PROJECT MUSE®

Visual Subject Analysis for Dublin Core Research / L'analyse
visuelle de sujet aux fins de la recherche Dublin Core

Jin Zhang, Xi Meng

Canadian Journal of Information and Library Science, Volume 37,
Number 3, September / septembre 2013, pp. 161-182 (Article)

Published by University of Toronto Press

DOI: <https://doi.org/10.1353/ils.2013.0014>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/520975>

Visual Subject Analysis L'analyse visuelle de for Dublin Core sujet aux fins de la Research recherche Dublin Core

Jin Zhang

School of Information Studies, University of Wisconsin–Milwaukee

jzhang@uwm.edu

Xi Meng

School of Information Resource Management, Renmin University of China

ximeng24@yahoo.com.cn

Abstract: The primary purpose of this article is to conduct a subject analysis on Dublin Core research, investigate subject topics related to Dublin Core research, and reveal their dynamics over time. Documents related to Dublin Core research were identified in authoritative and comprehensive databases of Web of Science, and subject terms were extracted from the relevant documents. These raw terms were regularized, and the multidimensional scaling (MDS) visualization analysis method was applied to reveal semantic relationships among subject terms. The temporal analysis on the related subject terms added a unique dimension to the study. Three periods (from 1997 to 2001; from 2002 to 2006; and from 2007 to 2011), in addition to the entire period (from 1997 to 2011), were analysed and compared in the visual contexts. Obsolete topics, newly emerging topics, and basic topics on Dublin Core research were identified and analysed in temporal subject analysis. Topic changes in the three periods are shown. The findings of this study reveal the hidden patterns of subject associations, illustrate themes of Dublin Core research and their dynamics over time, and shed light on the understanding of Dublin Core research.

Keywords: Dublin Core, metadata, visualization, subject analysis, MDS, temporal analysis

Résumé : L'objectif principal de cette étude est de procéder à des analyses de sujet dans des recherches Dublin Core, d'enquêter sur des thématiques de sujets liées à la recherche Dublin Core et de mettre en lumière leur dynamique dans le temps. Les documents relatifs à la recherche Dublin Core ont été identifiés dans des bases de données complètes et faisant autorité du Web of Science, les termes de sujets ont été extraits des documents pertinents, les termes bruts ont été régularisés, et la méthode d'analyse visuelle d'échelonnage multidimensionnel (MDS) a été appliquée afin de faire apparaître les relations sémantiques entre les termes de sujets. L'analyse temporelle sur les termes de sujets connexes a ajouté une dimension unique à l'étude. Trois périodes (1997–2001, 2002–2006, et 2007–2011), en plus de la période entière (1997–2011), ont été l'objet d'analyses comparatives dans les contextes visuels. Des sujets désuets, des sujets émergents, et des sujets de base de la recherche Dublin Core ont été identifiés et analysés dans l'analyse temporelle d'objet. Les changements

de sujet dans les trois périodes ont été présentés. Les résultats de cette étude révèlent qu'il existe des schémas cachés dans les associations de sujet; ces mêmes résultats illustrent les thèmes de la recherche Dublin Core et leur dynamique au fil du temps, et ils permettent de mieux comprendre la recherche Dublin Core.

Mots-clés : Dublin Core, métadonnées, visualisation, analyse de sujet, MDS, analyse temporelle

Introduction

As the number of digital information resources, especially Web-based resources, grows exponentially, effective and efficient access to these information resources is becoming increasingly important for end-users. Simply posting a web page on the Internet does not mean that the web page can be retrieved effectively by end-users. Without an information organization mechanism, these information resources may never be retrieved and browsed by Internet users. Metadata attached to these information resources, such as Dublin Core metadata, play a vital and essential role in attacking this thorny issue. The subject surrogates and nonsubject surrogates of a described digital resource are identified and used as access points in an information retrieval system. Users can easily access information resources through these value-added access points in the information retrieval system. Retrieved results are more relevant to users' information needs, and the quality of the original digital information resource is greatly improved in terms of information retrieval. For instance, [Zhang and Dimitroff \(2005a, 2005b\)](#) showed that metadata such as Dublin Core boost the visibility of websites in a returned results list from a search engine when the metadata are embedded into the websites properly. Dublin Core metadata are already widely used in the United States, Australia, Canada, Denmark, Finland, Sweden, Germany, Ireland, New Zealand, and the United Kingdom ([Weibel 2009](#)).

Information visualization methods can be used for object/subject clustering analysis, and they transcend traditional clustering analysis methods. The visual presentation generated by an information visualization approach illustrates object relationships in a two- or three-dimensional space. As a result, people can observe multiple perspectives of relationships among objects in an intuitive and vivid way. It shows not only connections among the observed objects but also contexts where the objects are located and how they are connected. In addition, some visualization environments even offer an interactive means which allows people to actively explore information in the visual presentation. For instance, people can rotate a visual presentation in a three-dimensional environment and pick an angle relevant to them to observe relationships among projected objects.

Dublin Core has a close relationship with information organization and knowledge organization research. Traditional subject schema (structure) and vocabulary control methods can be used to enhance metadata performance. Communication and computer science underlie the technical support for metadata implementation and operation on the Internet. Research on Dublin Core is interdisciplinary in nature. It involves not only library and information science, but also knowledge

management, computer science, information architecture, and search engine optimization. These subjects or topics are associated and form a hidden pattern of relationships in related research literature. The way they are associated and the degree to which they are semantically connected are invisible in these hidden patterns.

Research on Dublin Core is a relatively new field. Since this research has an inherent and natural relationship with information technology and other disciplines, it is not surprising that new information technology, applications, and methods would have a strong impact on Dublin Core. Therefore, research subjects or topics related to Dublin Core may change with time. It is important for researchers to see subjects and topics change over time because it would tell them which topics are obsolete and which topics are newly emerging. Identification of related research topics and revelation of related subjects or topics would definitely enable researchers to find new subjects and topics related to Dublin Core.

The research questions of this study are:

1. What are the subjects or topics related to Dublin Core?
2. What are the semantic connections among related subjects or topics?
3. What are the temporal subject changes in terms of research on Dublin Core?

