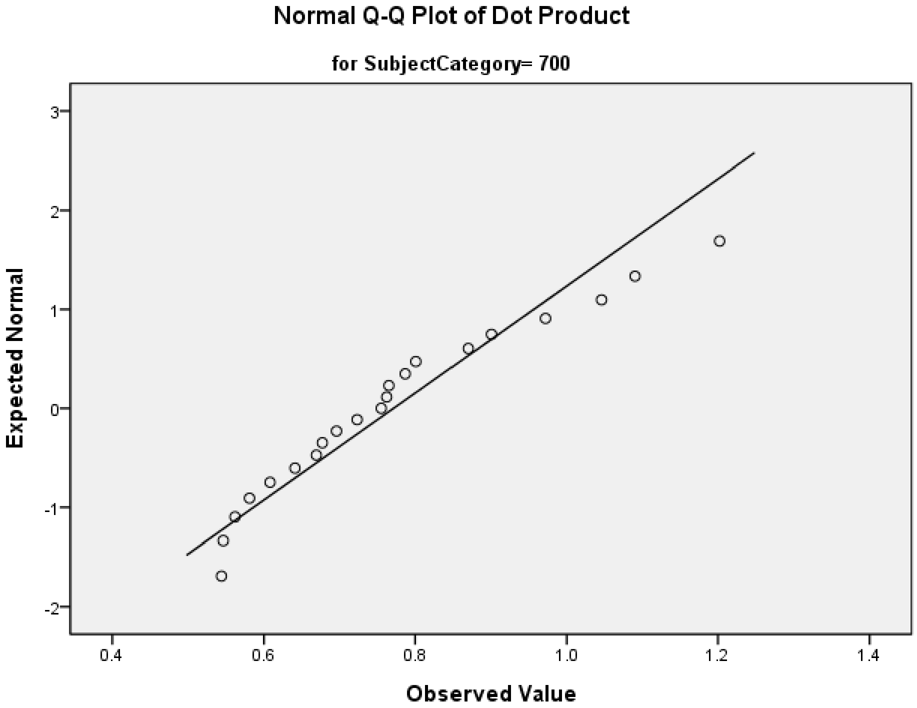


Figure 93. Normality Q-Q Plots in the dot product similarity measure (700 subject)

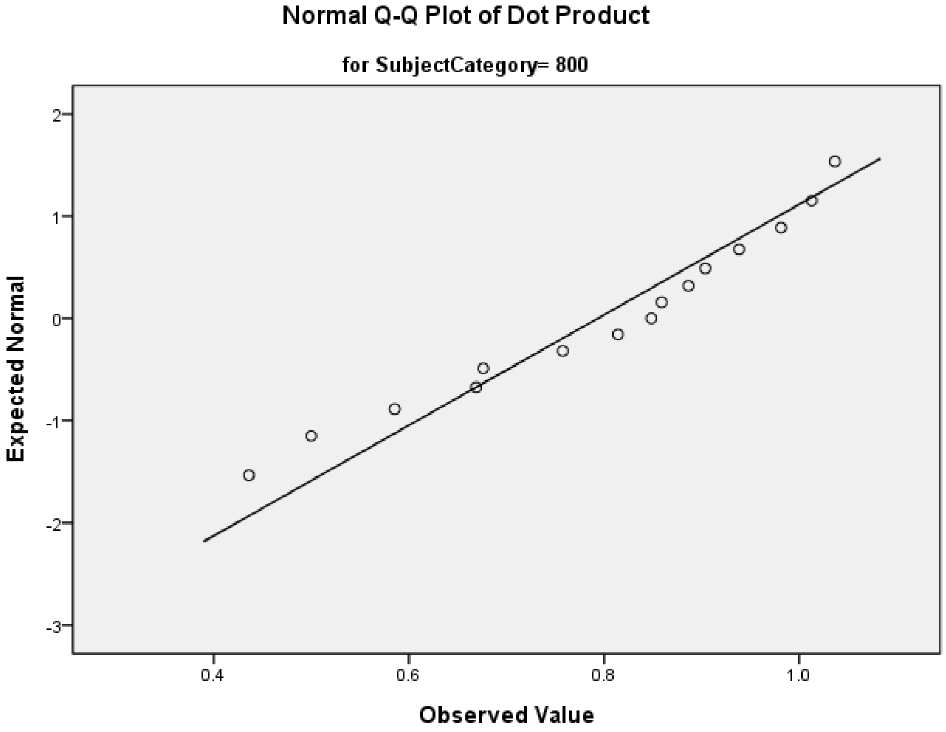


Figure 94. Normality Q-Q Plots in the dot product similarity measure (800 subject)

