MATHEMATICS SUBJECT CLASSIFICATION and related classifications in the Digital World

Antonella De Robbio
e-mail: derobbio@math.unipd.it
Dario Maguolo
e-mail: dario@math.unipd.it
Mathematical Seminar Library, University of Padova-(Italy)

Alberto Marini
e-mail: alberto@iami.mi.cnr.it
Institute for the Applications of Mathematics and Informatics, National Council of Researches (IAMI-CNR), Milano (Italy)

Connecting classifications in the digital world

In the ever more pervasive and interconnected world of digital information, reliable connections among such knowledge representation and lexical, as well as information retrieval, tools as classifications, lists of subject headings, thesauri, terminological collections and ontologies, are a necessity for networked knowledge-based activities.

Users in different settings and with different demands and expectations want to fulfil their information needs wherever information is available, possibly cutting costs and times, regardless of the heterogeneity of sources: from quite specialized databases or dedicated portals to general online library catalogues or Web search engines, from reference (metadata) databases to full-text digital libraries, from e-journal aggregators to preprint servers and authors' self-archives (commonly called *e-print systems*).

The organization, the functionalities and the interaction modes exploited by networked digital libraries may be completely different from those generally met with in traditional paper-based libraries. Moreover, just on the line of e-print systems, the development of technical mechanisms and organizational structures to support their interoperability, which is promoted by the Open Archives Initiative (OAI) [see the Website www.openarchives.org], is making them evolve into genuine building blocks of a transformed scholarly communication model, radically different form the traditional one, which is dominated by the heavy mediation business of academic publishing companies.

On the other hand, users do not want to change their mind to meet the particular way of storing, indexing and presenting information for any source they face: this should be automatically worked out by the system. But such a task is not trivial. As for subject indexing, different classifications, thesauri or otherwise structured terminologies, while insisting over the same area, can keep presenting strong linguistic (which can not be worked out by mere translation), structural and semantic disagreements, in spite of any effort for harmonization. Dramatic disagreements are evidenced in passing from the specialized world of discipline-oriented classifications to general classifications widely used in public, school or even general academic libraries, such as Dewey Decimal Classification, Universal Decimal Classification or Library of Congress Classification.

Misinterpretations are easy to occur when the same words are used in different contexts or for different purposes. Nevertheless, effective connections between classifications are feasible,

provided that the objects each classification or the like refers to are identified unambiguously by means of a suitable representation language.

Classifications: functions, structure and dynamics

Classifications can be viewed as abstract structured spaces, or models for arranging material spaces, where respectively immaterial or material objects can get a location according to selected characteristics, so that objects of interest can be found just by choosing and moving along the paths provided by the space structure or the concrete arrangement defined by the model. Typical material objects being located by means of a classification are books shelved in a library, or even bibliographic entries in printed indexes; as for the immaterial, we can think to fields or disciplines of human knowledge or activity, to concepts and objects of a certain field or discipline, or generally to subjects of documents abstractly taken as information units.

Even if a classification is not involved in physical space arrangement, but works as a pure information device, e.g. in computer-managed bibliographic records, an appropriate semantics can be given in terms of some notion of space, possibly less constrained and more complex than usual material ones. Anyway, the classification space is the form of a container, a grossly operative space for concept packaging and package linking; it is quite different from the space of objects as they are actually intended by classification users, a space that can be more or less definitely taken off from the classification like a conceptual space of true effective meanings.

While visiting a classification space, structured descriptions, semistructured or unstructured textual descriptions, images and other forms of display help, according to specific conventions, to recognize places and objects located therein, or to switch form one place to another. Any place, and therefore any object, is assigned one or more addresses, which are numbers or character strings suited to identify the place through an encoding of the path(s) to reach it.