This study uses a visualization method to demonstrate the semantic connections among subjects related to research on Dublin Core. The primary aim of this study is to reveal the hidden patterns of subject associations, illustrate semantic subject structures and their dynamics over time, and shed light on the understanding of semantic relationships in research on Dublin Core.

Related research

Information regardless of its physical form shares three fundamental features or attributes: subject characteristics, non-subject characteristics, and relationships with other information. From the information retrieval perspective, effective description of these three attributes would enhance information organization and therefore information discovery.

Metadata have been becoming more critical for information organization in the network environment, and interoperability is one of the challenges for information exchange. Metadata schemas' mapping methods to enhance interoperability are necessary to achieve universal availability of resources, to improve simple resource description records, and to enable searching across syntaxes and databases (Chandrakar 2005). Comparison studies between Dublin Core and other metadata schemas such as the RDF Schema, XML, Document-Type Definition (DTD), Document Content Description (DCD), and Schema for Object-Oriented XML, MPEG-7, EAD, MODS, VRA, and TEI have been hot research topics (Howarth 2003; Hunter and Armstrong 1999; Greenberg 2001).

Dublin Core is one of the most widely used metadata standards. It provides a standard method of describing a wide range of different types of digital information resources, allowing these diverse types to be retrieved through a single searching process (Howarth 2003). Dublin Core metadata set was initially developed during 1995 and 1996 as a generic metadata standard for libraries,

archives, governments, and other publishers of online information to enhance online information access (Weibel et al. 1995, 1998). Conferences on Dublin Core were held on a regular basis to address issues related to Dublin Core standard revision, implementation, challenge, application, and so forth. Dublin Core was approved as the NISO standard (Z39.50) in 2001 (National Information Standards Organization 2007). The standard was updated as an ISO standard (15836) entitled *Information and Documentation-Dublin Core Metadata Element Set* recently. A survey (Ma 2007) found that Machine Readable Cataloging (MARC) format was the most widely used metadata schema (91%), followed by Encoded Archival Description (EAD) (84%), then Unqualified Dublin Core (78%), and Qualified Dublin Core (67%).

Several studies have been done on the impact of Dublin Core implementation on web pages and their visibility in search engine ranking lists. Sokvitne (2000) focused much of his analysis on the evaluation of the effectiveness of title, creator/publisher, and subject in Dublin Core metadata for retrieval. In this study, metadata from Australian government and educational organizations were selected and analysed. The extent to which the three metadata elements facilitated information access and retrieval was evaluated. The findings showed that the title element enhanced information access, the access capacity of the creator/publisher element was weakened because of inconsistent name formats, and broad terms in the subject element resulted in improved precision. Zhang and Dimitroff (2004) examined the performance of major search engines with regard to two groups of web pages: those with metadata and those without metadata. In their study, seven major search engines, including AllWeb, EntireWeb, Lycos, AltaVista, Infospace/Fast, Google, and Yahoo, were investigated. Metadata were embedded in one group of websites, and selected terms were assigned to the subject element. Metadata were not assigned to the other group of the same websites, and this group served as a control group in the experimental study. Website visibility on a search engine retrieval list was used to evaluate the effectiveness of metadata. The findings demonstrated that the websites with metadata outperformed the websites without metadata among the investigated search engines. A similar study (Mohamed 2006) was conducted to investigate the impact of adding metadata elements to web pages on search engine results ranking. Three search engines (AltaVista, Infoseek, and HotBot) were used in an experimental study, and he concluded that adding Dublin Core metadata elements to web pages raised their rank orders on the search engine results lists.

Darmoni et al. (2001) pointed out that building an interdisciplinary, international consensus around a core element set is the central feature of Dublin Core Metadata Initiative (DCMI). To make Dublin Core applicable in more fields and describe complex information objects, DCMI established several work groups to expand Dublin Core metadata based on demands of different areas, including DC-Agent, DC-Citation, DC-Library, DC-Education, and DC-Government. Sutton (1999) introduced the conceptual foundations for the Gateway to Education Material (GEM) framework, which adapted Dublin Core Element Set to meet the needs of the education domain. Allinson (2008) discussed

the Scholarly Works Application Profile (SWAP), which uses a Dublin Core application profile to describe scholarly texts that support the growing corpus of scholarly resource types. Dublin Core was also used for describing Internet medical information so as to promote search access to Internet medical documents (Patridge and Namulanda 2008; Eysenbach et al. 2001; Malet et al. 1999). A Dublin Core-based metadata schema was also established for a unified e-Government information resource description framework (RDF) and the problems associated with metadata application across government agencies (Amara-vadi 2005; Park et al. 2009; Devey and Côté 2006).

Bibliometric analysis methods and citation analysis methods are usually used to discover disciplines or fields related to a specific discipline, reveal significant contributors, and unveil emerging interdisciplinary areas. Many studies used bibliometric analysis methods to ascertain interdisciplinary structures and revealed hidden patterns in related fields such as education and biochemistry at journal level or article level (Leydesdorff 2006; Porter et al. 2007; Adams, Jackson, and Marshall 2007; Van Raan and Van Leeuwen 2002). An example of the bibliometric analysis method can be found in the investigation by Archibald and Line (1991) of nine subject areas defined by the Dewey Decimal Classification through the journals and their articles. Similarly, Chua and Yang (2008) investigated the subjects of journals by focusing on the keywords of articles published in the *Journal of the American Society for Information Science and Technology* in two decades (1988–1997 and 1998–2007) and discovered a topic shift from general information science toward specific sub-disciplines. Milojević (2009) also used the bibliometric method to identify social and cognitive structures of nanoscience and their dynamics over time.

Effective access to Internet information resources has become a very urgent necessity for researchers in LIS and in other fields. Dublin Core, a powerful metadata standard used to describe Internet resources, has already manifested its interdisciplinary and dynamic nature. Although there have been a lot of studies on Dublin Core implementation, application, expansion, use, and so forth, most of them focus on a specific issue of Dublin Core, and few of them focus on revealing related topics and relationships among the research studies on Dublin Core.