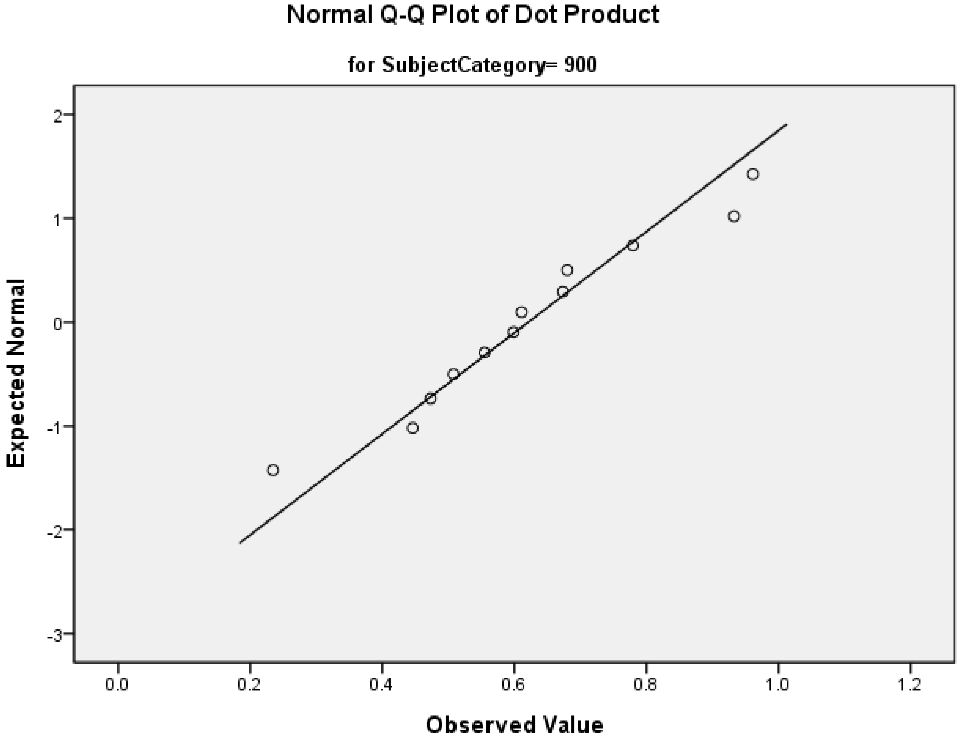


Figure 95. Normality Q-Q Plots in the dot product similarity measure (900 subject)

F.3 Normality Q-Q Plots in the Distance metrics (10 subjects)

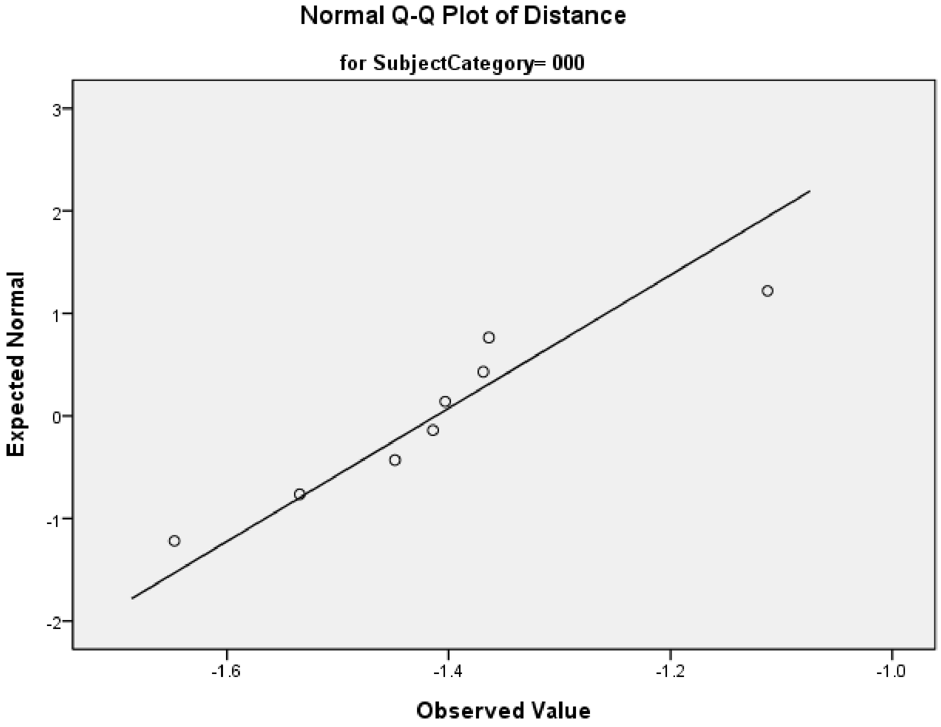


Figure 96. Normality Q-Q Plots in the distance metrics (000 subject)

