INTEROPERABILITY BETWEEN CLASSIFICATION SYSTEMS USING METADATA

Rosa SAN SEGUNDO

Carlos III University of Madrid

rsan@bib.uc3m.es

Metadata are structures which catalogue, classify, describe and articulate electronic information. The *Subject* element of Dublin Core is used for classification systems and subject headings. There are five ways of applying semantic interoperability: interoperability between controlled vocabularies in the same language; between controlled vocabularies in different languages and classification systems; between subject headings and classification systems; between classification systems; and between languages. The relations between diverse types of standards or systems present diverse difficulties. The electronic information container, which is Internet, guarantees the trend to try and achieve the interoperability of content analysis, whether it be between classification systems, or subject headings. The organisation of information in a physical format has transferred its organisational forms to the structuring of electronic information. The digital format transforms the organisational form itself. If, in information *the message is the medium*, in organisation *the structure is the medium*.

1. Introduction

Metadata are data which describe other electronic information data[1], such as web page references, correlation to the bibliographical description of web pages and electronic resources. They represent electronic information. Not only do they embrace formal descriptions, locate target objects on Internet and establish links[2], but they also encompass content.

Metadata are structures which catalogue, classify, describe and articulate electronic information. Their primary function is to identify and locate information data. They also encompass the semantic and syntactic function of electronic information and, moreover, address the organisation and visualisation of electronic information. All of the above converges in one primordial function: to attempt to facilitate interoperability both of formal structures as well as of content.

Several metadata models exist which include those that refer to library regulations such as the MARC format. There are also ones which refer to the adoption of a formal language such as XML or SGML, or metadata structures which offer great flexibility such as the Dublin Core, which has been widely accepted in bibliographical and library applications where it is used as an element heading in HTML.

The 15 elements which comprise the Dublin Core format are used as descriptors. The Subject element is used for classification systems and subject headings.

One of the most important aspects of the Dublin Core metadata is that the SUBJECT element comprises key words and includes a description of the full contents of the document or the source. In other words, it comprises the classes of a classification system, descriptors taken from a Thesaurus, subject headings and others. Furthermore, the "type" and "scheme" qualifiers provide information related to classifying systems via the <METANAME="DC.Materia"CONTENT=" element.

The "scheme" and "type" qualifiers[3] identify the system used in the description of each element. "Schemes" refer to the general standards or, in other words, the classification systems used and their editions, as well as to the lists of subject headings. The scheme, moreover, allows optional qualifiers to be used for each element which make it possible to indicate the regulations used in the event of regular bibliographical description standards having been used. Consequently, the "Schemes" can be expressed online and the "Types" refer to the data or elements of these general standards. "Types" cannot generally be expressed like this. Moreover, the Dublin Core elements are optional and repeatable and any possible ambiguities that may arise with the Dublin Core elements can be avoided by using HTML links.

The syntax of the Dublin Core subject element is expressed in the following way[4] :

<META NAME="DC.Subject"CONTENT="(Scheme=LCSH)Nursing-Dictionaries

<META NAME="DC.Subject"CONTENT="(Scheme=UDC)

946.O"1936/1939"

<META NAME="DC.Subject"CONTENT="(Scheme=CDD) 398.2

<META NAME="DC.Subject"CONTENT="(Scheme=CDD22) 520.60

<META NAME="DC.Subject"CONTENT="(Scheme=LCC) KDK710

<META NAME="DC.Subject"CONTENT="(Scheme=CC) O123,3J47

The subject element encompasses the document content and correlates to the fields of the MARC format. In other words, the 080 field is used to state the notation related to a classification system, and the 600 fields to the reduction of the document contents in alphabetical form. And, specifically: 653 for entry by key word, Index Term, Uncontrolled; 650 for entry using subject heading LCSH, MeSH; 050 Call Number/Classification number; 082 Dewey Decimal Call Number/Classification number; 080 Universal Decimal Classification Number.

It is essential to make subject access compatible. In other words, to achieve interoperability between the diverse knowledge organisation systems and the controlled vocabularies in order to meet users' new requirements and to be able to correct many current access defects within the wide electronic context by means of content.

Consequently, an attempt will be made to create multilingual vocabularies using already existing tools by means of which there is already a good deal of processed, structured and organised information. These include already created systems, such as the DDC, the UDC, LCSH or others, taking advantage of their multilingual capacity when they have it. Furthermore, the need to articulate semantic interoperability or equivalency between classification systems and subject headings exists and is already inevitable.

In order to establish this concordance, it is necessary to use authority registries, in an automated manner, since comparing terms manually requires great intellectual effort whereas comparing and managing large databases is much more profitable although it is still checked manually. Many projects already exist which compare both languages and structures, and many of them already include multiple structures or, in other words, they incorporate diverse tools for their equivalency.