Research method description

Temporal analysis: Partition of the subject analysis periods

The first workshop on Dublin Core was held in 1995 by Online Computer Library Center (OCLC), a library consortium (Weibel, Iannella, and Cathro 1997). The earliest research papers on Dublin Core in the investigated databases appeared in 1997 (Desai 1997), which is consistent with the time of Dublin Core creation because it is very normal that research papers related to a new concept or theory be published two years after the new concept or theory is introduced. Therefore the starting point of this article's temporal analysis was 1997 and ended in 2011. This study divided the entire investigated period into three sub-periods of five years each.

- Period I: 1997–2001
- Period II: 2002–2006
- Period III: 2007–2011

Investigators retrieved, collected, and analysed documents related to Dublin Core separately for each period, and as a result, data from the three periods can be compared and differences can be identified.

Data collection

Web of Science was selected for this study because research papers on Dublin Core are well covered in Web of Science. “Dublin Core” was the query phrase used to retrieve all related documents from all Web of Science databases. There was no limitation placed on search fields because search results from the databases needed to be as inclusive as possible for related subject analysis. After the query was submitted to the system and relevant documents were retrieved, results were recorded separately for the three different periods.

Subject term extraction

Investigators extracted subject terms from the abstract field, title field, author-assigned keyword field, and subject field in a retrieved document. It was from these subject terms that related subjects or topics were revealed and identified in visual term analysis. If the databases did not provide detailed bibliographic information in a retrieved document record, the investigators checked the document’s full text for more information.

All the extracted terms formed a term master file. Synonyms, plurals, and abbreviations of raw terms were regularized, and the regular term became the standard entry term in the term master file. Consequently, terms related to the same concept were consolidated in one regular term in this process. The regularization process enhanced the quality of term clustering analysis because the terms related to the same concept were treated as one term. Next, the term frequency of each regular term was tallied. As a result, the term master file had three columns for each regularized term: first for a regularized term, second for its frequency, and third for the document ID, which was used to trace from where the term was extracted.

To achieve a sound and plausible clustering analysis result, a frequency cut-off point was set to eliminate less-related terms. Low-frequency terms make no contributions to the term clustering analysis because they are not associated with other terms and term clustering analysis is based on term associations. It is clear that the higher the term frequency, the more relevant the term is to Dublin Core, and vice versa. The term frequency cut-off point for this research was set to 2. Terms whose frequencies were less than 2 were removed from the term master file.

Term-term proximity matrix definition

The completed term master file can be converted to a term-document matrix (Equation (1)). The rows are the subject terms from the term master file, while the columns are the documents which contain these subject terms.

$$Matrix_1 = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \quad (1)$$

In Equation (1), m is the number of the subject terms in the term master file, n is the number of the retrieved documents that include the term, and a_{ij} represents the frequency of term i in document j . The matrix is an $m \times n$ matrix. In fact, Equation (1) corresponds to a high dimensional vector space defined by the retrieved document set and related terms. Each period in this study corresponds to one subject clustering analysis and its matrix. This study also includes a matrix for the entire period of the study. As a result, there are four term-document matrixes.

Each term-document matrix was converted to a new term-term matrix. The term-term matrix was used as input data for the visual subject term analysis, and a similarity measure was selected to measure similarity between two terms in Equation (1). There are many similarity measures available such as the cosine measure and the distance measure. Each similarity measure has its strengths and weaknesses, and its performance varies in different data sets (Korfhage 1997). A pilot study showed that the distance-based similarity measure was the most suitable for analysis of regularized terms in Dublin Core metadata sets. Therefore, it was selected for study. The distance-based similarity measure is defined as follows:

$$S(T_i, T_j) = \frac{1}{c^{D(T_i, T_j)}} \quad (2)$$

$$D(T_i, T_j) = \left(\sum_{k=1}^n (a_{ik} - a_{jk})^2 \right)^{1/2} \quad (3)$$

where T_i and T_j are two terms, c is a positive constant whose value is equal to 1.4 in this study, and a_{ik} and a_{jk} are cells of the matrix defined in Equation (1). In fact, T_i and T_j correspond to row i and row j , respectively, in the matrix defined in Equation (1). Equation (2) defines similarity between the two terms T_i and T_j . Equation (3) defines the Euclidean distance between the two terms T_i and T_j . Both Equations (2) and (3) suggest that the similarity value between two terms ranges from 0 to 1, 0 for the weakest similarity and 1 for the strongest similarity.

The generated term-term matrix (Equation (4)), which provides the proximity between two terms, is defined as follows. The proximity or similarity between two terms is determined by the strength between the two terms in the highly dimensional space defined by Equations (1), (2), and (3). The size of the matrix is determined by the number of extracted terms.

$$Matrix_2 = \begin{pmatrix} b_{11} & \dots & b_{1m} \\ b_{21} & \dots & b_{2m} \\ \dots & \dots & \dots \\ m_{m1} & \dots & b_{mm} \end{pmatrix} \quad (4)$$

Since the similarity between one term and itself is always equal to 1, then:

$$b_{kk} = 1 \quad 1 \leq k \leq m \quad (5)$$

Other cells' definitions are shown in Equation (6).

$$b_{ij} = S(T_i, T_j), \quad i \neq j, \quad 1 \leq i, j \leq m \quad (6)$$

Because the similarity between term A and term B is equal to the similarity between term B and term A, the matrix is an $m \times m$ symmetric matrix. In other words the following equation always holds.

$$b_{ij} = b_{ji} \quad 1 \leq i, j \leq m \quad (7)$$

Four term-term matrixes for the three different periods and one for the entire period were produced. Each matrix served as input data for a visual clustering analysis.

Multidimensional scaling visualization analysis method

Multidimensional scaling (MDS) visualization analysis method is a mature and widely used collection of statistical techniques allowing three-dimensional visualization of information. The MDS method has many advantages: (1) Data used in multidimensional scaling analysis are relatively free of any distributional assumptions such as normal distribution. (2) The method can handle various types of data including ordinal, interval, and ratio-level data, unlike some visualization methods, which are only applied to interval data and ratio data. (3) This method is well established and is widely used in many application domains such as in medical, health, business, and other fields. (4) Many commercial statistical software packages (e.g., SPSS) and non-commercial software packages incorporate MDS analysis (Zhang et al. 2008). The ALSCAL procedure found in the SPSS (version 19) software package incorporates MDS analysis and was used for data analysis in this study.