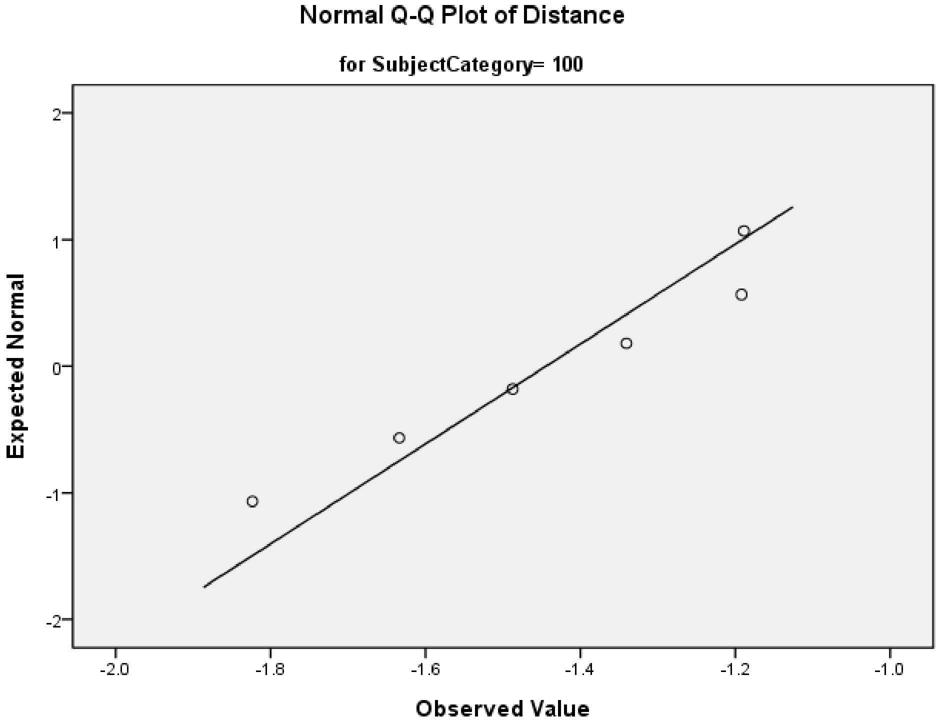


Figure 97. Normality Q-Q Plots in the distance metrics (100 subject)

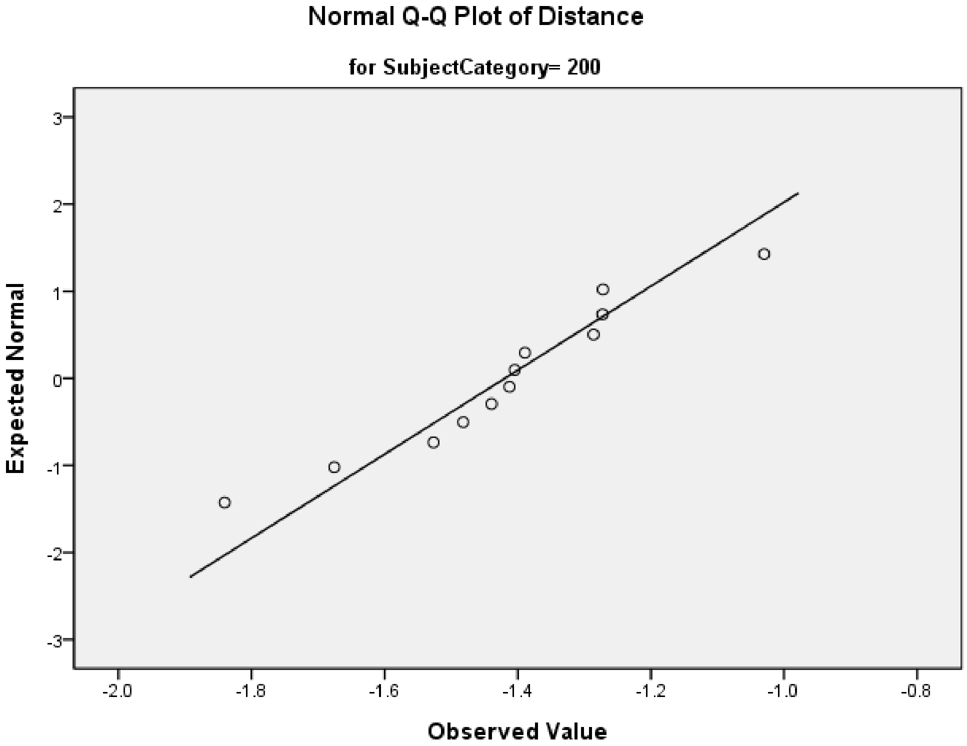


Figure 98. Normality Q-Q Plots in the distance metrics (200 subject)