2. Ways of applying semantic interoperability

The semantic interoperability are the differents concepts of the subject gateways created like an instruments to establishment equivalences between classification systems, subject headings and subject descriptions which are the structure which will form the backbone the access since differents forms of the so-called digital or virtual libraries. At all times, this points to the universal virtual library or universal bibliographical catalogue projected by Otlet and La Fontaine.

Interoperability between subject headings and classification systems may also articulate interconnection between diverse document languages in different tongues. In information recovery, users should not have to concern themselves with the articulation of different recovery languages, but the ideal thing would be for the user to formulate one single search instead of formulating it in different ways and in different catalogues. The ideal thing would be to make different controlled vocabularies and classification systems interoperable.

One of the primary objectives of library catalogues has been interoperability between conventional classification systems such as UDC, DDC, LCC, and others published with subject headings linked to classification numbers, known as a chain procedure masterminded by Ranganathan. The system which impacted most was Dewey Decimal Classification, already in its 22nd edition, which is linked to the subject heading list of the Library of Congress in Washington. The electronic version of the subject headings of the Library of Congress in Washington (LCSH) are also linked to the classification of the same library.

2.1.Interoperability between subject headings in the same language

The interoperability between subject headings in the same language, for example Library of Congress Subject Headings (LCSH) and the Medicine Subject Headings (MeSH), where an attempt was made to integrate subject headings in online catalogues,

is worthy of mention. This instance of interoperability was methodologically exported to other subject headings, as it establishes relations between authority files, databases and metadata. It makes use of the MARC format and establishes relations between authority records of almost 10,000 registries, although the syndetic structure of LCSH cannot be fully completed. This has to be done manually [5].

The metathesaurus of the National Library of Medicine UMLS Unified Medical Language System of the USA has made over thirty vocabularies, subject headings, thesaurus terms and classification systems interoperable. In order to do this, it uses processing techniques based on lexical units. It is based on the construction of a specialised word list which includes over 180,000 entries and includes verbs, nouns, adjectives, etc. on Biomedicine. It also has a semantic network which includes 132 semantic types. This semantic network establishes the categorization of all the components of the metathesaurus, with 53 links between the semantic types or nodes in the network. The links are the relations between them[6].

2.2. Interoperability between Multilingual subject headings

Multilingual subject headings are established using a system of linguistic equivalencies, making use of the establishment of relations between terms in different languages. Nevertheless, comparing terms is not an easy task. If the comparison is established merely between the terms of the subject heading lists in different languages, it is called terminological comparison and addresses linguistic problems. If the comparison is based on the establishment of equivalency between equivalent authority registries in different languages, it shall be called semantic equivalency and semantic problems are addressed. And if equivalency is established by means of application, it will deal with syntactic equivalency and will tackle technology-related aspects.

The European multilingual project MACS (Multilingual Access to Subject) on interoperability was created in 1997 at the Conference of European National Libraries (CENL) to try and lessen the problem of multilingual access to European databases. Switzerland, as a multilingual country, was especially interested in the implementation of this project. It establishes interoperability between three lists of subject headings: Library of Congress Subject Headings (LCSH), RAMEAU, and SWD Deuthe Wisdon and also enables a common list of all three to be accessed, namely, SHL or Subject Heading Languages. The framework of this interoperability is articulated by means of a system of links and searches can be formulated in the four national libraries or in only one of them. The search can be undertaken using a list of subject headings or in all SHL. This also means that a search can be formulated in a library catalogue in another language. Furthermore, recovered registries can be visualised using some of the MARC formats such as USMARC, UNIMARC and also MAB and in three languages: English, French and German. All of the above has been undertaken and financed jointly by three National Libraries: the National Swiss Library, the National French Library, Die Deutche bibliotek and

the British Library. MACs are currently restricted to authorised headings and are only applicable as a dictionary of subject heading languages. The future incorporation of links to other elements of authority registries may result in the creation of a virtual multilingual authority file [7].

The multilingual database on French monumental heritage, MERIMEE, encompasses religious, civil, school, military and industrial architecture and aims to articulate interoperability between controlled vocabularies in different languages. It embraces three areas: the inventory undertaken by regional services, dossiers and old inventories included in PREDOC and historic monuments under protection since 1913 including the decree that registered and classified them. Thus the MEREMEE database comes within the context of five databases including THESAURUS, PALISY, MEMOIRE and ARCHIDOC[8].

It's exactly interoperability beteewn subject headings in diferents languages. For examplethere is an European project: HEREIN, the European information network on cultural heritage policies, lies in the Council of Europe and is financed by the European Union. It is sponsored by six countries which are Spain, France, Ireland, the United Kingdom, Norway and Hungary, and later Belgium.