The quality of results in a MDS analysis is usually measured by two indicators: one is the stress value, and the other is the squared correlation index (RSQ)

(R^2). Many factors affect these indicators. For instance, low dimensionality of a data set and loose associations among projected objects can lead to a high stress value and/or low R^2 value. Generally speaking, the lower the stress value (or the higher the R^2 value), the better the MDS result is. In other words, a low stress value (or a high R^2 value) corresponds to a plausible and reliable MDS analysis result. An analysis result with a stress value less than 0.10 and R^2 value more than 0.90 is usually considered reliable and plausible. SPSS provides both of these indicators for decision making.

Results and discussion

The Web of Science database returned 146 articles from the query “Dublin Core.” Thirty-three were removed from the data set as irrelevant. They addressed issues like the transportation policy of the city of Dublin, research on Dublin’s rental market, and references to Trinity College Dublin. Term extraction was applied to the remaining 113 relevant articles. This extraction generated 625 raw (not regularized) terms, which averaged about 6 terms per article.

These 625 raw terms extracted from the relevant retrieved documents were merged by regularizing into a list of 203 terms. For instance, the two terms *Web-search engine* and *search engine* were merged into *search engine*; *quality control process* and *quality control* were regularized as *quality control*; and *Internet resource* and *Web resources* were combined into *Web resources*. This merged list of 203 terms, known as the term master file, ranked the terms in descending order of their frequencies of appearance in the retrieved document sets. Based on the predefined frequency cut-off point value ≥ 2 , 123 terms were excluded from the final term master file list of 80 subject terms. Of these 80 subject terms, 58 terms were found in period I, 67 terms in period II, and 42 terms in period III. Consequently, an 80×80 term-term matrix was produced for the entire period, a 58×58 term-term matrix for period I, a 67×67 term-term matrix for period II, and a 42×42 term-term matrix for period III in this study. Final subject terms and their frequencies in the three periods are listed in [table 1](#).

Analysis of the subject terms and their frequencies was applied to the term master file. This was followed by term clustering analysis to enable people to observe and understand the semantic relationships among subject terms. Term clustering analysis was performed separately for each period and for the entire period from 1997 to 2011. This last analysis was included to demonstrate an overall picture of term associations for Dublin Core study.

[Table 2](#) shows that the number of the relevant documents and the extracted subject terms for period I were 43 and 58, respectively; the number of the relevant documents and the extracted subject terms for period II were 47 and 67, respectively; and the number of the relevant documents and the extracted subject terms for period III were 23 and 42, respectively. The numbers reached their peaks in period II and then plummeted in period III ([figure 1](#)). These numbers suggest that period I is an initial phase for Dublin Core research, period II is its booming phase, and period III is its stable and mature phase.

Table 1: Final subject terms and their frequencies in three periods

Terms	Period I	Period II	Period III	Terms	Period I	Period II	Period III
Dublin Core	21	26	11	EAD	2	1	1
metadata	13	19	9	multimedia	3	1	
XML	9	10		W3C	2	2	
digital library	4	8	5	quality control		1	3
RDF	9	6	1	classification	1	3	
indexing	9	5		bibliographic	1	2	
				control			
DCMES	7	2	4	data handling		2	1
semantics	3	8	2	digital repository			6
digital content	5	8		metadata format	1		2
cataloguing	4	4	3	metadata classification		3	
				schema			
standard	3	6	2	open access		1	2
interoperability		6	3	SGML	3		
MARC	3	4	2	VRA Core	2	1	
Semantic Web		8	1	Web service	1	2	
search engine	4	5		Z39.50	3		
controlled	4	2	2	abstracting	2		
vocabulary							
DCMI	4	4		AACR	3	1	1
information retrieval	3	4	1	abstract model		1	1
metadata schema	3	4	1	bibliographic record	1	1	
World Wide Web	10	4	1	bibliographic citation	1	1	
Web resource	8	4	1	collection-level		2	
				metadata			
application profile		4	3	DSpace			2
database	1	3	3	digital divide		2	
management							
information	3	2	2	hypertext	2		
management							
metadata standard	3	4		institutional repository			2
meta tag	3	2	2	keyword		2	
MPEG-7	3	4		museum	1	1	
resource discovery	6			metadata creation	1	1	
e-Government		3	2	MODS		1	1
electronic resource	1	3	1	Open Archives		2	
				Initiative			
HTML	4	1		OAI-PMH		1	1
information system	1	3	1	photograph	1		1
metadata element	1	3	1	resource description	1	1	
set							
MeSH	5			subject gateway	1	1	
ontology		4	1	schema	2		
Website	1	4		subject heading	1	1	
authoring	2	2	1	video	1	1	
audiovisual library	2	2		Web mining	1	1	
digitization		4		health information		1	1
domain ontology		3	1	image	1		1

Table 2: Summary of retrieved documents and extracted terms

	Period I	Period II	Period III
Number of relevant document retrieved	43	47	23
Number of extracted and regulated subject terms	58	67	42

Figure 1: Changes in the number of related documents and number of terms in the three periods

Subject term analysis

Not only did the frequency of documents and terms change between periods, but the subject terms also changed. Appearing only in period I, the following subject terms were phased out of the subsequent periods: *abstracting*, *hypertext*, *resource discover*, *MeSH*, *schema*, *SGML*, and *Z39.50*. Related topics such as *hypertext*, *SGML*, and *Z39.50* introduced in the last century are now relatively obsolete. The terms *digitization*, *metadata classification schema*, *collection-level metadata*, *keyword*, and *digital divide* did not appear in either period I or III. Appearing only in period II, this group of terms demonstrates the impact of digitization on Dublin Core research.

Some terms (*application profile*, *abstract model*, *data handling*, *domain ontology*, *e-Government*, *interoperability*, *MODS*, *open access*, *ontology*, *OAI-PMH*, *Semantic Web*, *quality control*, *health information*) appeared in both periods II and III. Other terms (*digital repository*, *DSpace*, *institutional repository*) occurred only in period III and represent the newly emerging themes in research on Dublin Core. This research has shifted from general research to more specific applications in domains such as e-Government and health informatics. One of the distinct themes is the application of Dublin Core to information repositories. Dublin Core as an information organization mechanism has been used to improve access to a variety of information repositories. It is not surprising that new information technology like abstract modelling, ontology, and Semantic Web are associated with it.