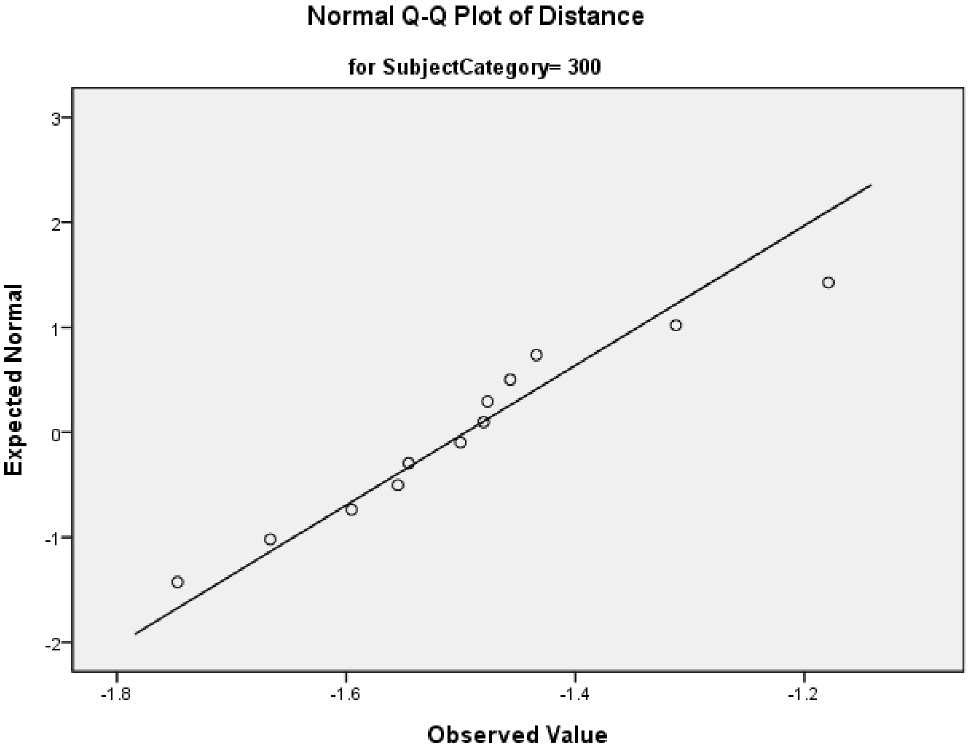


Figure 99. Normality Q-Q Plots in the distance metrics (300 subject)

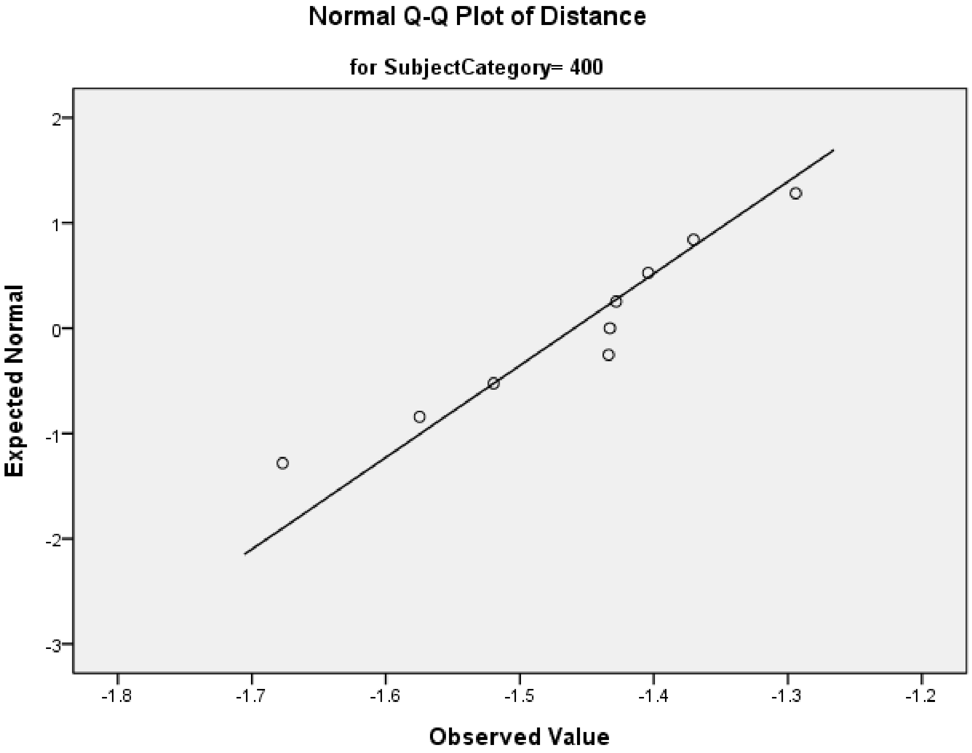


Figure 100. Normality Q-Q Plots in the distance metrics (400 subject)