It aims to articulate a new framework of collaboration on the subject of heritage among these countries. Its primary objective is to make it possible to interchange information on heritage policies. It also has other objectives such as collaboration in working groups via forums, associations, and collaboration in other heritage areas such as archives, libraries and museums. It makes these countries' heritage policy reports, which encompass access, protection and conservation, available. In order to achieve this, a thesaurus of key words used in diverse documentation such as heritage policy reports and others was drawn up. It was not based on any already existing thesauruses, but attempted to create a specialised multilingual thesaurus in three languages: English, French and Spanish, and it may be extended to other languages. It establishes hierarchical, equivalent and associative relations[9].

2.3. Interoperability between subject headings and classification systems in the same language

The advantage of Dewey for Windows 22, which articulates edition 22 of the Dewey Decimal Classification Tables, is that it is used in a large number of libraries in 135 countries and has been translated into 30 languages. In the USA, it is used in 95% of public and school libraries in addition to a high number of university and specialized libraries.

Its alphabetical index was one of the tools which introduced Dewey Decimal Classification into the framework of modern classification systems as, in addition to incorporating a new contemporary model for systematising science, where Theology no longer had central position, it also included recently created new sciences and

disciplines, such as Social Sciences. As regards the formal and structural aspects of the system itself, it included divisions and subdivisions, an explanation of the system, collective signs, notations, and finally an alphabetical index with references to all the classifying numbers. This alphabetical index would be the backbone to interoperability in classification systems from where the leap could be made to natural language although, on numerous occasions, the alphabetical indices of charts are called natural language[10].

Interoperability ranges from classifying numbers to alphabetical titles of charts, to notations of the same and even to the Library of Congress Subject Headings (LCSH), as already occurred with the last two editions of the *r* charts which could be used to produce these leaps. These can also be established between diverse classification systems and between diverse languages of the same classification system and between alphabetical and systematic systems. Thus, content interoperability is very broad-ranging.

Interoperability between Library of Congress Subject Headings (LCSH) and Library of Congress Classification (LCC) has been articulated in the so-called Classification Plus (in CD-ROM) and is also available in Classification Web, an interface which is being developed in the Library of Congress[11].

There is also another product which is based on the abbreviated edition of the UDC, which correlates and translates to the General Finish Subject Headings[12].

In Spain, interoperability is being undertaken between the subject heading list of the National Library in Madrid from the authority file and the UDC, a project which is sponsored by AENOR[13].

At this moment there are some project of interoperability between subjects headings and classification in the same language

2.4. Interoperability between subject headings and classification systems in different language

Although the British project, *HILT High Level Thesaurus Project* on a teaching thesaurus is originally from Great Britain, it also embraces Australia, Canada and the United States. It makes diverse controlled vocabularies interoperable, and tries to facilitate both search and subject navigation. Its interoperability encompasses the LCSH, the DDC and the UNESCO Thesaurus, UDC and AAT (Art and Architecture Thesaurus)[14].

The DARPA project deals with interoperability and comparisons using metadata between controlled entry languages and recovery languages and between controlled vocabularies and metadata vocabularies [15].

The European RENARDUS project addresses interoperability between specific classification systems, different controlled vocabularies in different languages and their

convergence in a common classification system such as Dewey Decimal Classification. It is a programme based on technical models and computer tools; consequently interoperability is primarily based on computing tools[16].

The most significant Polish project on interoperability between controlled vocabularies merges several controlled vocabularies such as SHL (subject headings language), TCT (Thesaurus of Common Topics), UDC (Universal Decimal Classification) and PTC (Polish Themaic Classification).

2.5. Interoperability between classification systems

The American Mathematics Society (AMS) is working on interoperability between Mathematics Subject Classification (MSC) and the DDC, specifically with class 510 related to mathematics in the State University of New Cork in Albany.

There is also an interoperability project being run between the Swedish Classification System (SAB) and the DDC. The project is financed by the Royal Library, the National Library of Sweden. The chart conversion is published by the library and can be found online[17].

3. The outlook for interoperability in classification systems

The relevance of the Dublin Core format lies in three aspects. Firstly, the Dublin Core subjects element; secondly, the syntax developed in HTML and, finally, management as a Web page with a structure proper to classification systems.

To sum up, the immediate future points to the creation of, access to and methods for organising the data from different materials into classified online libraries. On presenting a novel hypertextual and non-linear architecture, the organisation of information will furnish the organisation of knowledge with a new paradigm.

Furthermore, many pages have links with subjects (key words) as metaelements as they are used by large search engines. However, as yet no widely-used search engines undertake their searches using meta-questions, as there do not exist metadata indices of special relevance. Nevertheless, the Dublin Core Subject meta-question comprises the same information as other electronic documents processed using other systems. In web page design, HTML and metaidentification or metasignposting are used, that is to say, the key words function as metasignposts as they are used to index a web page. An attempt has been made to draw up an index of pan-thematic metadata which includes metainformation highlighted in the subject element. Scheme semantics and ontology are already essential; metadata registries contain information about the semantics, structure and syntax of metadata elements.