Several terms(*cataloguing, controlled vocabulary, Dublin Core, digital library, database management, DCMES, EAD, electronic resource, information system, information management, World Wide Web, information retrieval, metadata, metadata element set, meta tag, MARC, AACR, metadata schema, RDF, standard, semantics, Web resource, authoring*) appeared in all the three periods. This suggests that these subject topics are fundamental and essential to research on Dublin Core. MARC, EAD, AACR, electronic resource, information retrieval, cataloguing, and controlled vocabulary are traditional library and information science research topics. These topics have a natural and inherent relationship with Dublin Core because they address information description, information organization, information discovery, and information access.

Entire period (1997 to 2011) MDS analysis

Figure 2 shows MDS analysis result for the entire period from 1997 to 2011. Both the stress value (0.01826) and R^2 value (0.99921) indicate that the MDS analysis result (see table 3) is good and acceptable. Since the visual space is limited and the number of displayed terms is relatively large, the terms from the master file list were grouped into 10 cluster themes coded from C1 to C10.

The theme of C1 can be characterized as digitization, content, description, and Web; theme of C2 as retrieval, interoperability, and standard; theme of C3 as management, and metadata structure; theme of C4 as ontology, application, and Web; theme of C5 as multimedia, control, and system; theme of C6 as standard; theme of C7 as repository; theme of C8 as access, service, and control; and theme of C9 as application, bibliography, subject, description, and multimedia. C10 contains the query term, *Dublin Core* (k14), and its related general term, *metadata* (k35).

Notice that C10 is far away from other clusters. *Dublin Core* is the central term in this study with each of the extracted terms closely related to it. Without the term *Dublin Core*, the other terms would not have been included in the study. Therefore, *Dublin Core* tends not to cluster with any of these terms. At the same time, the term *metadata* usually co-occurs with the term *Dublin Core* in the retrieved documents. As a result they are projected together onto the same area in the visual space.

It is not surprising that various information standards, protocols, and rules like Resource Description Framework (RDF), Machine-Readable Cataloging (MARC), Moving Picture Experts Group (MPEG), Hyper Text Markup Language (HTML), Encoded Archival Description (EAD), Standard Generalized Markup Language (SGML), Visual Resources Association Code (VRA Code), Anglo-American Cataloguing Rules (AACR), Metadata Object Description Schema (MODS), and Open Archives Initiative -Metadata Harvesting Protocol (OAI-PMH) were associated with Dublin Core research. These information standards, protocols, and rules range from bibliographic information, to archival, to Internet information, to moving picture, to image, and so forth. and are related to Dublin Core metadata standards.

The basic terms that appeared in all the three periods were distributed in 6 of the 10 clusters (C1, C2, C3, C4, C5, and C10). Since Dublin Core is

Figure 2: Visual display for the entire period (1997–2011)

C1: digital library (k15), DCMES (k20), indexing (k29), RDF (k58), XML (k73), Web resource(k71), digital content (k79), World Wide Web (k67), semantics (k62)

C2: cataloguing (k9), controlled vocabulary (k10), interoperability (k28), information retrieval (k33), MARC (k41), Semantic Web (k58), search engine (k59), standard (k61)

C3: DCMI (k12), database management (k16), information management (k31), metadata standard (k36), meta tag (k39), metadata schema (k43), MPEG-7 (k46), resource discovery (k55)

C4: application profile (k1), AACR (k4), digital repository (k17), domain ontology (k21), e-Government (k23), EAD (k24), HTML (k27), MODS (k48), ontology (k50), VRA Core (k66), Website (k70), W3C (k72), classification (k78), authoring (k80)

C5: audiovisual library (k2), digitization (k18), electronic resource (k25), information system (k30), multimedia (k37), metadata element set (k38), MeSH (k40), quality control (k75)

C6: SGML (k63), Z39.50 (k74), metadata classification schema (k45), metadata format (k44)

C7: DSpace (k19), institutional repository (k32)

C8: data handling (k13), bibliographic control (k7), Web service (k69), opens access (k49)

C9: open archive initiative (k51), bibliographic record (k6), bibliographic citation (k8), abstract model (k5), OAI-PMH (k52), museum (k42), Web mining (k68), resource description (k56), subject gateway (k57), collection-level metadata (k11), metadata creation (k47), digital divide (k22), subject heading (k64), schema (k60), video (k65), abstracting (k3), photograph (k53), image (k77), health information (k76), hypertext (k26), keyword (k34)

C10: Dublin Core (k14), metadata (35)

Table 3: Summary of the research periods investigation

Periods	RSQ (R^2)	Stress Value	Number of Keywords	Number of Clusters
1997–2011	0.99921	0.01826	80	10
1997–2001	0.99834	0.02576	58	6
2002–2006	0.99857	0.02445	67	7
2007–2011	0.99436	0.04752	42	7

implemented and operated on the Web, it is natural that Web-related terms like *Semantic Web*, *World Wide Web*, *Web resource*, *website*, *Web service*, and *Web mining* show up in 5 of the clusters in figure 2.

Period I (1997–2001) MDS analysis

Figure 3 shows MDS analysis result for period I from 1997 to 2001. Both the stress value (0.02576) and R^2 value (0.99834) indicate that the MDS analysis

Figure 3: Visual display for period I (1997–2001)

C1: DCMES (k20), indexing (k29), RDF (k54), World Wide Web (k67), Web resource (k71), XML (k73)

C2: abstracting (k3), cataloguing (k9), controlled vocabulary (k10), DCMI (k12), digital library (k15), HTML (k27), information retrieval (k33), MeSH (k40), resource discovery (k55), search engine (k59), metadata standard (k36)

C3: Z39.50 (k74), SGML (k63)

C4: audiovisual library (k2), AACR (k4), information management (k31), multimedia (k37), meta tag (k39), MARC (k41), metadata schema (k43), MPEG-7 (k46), standard (k61), semantic (k62), digital content (k79)

C5: subject gateway (k57), subject heading (k64), hypertext (k26), electronic resource (k25), resource description (k56), database management (k16), metadata format (k44), VRA Core (k66), bibliographic control (k7), information system (k30), authoring (k80), bibliographic record (k6), EAD (k24), bibliographic citation (k8), photograph (k53), metadata element set (k38), classification (k78), image (k77), Website (k70), metadata creation (k47), Web service (k69), museum (k42), Web mining (k68), schema (k60), video (k65), W3C (k72)

C6: Dublin Core (k14), metadata (k35)

result (see [table 3](#)) is good and acceptable. Six clusters were created for this period.