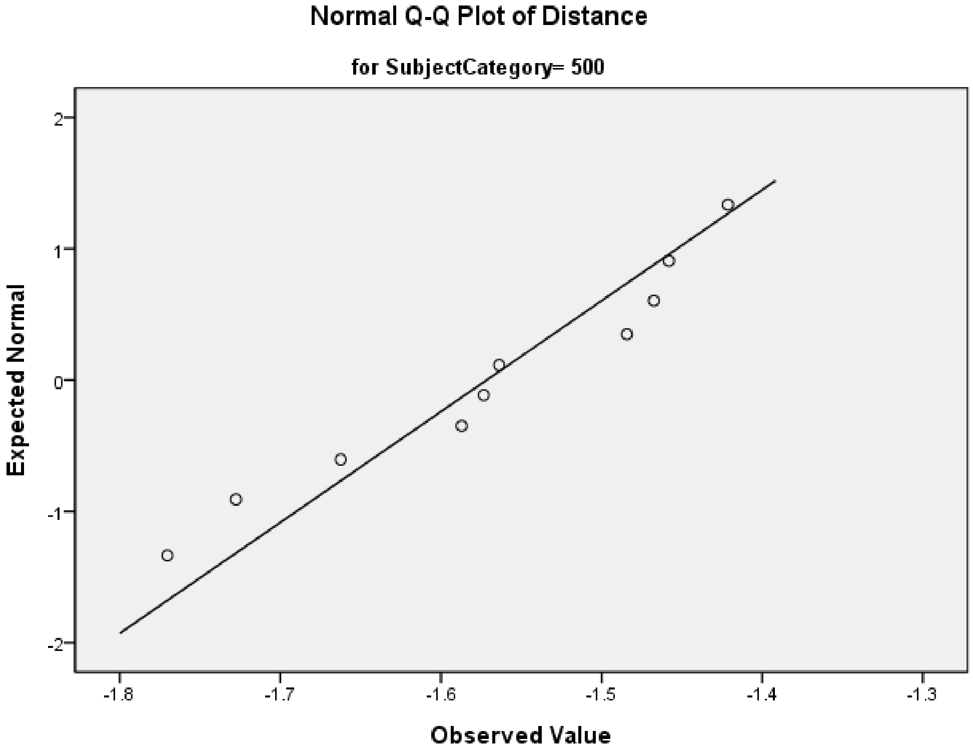


Figure 101. Normality Q-Q Plots in the distance metrics (500 subject)

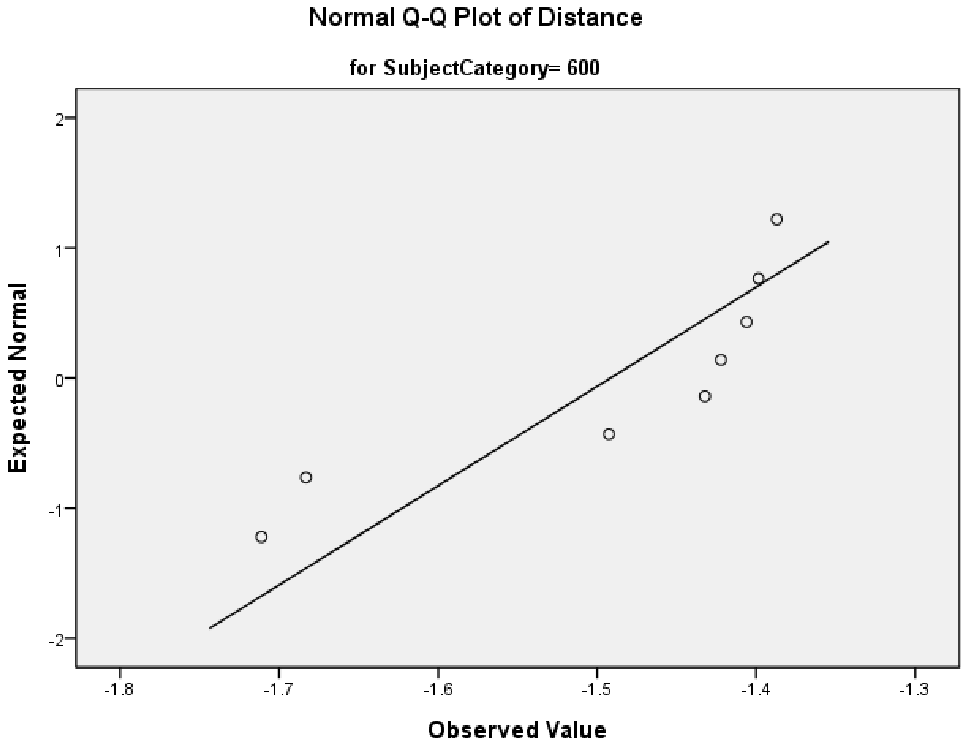


Figure 102. Normality Q-Q Plots in the distance metrics (600 subject)