The relations between diverse types of standards or systems present diverse difficulties. The electronic information container, which is Internet, guarantees the trend to try and achieve the interoperability of content analysis, whether it be between

classification systems, or subject headings. Correlation between both can never be articulated by the positivist paradigm of total equity[18].

The organisation of information in a physical format has transferred its organisational forms to the structuring of electronic information. The new material substratum of electronic information conforms and delimits this new organisation. Thus, both physical organisation on shelves and its correlation to physical organisation in a catalogue correlate to the electronic organisation of information. The digital format transforms the organisational form itself. If, in information *the message is the medium,* in organisation *the structure is the medium.*

---

[1] SAN SEGUNDO, Rosa. *Organización del conocimiento en Internet. Metadatos bibliotecarios Dublin Core.* En: VI JORNADAS Españolas de Documentación, Valencia 1998. --Valencia : FESABID, 1998; P.805-817 http://www.florida-uni.es/~fesabid98/Comunicaciones/r-sansegundo.htm

2 SAN SEGUNDO, Rosa. *A new concept of knowledge.* ONLINE Information Review. 2002, Vol. 26 No.4. *http://thesius.emeraldinsight.com/vl=5405326/cl=115/nw=1/rpsv/cw/www/mcb/14684527/v26n4/contp1-1.htm*

[3] WEIBEL, S.; GLDBY, J.; MILLER, E. OCLC/NCSA *Metadata Workshop Report*

4 Dublin Core creation. http://www.sics.se/-preben/DC

[5] OLSON, Tony. *Integrating LCSH and MeSH in information systems.* IFLA, Dublin, OCLC, 2001

OLSON, Tony. *The Integration of information Languages and interoperability.* ALA Annual Conf. 2002

[6] Oficial page of Project *NATIONAL libray of Medicine Fact Sheeet: UMLS Metatesauro* http://www.nlm.nih.gov/factsheets/umlsmeta.htm

[7] LANDRY. Patrice *MACS Project : Multilingual Access to Subjects (LCSH, RAMEAU, SWD)* Swiss National Library Switzerland. 66th IFLA Council and General Conf.

[8] *MERIMEE* http://wwwculture.gouv.fr/documentation/merimee/accueil.html

[9] *European Heritage Network* http://www.european-heritage.net/en/index.html

[10] *DEWEY Decimal Classification 22* http://www.oclc.org/dewey/enhancements/enhancement03.htm

[11] *RESEARCH on Interoperability of Metadata in Classification Schemes construction of automatic mapping system between CLC and DDC*
http://org/metadataresearch/dcconf2004/papers/Paper_12.pdf

[12] HIMANKA , Janne ; VESA, Kauto. *Translation of the finifs Abrieged Editon of UDC into General Finish Subject Headings. In*: International Classificaiton 1992, 19 (3), p. 31-134

[13] Whit Carlos III University of Madrid, and National Library of Madrid SAN SEGUNDO MANUEL, R. *La clasificación automática mediante la CDU con el procedimiento en cadena.* En: I JORNADAS de Tratamiento y Recuperación de Información (JOTRI 2002). Valencia : Universidad Politécnica, 2002 ; p. 53-59 http://www.fiv.upv.es/jotri/Ponencias/Clasificacionauto.pdf

[14] HILT High Level Thesaurus Project http://hilt.cdlr.strath.ac.uk/index.html

[15] BUCKLAND, M Et el. *Mapping entry vocabulary to unifamiliar metadata D-lib Magazine* 5 (1) http://www.diib.org/dlib/january99/buckland/01/buckland.html

[16] RENARDUS http://www.renardus.org/ KOCH, Traugott, Neuroth, Heike and Day, Michael (2001). Renardus: Cross-browsing European subject gateways via a common classification system (DDC). In: "Subject Retrieval in a Networked Environment". Proceedings of the IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing and the IFLA Section on Information Technology, 14-16 August2001, Dublin, OH,USA U CIM Publications - New Series Vol. 25, Muenchen 2003. pp 25-33. Manuscript at: http://www.lub.lu.se/~traugott/drafts/preifla-final.html

[17] *CONVERSION between DDC 21 and SAB. National* Library of Sweden, IFLA, Secction de Classification and indexing, 2001, newsletter, n.24, http:www.kb.se/bus/konverteringstabelleng.htm

[18] SAN SEGUNDO, Rosa. *New Conception of Representation of Knowledge.* En: Knowledge Organization. International Journal. , Vol 31, 2004, n.2, p. 106-111