In [figure 3](#), cluster C1 is characterized as Internet, resource, and description; C2 is characterized as retrieval, vocabulary control, description, and resources; C3 is characterized as standard; C4 is characterized as multimedia, description, and schema; C5 is characterized as bibliography, subject, multimedia, resources, Web, metadata structure, and creation; and C6 consists of Dublin Core and metadata.

Points of interest for period I include the following:

- C4 and C5 include all multimedia related terms. C4 includes audiovisual library, multimedia, and MPEG-7; and C5 includes image and video.
- C2 and C5 cluster traditional information organization terms in library and information science. C5 includes subject gateway, subject heading, resource description, bibliographic control, bibliographic, record, EAD, bibliographic citation, and classification; and C2 contains abstracting, cataloguing, and controlled vocabulary.
- Some terms appear only in period I. These include abstracting, resource discovery, and MeSH from C2 and all of C3 terms (SGML and Z39.50).
- Semantics emerges as a theme.
- Museum becomes the first metadata application domain (period I, C5).

Period II (2002–2006) MDS analysis

[Figure 4](#) shows MDS analysis result for period II from 2002 to 2006. Both the stress value (0.02445) and R^2 value (0.99857) from this data set indicate that the MDS analysis result is good. Seven clusters are created in this period.

Cluster C1 is characterized as semantic and digitization; C2 is characterized as application, ontology, and system; C3 is characterized as Web, ontology, retrieval, and standard; C4 is characterized as subject, bibliography, and application; C5 is characterized as application, process, bibliography, metadata creation, Web, and multimedia; C6 is characterized as retrieval and organization; and C7, like Period I, still consists of Dublin Core and metadata.

Key elements of period II are as follows:

- E-Government in C2 and health information in C5 emerge as new metadata application domains.
- Ontology, which covers knowledge representation and description, emerges as a theme. It indicates that Dublin Core research has moved to a new direction, which reveals sophisticated, complicated, and semantic relationships of the described objects.
- In this period, Dublin Core was widely applied to a variety of digital resources. As a result, issues related to Dublin Core implementation, creation, operation, and application such as metadata handling and quality control, open access, and interoperability with other information standards or protocols popped up.
- This was the booming period (2002–2006) of emerging research on Dublin Core for this study.

Figure 4: Visual display for period II (2002–2006)

C1: digital library (k15), semantics (k62), XML (k73), Semantic Web (k58), digital content (k79)

C2: database management (k16), domain ontology (k21), e-Government (k23), electronic resource (k25), information system (k30), metadata element set (k38), classification (k78)

C3: application profile (k1), cataloguing (k9), DCMI (k12), digitization (k18), information retrieval (k33), metadata standard (k36), MARC (k41), metadata schema (k43), MPEG-7 (k46), ontology (k50), World Wide Web (k67), Website (k70), Web resource (k71)

C4: AACR (k4), abstract model (k5), bibliographic record (k6), bibliographic citation (k8), collection-level metadata (k11), EAD (k24), HTML (k27), museum (k42), metadata classification schema (k45), metadata creation (k47), MODS (k48), open access (k49), OAI-PMH (k52), resource description (k56), subject gateway (k57), subject heading (k64), VRA Core (k66), Web mining (k68), quality control (k75), C5: audiovisual library (k2), bibliographic control (k7), data handling (k13), DCMES (k20), digital divide (k22), controlled vocabulary (k10), information management (k31), keyword (k34), meta tag (k39), open archives initiative (k51), video (k65), Web service (k69), W3C (k72), health information (k76), authoring (k80), multimedia (k37)

C6: interoperability (k28), indexing (k29), RDF (k54), search engine (k59), standard (k61)

C7: Dublin Core (k14), metadata (k35)

Period III (2007–2011) MDS analysis

Figure 5 shows MDS analysis result for period III (2007–2011). Both the stress value (0.04752) and R^2 value (0.99436) indicate that the MDS analysis result (see table 3) is good and acceptable. Seven clusters are created for this period.

Cluster C1 is characterized as creation, management, and digitization; C2 is characterized as repository; C3 is characterized as standard, retrieval, description, and application; C4 is characterized as creation, retrieval, Web, standard, multimedia, and application; C5 is characterized as application, ontology, Web, and control; C6 is characterized as schema; and C7, like periods I and II, consists of Dublin Core and metadata.

Key points for period III include the following:

- A salient theme of institutional repository emerges. Other related terms to this theme include digital repository and DSpace. In fact, DSpace and institutional repository form a clear cluster (C2) in figure 5.
- This period represents a stable and mature phase of research on Dublin Core.

Summary of visual MDS analysis results for the three periods and the entire period is listed in table 3. Note that all values of RSQ (R^2) in this study were larger than 0.90 and all stress values were smaller than 0.10, meaning that each of the results was sound and plausible.

Conclusion

Metadata such as Dublin Core have been used to enhance digital information access in an electronic environment, especially for information resources on the Internet. Dublin Core plays an important role in organizing digital information resources and improving visibility of digital information in a search engine ranking list. For these reasons, research on Dublin Core has attracted the attention of many fields such as library and information science, computer science, and communication. Each field has made remarkable contributions to this research. Revelation of involved subject topics and their semantic relationships would offer insight into Dublin Core research and help people better understand it.

Toward this aim, research papers on Dublin Core published between 1997 and 2011 were retrieved from the authoritative, comprehensive, high-quality, and multidisciplinary bibliographic databases of Web of Science. After identification and selection of relevant articles, related subject terms were extracted, regularized, and organized into data sets to which the MDS visualization analysis method was applied. This enabled the investigation of semantic relationships among related subject terms.

These data sets sorted the retrieved documents into three time periods: period I (1997–2001), period II (2002–2006), and period III (2007–2011), which enabled the observation of the evolution of subject topics used in research on Dublin Core. Combining these data sets presented an overall picture of all related terms. This study identified and analysed obsolete subject topics, newly emerging subject topics, and basic subject topics in Dublin Core research.