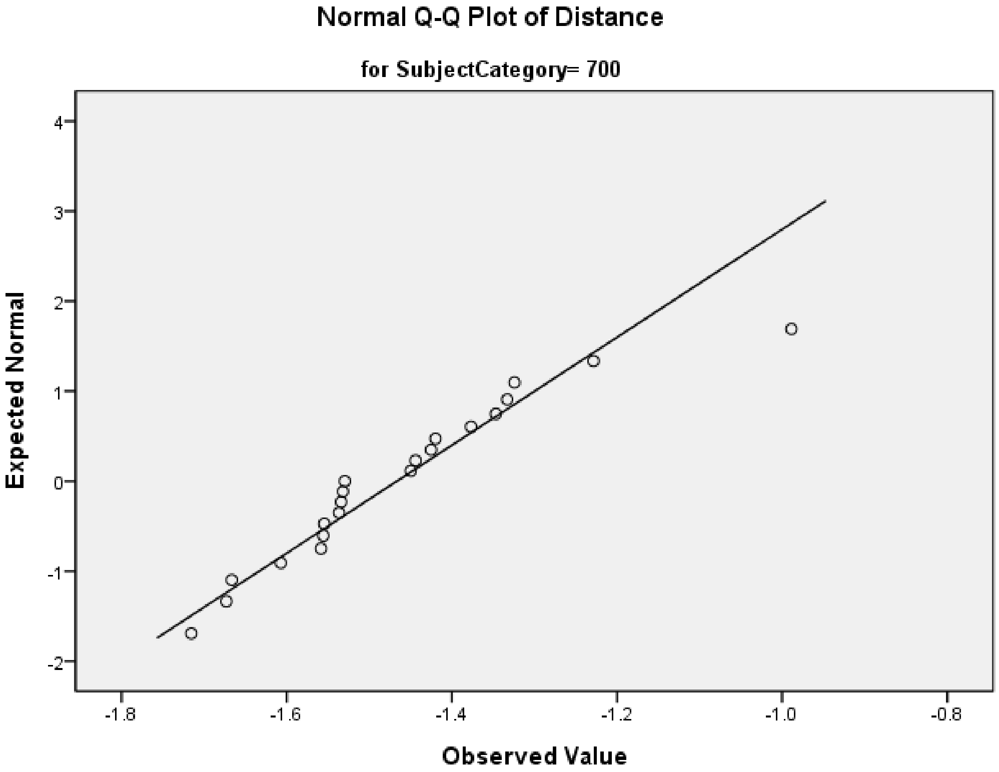


Figure 103. Normality Q-Q Plots in the distance metrics (700 subject)

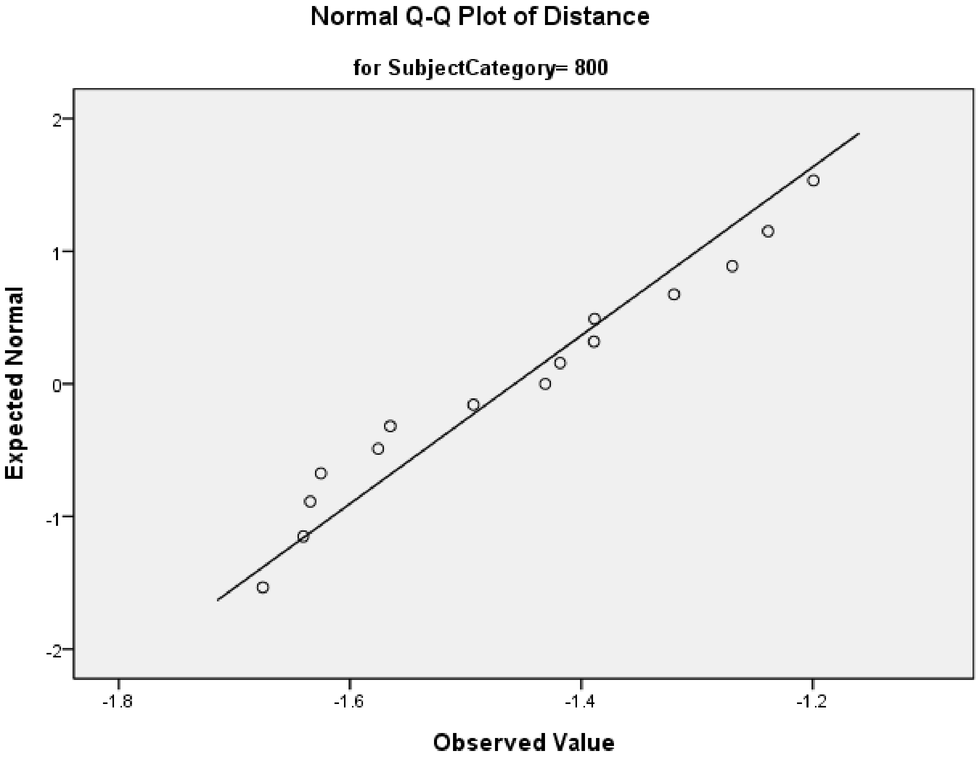


Figure 104. Normality Q-Q Plots in the distance metrics (800 subject)