Figure 5: Visual display for period III (2007–2011)

C1: cataloguing (k9), database management (k16), digital library (k15), digital repository (k17), DCMES (K20), interoperability (k28)

C2: DSpace (k19), institutional repository (k32)

C3: controlled vocabulary (k10), e-Government (k23), information management (k31), meta tag (k39), MARC (k41), metadata format (k44), open access (k49), standard (k61), semantics (k62)

C4: abstract model (k5), data handling (k13), information retrieval (k33), metadata element set (k38), OAI-PMH (k52), photograph (k53), World Wide Web (k67), Web resource (k71), authoring (k80), image (k77)

C5: application profile (k1), domain ontology (k21), EAD (k24), electronic resource (k25), information system (k30), ontology (k50), RDF (k54), Semantic Web (k58), health information (k76), quality control (k75)

C6: AACR (k4), MODS (k48), metadata schema (k43)

C7: Dublin Core (k14), metadata (k35)

After extraction from the three databases of Web of Science, regularization of terms, and application of the frequency cut-off point, 80 related terms formed the term master file. Periods I, II, and III included 58, 67, and 42 regularized terms, respectively, which were obtained from 43, 47, and 23 retrieved documents, respectively. Both the number of the relevant documents and the number of the extracted terms suggest that period I is the initial phase, period II is the booming phase, and period III is the mature phase for research on Dublin Core. After one decade, this research has stabilized although new themes keep appearing.

Application domains found in this study include museum, archival, health information, and e-Government. Salient themes are ontology, institutional repository, and semantics.

In MDS analyses, term relationships for the three periods and for the entire period were illustrated in separate three-dimensional visual spaces. All the values of RSQ (R^2) in this study were larger than 0.90, and all stress values were smaller than 0.10, meaning that each of the results was sound and plausible.

The implications of this study are twofold. The first one is at a macro-level. Dublin Core has a close relationship with other data exchange standards. Studies on the comparisons and compatibility between Dublin Core and other standards are no longer current research focuses. Application of Dublin Core to emerging areas will attract the attention of researchers. Using new information technology such as semantic analysis to effectively manage and control access to Internet information as well other electric information resources will continue to be a major theme of Dublin Core study.

The second implication of this study is at a micro-level. The results of this study can be used to revise and enrich existing thesaurus or subject headings. This study revealed many terms relevant to Dublin Core. After these terms are examined and evaluated, meaningful terms can be added to the thesaurus or subject headings. In addition, the results can also be used to improve the effectiveness of information retrieval systems by providing users with more terms relevant to Dublin Core.

An information visualization method like MDS technique is a unique way to present information and analyse information. It has both strengths and weaknesses. The visualization space is a foundation of information presentation where objects are projected, connections are illustrated, and analysis is conducted. MDS analysis is weakened when it is hard for people to fully understand the visual space and the relationships among the projected objects in that space. Visualization space needs to be defined in a way that can be easily understood for the method to be effective.

Future research directions include, but are not limited to, adding new period(s) to the temporal subject analysis to obtain new topics, applying the same research method to other specialized metadata schemas, and making comparisons with the findings of this study.

References

- Adams, Jonathan, Louise Jackson, and Stuart Marshall. 2007. "Bibliometric Analysis of Interdisciplinary Research." *Report for Higher Education Funding Council for England*, http://webarchive.nationalarchives.gov.uk/20060829143900/hefce.ac.uk/pubs/rdreports/2007/rd19_07/.
- Allinson, Julie. 2008. "Describing Scholarly Works with Dublin Core: A Functional Approach." *Library Trends* 57 (2): 221–43. <http://dx.doi.org/10.1353/lib.0.0034>.
- Amaravadi, Chandra S. 2005. "Digital Repositories for E-Government." *Electronic Government* 2 (2): 205–18. <http://dx.doi.org/10.1504/EG.2005.007095>.
- Archibald, G., and M.B. Line. 1991. "The Size and Growth of Serial Literature 1950–1987, in Terms of the Number of Articles per Serial." *Scientometrics* 20 (1): 173–96. <http://dx.doi.org/10.1007/BF02018154>.
- Chandrakar, Rajesh. 2005. "An Approach to Mapping CCF to Dublin Core." *Electronic Library* 23 (5): 577–90. <http://dx.doi.org/10.1108/02640470510631290>.
- Chua, Alton Y. K., and Christopher C. Yang. 2008. "The Shift Towards Multi-disciplinarity in Information Science." *Journal of the American Society for Information Science and Technology* 59 (13): 2156–70. <http://dx.doi.org/10.1002/asi.20929>.
- Darmoni, S.J., B. Thirion, J.-P. Leroy, M. Douyère, and J. Piot. 2001. "The Use of Dublin Core Metadata in a Structured Health Resource Guide on the Internet." *Bulletin of the Medical Library Association* 89 (3): 297–301.
- Desai, Bipin C. 1997. "Supporting Discovery in Virtual Libraries." *Journal of the American Society for Information Science* 48 (3): 190–204. [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199703\)48:3<190::AID-AS12>3.0.CO;2-S](http://dx.doi.org/10.1002/(SICI)1097-4571(199703)48:3<190::AID-AS12>3.0.CO;2-S).
- Devey, Margaret, and Marie-Claude Côté. 2006. "The Development and Use of Metadata Application Profiles: The Government of Canada Experience." *Serials Librarian* 51 (2): 103–15. http://dx.doi.org/10.1300/J123v51n02_08.
- Eysenbach, G., C. Koehler, G. Yihune, K. Lampe, P. Cross, and D. Brickley. 2001. "A Metadata Vocabulary for Self- and Third-party Labeling of Health Web-sites: Health Information Disclosure, Description and Evaluation Language (HIDDEL)." www.ncbi.nlm.nih.gov/pmc/articles/PMC2243523/pdf/procamiasymp00002-0208.pdf.
- Greenberg, Jane. 2001. "A Quantitative Categorical Analysis of Metadata Elements in Image-applicable Metadata Schemas." *Journal of the American Society for Information Science and Technology* 52 (11): 917–24. <http://dx.doi.org/10.1002/asi.1170>.
- Howarth, Lynne C. 2003. "Designing a Common Namespace for Searching Metadata-enabled Knowledge Repositories: An International Perspective." *Cataloging & Classification Quarterly* 37 (1/2): 173–85. http://dx.doi.org/10.1300/J104v37n01_12.
- Hunter, Jane, and Liz Armstrong. 1999. "A Comparison of Schemas for Video Metadata Representation." In *The 8th International World Wide Web Conference*, Toronto, 1999, Amsterdam: Elsevier, 1431–51. [http://dx.doi.org/10.1016/S1389-1286\(99\)00053-5](http://dx.doi.org/10.1016/S1389-1286(99)00053-5).
- Korfage, Robert R. 1997. *Information Storage and Retrieval*. New York: Wiley.
- Leydesdorff, Loet. 2006. "Can Scientific Journals Be Classified in Terms of Aggregated Journal-Journal Citation Relations Using the Journal Citation Reports?" *Journal of the American Society for Information Science and Technology* 57 (5): 601–13. <http://dx.doi.org/10.1002/asi.20322>.