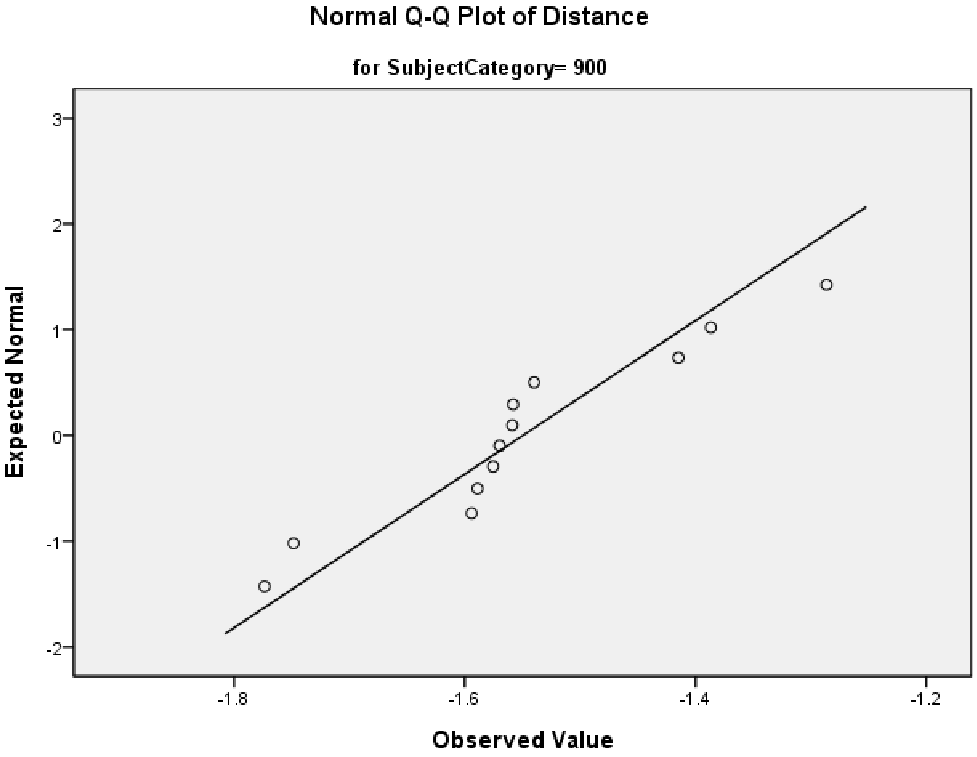


Figure 105. Normality Q-Q Plots in the distance metrics (900 subject)

APPENDIX G. A SAMPLE OF CODED WEB DOCUMENT BASED ON FRBR ATTRIBUTES

Table 50. A sample of coded web document based on FRBR attributes

Subject	Title	Tags	Work					Expression						N/A*
			WT	WF	WD	WI	WC	EF	ED	EL	ES	EU	ET	
890 Poetry, general resources	Modern Haiku, http://www.modernhaiku.org/	haiku	1											
		poetry /poems		1										
		japan					1							
		literature												S
		magazine		1										
		writing												S
		journal		1										
		words						1						
		review		1										
		world					1							
		creative writing												S
		online												F

*: Not Applicable

S: subject, U: utilization, F: feature

APPENDIX H. RESULTS OF INTERCODER RELIABILITY TEST BY SUBJECT AREAS

<<000>>

```
Run MATRIX procedure:

coder reliability

Holsti         .9697
Scott pi       .9356
Kappa          .9357
Alpha          .9366

------ END MATRIX -----
```

Table 51. Crosstabulation of coded data (000 subject)

		Coder B					Total
		0	1	2	4	5	
Coder A	0	23	0	0	1	0	24
	1	0	3	0	0	0	3
	2	0	0	2	0	0	2
	4	0	0	0	1	0	1
	5	0	0	0	0	3	3
Total		23	3	2	2	3	33

0: N/A

1: WT

2: WF

4: WI

5: WC

<<100>>

```
Run MATRIX procedure:

coder reliability

Holsti        1.0000
Scott pi      1.0000
Kappa         1.0000
Alpha         1.0000

------ END MATRIX -----
```

Table 52. Crosstabulation of coded data (100 subject)

		Coder B				Total
		0	1	2	6	
Coder A	0	19	0	0	0	19
	1	0	2	0	0	2
	2	0	0	6	0	6
	6	0	0	0	3	3
Total		19	2	6	3	30