- Ma, Jin. 2007. *Metadata, SPEC Kit 298*. Washington, DC: Association of Research Libraries.
- Malet, Gary, Felix Munoz, Richard Appleyard, and William Hersh. 1999. "A Model for Enhancing Internet Medical Document Retrieval with 'Medical Core Metadata'." *Journal of the American Medical Informatics Association* 6 (2): 163–72. <http://dx.doi.org/10.1136/jamia.1999.0060163>.
- Milojević, Staša. 2009. "Big Science, Nano Science?: Mapping the Evolution and Socio-cognitive Structure of Nanoscience/Nanotechnology Using Mixed Methods". PhD dissertation, University of California, Los Angeles.
- Mohamed, Khaled A. F. 2006. "The Impact of Metadata in Web Resources Discovering." *Online Information Review* 30 (2): 155–67. <http://dx.doi.org/10.1108/14684520610659184>.
- National Information Standards Organization. 2007. "ANSI/NISO Z39.85-The Dublin Core Metadata Element Set.," https://www.ftb.ca.gov/aboutFTB/Projects/ITSP/Dublin_Core.pdf.
- Park, Eun G., Manon Lamontagne, Amilcar Perez, Irina Melikhova, and Gregory Bartlett. 2009. "Running Ahead Toward Interoperable E-Government: The Government of Canada Metadata Framework." *International Journal of Information Management* 29 (2): 145–50. <http://dx.doi.org/10.1016/j.ijinformgt.2008.06.003>.
- Patridge, Jeff, and Gonza Namulanda. 2008. "Describing Environmental Public Health Data: Implementing a Descriptive Metadata Standard on the Environmental Public Health Tracking Network." *Journal of Public Health Management and Practice* 14 (6): 515–25. <http://dx.doi.org/10.1097/01.PHH.0000338363.20962.f5>.
- Porter, Allan L., Alex S. Cohen, J. David Roessner, and Marty Perreault. 2007. "Measuring Researcher Interdisciplinary." *Scientometrics* 72 (1): 117–47. <http://dx.doi.org/10.1007/s11192-007-1700-5>.
- Sokvitne, Lloyd. 2000. "An Evaluation of the Effectiveness of Current Dublin Core Metadata for Retrieval." In *Victorian Association for Library Automation Conference in Victorian, Australia, 2000*. Melbourne: Victorian Association for Library Automation.
- Sutton, Stuart A. 1999. "Conceptual Design and Deployment of a Metadata Framework for Educational Resources on the Internet." *Journal of the American Society for Information Science American Society for Information Science* 50 (13): 1182–92. [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(1999\)50:13<1182::AID-ASI4>3.0.CO;2J](http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:13<1182::AID-ASI4>3.0.CO;2J).
- Van Raan, A.F.J., and Th.N. Van Leeuwen. 2002. "Assessment of the Scientific Basis of Interdisciplinary, Applied Research: Application of Bibliometric Methods in Nutrition and Food Research." *Research Policy* 31 (4): 611–32. [http://dx.doi.org/10.1016/S0048-7333\(01\)00129-9](http://dx.doi.org/10.1016/S0048-7333(01)00129-9).
- Weibel, Stuart L. 2009. "Dublin Core Metadata Initiative: A Personal History." In *Encyclopedia of Library and Information Science*, 3rd ed., ed. Marcia J. Bates and Mary Niles Maack, 1655–63. Boca Raton, Fla.: CRC Press.
- Weibel, Stuart L., Jean Godby, Eric Miller, and Ron Daniel. 1995. OCLC/NCSA Metadata Workshop Report. <http://xml.coverpages.org/metadata.html>.
- Weibel, Stuart L., Renato Iannella, and Warwick Cathro. 1997. The 4th Dublin Core Metadata Workshop Report. www.dlib.org/dlib/june97/metadata/06weibel.html.
- Weibel, Stuart L., John A. Kunze, Carol Lagoze, and Misha Wolf. 1998. RFC 2413-Dublin Core Metadata for Resource Discovery. <http://gamay.tools.ietf.org/html/rfc2413>.

- Zhang, Jin, and Alexandra Dimitroff. 2004. "Internet Search Engines Response to Metadata Dublin Core Implementation." *Journal of Information Science* 30 (4): 310–21. <http://dx.doi.org/10.1177/0165551504045851>.
- . 2005a. "The Impact of Webpage Content Characteristics on the Webpage Visibility in Search Engine Results (Part I)." *Information Processing & Management* 41 (3): 665–90. <http://dx.doi.org/10.1016/j.ipm.2003.12.001>.
- . 2005b. "The Impact of Metadata Implementation on the Webpage Visibility in Search Engine Results (Part II)." *Information Processing & Management* 41 (3): 691–715. <http://dx.doi.org/10.1016/j.ipm.2003.12.002>.
- Zhang, Jin, Dietmar Wolfram, Peiling Wang, Yi Hong, and Rick Gillis. 2008. "Visualization of Health Subject Analysis Based on Query Term Co-occurrences." *Journal of the American Society for Information Science and Technology* 59 (12): 1933–47. <http://dx.doi.org/10.1002/asi.20911>.