0: N/A

1: WT

2: WF

6: EF

<<200>>

```
Run MATRIX procedure:

coder reliability

Holsti        .8780
Scott pi      .8149
Kappa         .8153
Alpha         .8172

------ END MATRIX -----
```

Table 53. Crosstabulation of coded data (200 subject)

		Coder B						Total
		0	1	2	5	6	8	
Coder A	0	21	0	0	1	1	0	23
	1	0	5	0	1	0	0	6
	2	0	0	3	0	1	0	4
	5	0	0	0	5	0	0	5
	6	0	0	1	0	1	0	2
	8	0	0	0	0	0	1	1
Total		21	5	4	7	3	1	41

0: N/A

1: WT

2: WF

5: WC

6: EF

8: EL

<300>>

```
Run MATRIX procedure:

coder reliability

Holsti         .9091
Scott pi       .7903
Kappa          .7903
Alpha          .7934
------ END MATRIX -----
```

Table 54. Crosstabulation of coded data (300 subject)

		Coder B			
		0	1	2	Total
Coder A	0	23	0	1	24
	1	1	3	0	4
	2	0	1	4	5
Total		24	4	5	33

0: N/A

1: WT

2: WF

<<400>>

```
Run MATRIX procedure:

coder reliability

Holsti         .8788
Scott pi       .8124
Kappa          .8128
Alpha          .8152
------ END MATRIX -----
```

Table 55. Crosstabulation of coded data (400 subject)

		Coder B						
		0	1	2	5	6	8	Total
Coder A	0	17	0	0	1	0	0	18
	1	0	5	0	0	0	0	5
	2	0	0	4	0	2	0	6
	6	1	0	0	0	0	0	1
	8	0	0	0	0	0	3	3
Total		18	5	4	1	2	3	33

0: N/A

1: WT

2: WF

5: WC

6: EF

8: EL

<<500>>

```
Run MATRIX procedure:

coder reliability

Holsti         .8837
Scott pi       .7664
Kappa          .7671
Alpha          .7691

------ END MATRIX -----
```

Table 56. Crosstabulation of coded data (500 subject)

		Coder B						Total
		0	1	2	4	5	8	
Coder A	0	28	0	0	1	0	0	29
	1	1	3	0	0	0	0	4
	2	0	0	6	0	0	0	6
	5	0	0	0	0	1	1	2
	6	1	0	1	0	0	0	2
Total		30	3	7	1	1	1	43

0: N/A

1: WT

2: WF

4: WI

5: WC

6: EF

8: EL

<<600>>

```
Run MATRIX procedure:

coder reliability

Holsti         .9688
Scott pi       .9272
Kappa          .9273
Alpha          .9283

------ END MATRIX ----
```

Table 57. Crosstabulation of coded data (600 subject)

		Coder B				Total
		0	1	2	6	
Coder A	0	23	0	0	1	24
	1	0	4	0	0	4
	2	0	0	4	0	4
Total		23	4	4	1	32

0: N/A

1: WT

2: WF

6: EF

<<700>>

```
Run MATRIX procedure:

coder reliability

Holsti        .8649
Scott pi      .7822
Kappa         .7832
Alpha         .7836

------ END MATRIX -----
```

Table 58. Crosstabulation of coded data (700 subject)

		Coder B					Total
		0	1	2	5	6	
Coder A	0	38	1	0	0	0	39
	1	2	6	0	0	0	8
	2	3	0	10	0	1	14
	5	1	0	0	5	0	6
	6	2	0	0	0	5	7
Total		46	7	10	5	6	74

0: N/A

1: WT

2: WF

5: WC

6: EF

<<800>>

```
Run MATRIX procedure:

coder reliability

Holsti         .7391
Scott pi       .5941
Kappa          .5952
Alpha          .5970
------ END MATRIX -----
```

Table 59. Crosstabulation of coded data (800 subject)

		Coder B					Total
		0	1	2	5	6	
Coder A	0	31	1	1	1	3	37
	1	2	7	0	0	0	9
	2	2	0	8	0	3	13
	5	0	0	0	3	0	3
	6	3	0	0	0	2	5
	8	1	1	0	0	0	2
Total		39	9	9	4	8	69

0: N/A

1: WT

2: WF

5: WC

6: EF

8: EL

<<900>>

```
Run MATRIX procedure:

coder reliability

Holsti         .9074
Scott pi       .8348
Kappa          .8349
Alpha          .8363

------ END MATRIX -----
```

Table 60. Crosstabulation of coded data (900 subject)

		Coder B					Total
		0	1	2	5	8	
Coder A	0	32	0	1	1	0	34
	1	0	3	0	1	0	4
	2	0	0	7	0	0	7
	5	1	0	0	6	0	7
	6	1	0	0	0	0	1
	8	0	0	0	0	1	1
Total		34	3	8	8	1	54

0: N/A

1: WT

2: WF

5: WC

6: EF

8: EL