Obviously the descriptions, be they textual or otherwise performed, are not the objects or the structured space they refer to: they are means to orientate the user in the classification space. Descriptions refer to objects through the mediation of places that gather them, or channels that convey them, in order to meet some external specifications or constraints (human readability, manageability for use). So one description may refer to a collection of objects that are intended distinctly by the user, but are collected according to the classification organization. On the other hand, one object or place may be represented in different forms, still observing the linguistic or semiotic conventions of the classification. Thesauri and lists of subject headings, on the contrary, are worried to maintain a tight correspondence between objects and descriptions, at the price of bothering about preferred and non-preferred forms: but this amounts to constrain the variety of natural language to pass through the cog-wheel of machine identifiers. The addition of free text scope notes is a further signal of this blurring.

It's the role of addresses to guide the travel machinery: for this work there is no need to know why the traveller wants to reach a certain place, and to find what. So in subject classifications addresses (commonly named *classification codes* or *numbers*) are fundamental in their very form for material document shelving in material libraries, and lists of addresses are major means for subject indexing in bibliographic databases and online library catalogues; addresses encode and display the space structure, but they act as mere linking elements, without any real semantic content. The real carriers of semantic content are descriptions, and the classification organizes them inside a structure that exists independently from the actual forms of the addresses, i.e. from the forms that are fixed for representing the classification space structure in view of external reference and linking.

Moreover, while both descriptions and addresses can change, in time or across different linguistic, semiotic or encoding conventions, it is not necessary that they change in dependence from one another, or from the changes, transformations, births and deaths among the objects, the spaces and the ways objects and spaces are organized and perceived. Addresses may change while descriptions remain the same, or space structure, at least locally, is preserved; descriptions may change while objects remain the same; objects may change while addresses remain the same, and so on.
Different classifications that cover overlapping areas can exercise in time influence on one another, especially on structure and descriptions, in order to get similar or compatible views of the same objects, even if they are seen from different viewpoints or on different scales, and different groupings can be kept within each classification.

Displaying classification schemes: our achievements

While a number of approaches to the issues of connecting classifications or thesauri exploit statistical methods or neural network techniques, a different trend is oriented towards the analysis, modeling and support of conceptual organization by humans. The former can be very helpful, even in view of the latter; a well defined integration seems to be the recipe for the near future [see D01].
We are not working with statistical or neural methods; our first concern on these issues was directed to the generation of highly portable hypertexts and presentation modes, suitable to facilitate readability and discovery of meaning by humans in a generality of complex documentation structures as classification schemes, terminologies, metadata collections, etc. We are especially exploiting a presentation mode that allows moving to and fro parallel views of the same or similar structures; this proves very useful in our setting.
Such hypertexts are produced mainly by a pool of standard C programs, which operate only on sequential ASCII files and are aimed to the analysis and transformation of specific texts and to the generation of groups of syntactically simple but highly connected and JavaScript enriched HTML pages (*H-volumes*).
Various hypertextual frame presentations of the latest version of Mathematics Subject Classification, MSC2000 have been realized in this way and are collected in the *Mathematics Classification Page* at the server of the Math Department of Padova University (Web address: www.math.unipd.it/~biblio/math/index.html ; other access at the server of the Institute IAMI-CNR in Milano, Web address: www.iami.mi.cnr.it/~alberto/a0msc.htm). Here is a list of he groups of HTML pages, together with an indication of their features and their access address.

Group 1: Simple frame presentation
www.math.unipd.it/~biblio/math/mainb/mhbmain.htm
Group 2: Double view presentation
www.math.unipd.it/~biblio/math/doppiaeng/mhdmain.htm
Group 3: Simple frame presentation, including changes from MSC 1991
www.math.unipd.it/~biblio/math/complexc/mhcmain.htm
Group 4: Simple frame presentation, with links to subject specific pages of relevant Websites
www.math.unipd.it/~biblio/math/complexc/mhcmain.htm
Group 5: Simple frame presentation, Italian translation
www.math.unipd.it/~biblio/math/italiana/mhimain.htm

Group 6: Simple frame presentation, interleaved English and Italian texts
www.math.unipd.it/~biblio/math/it+eng/mhlmain.htm

We advanced on this line by laying down connections between classification numbers from the DDC 21 and MSC2000 schemes; a draft page in double frame presentation was then produced and is visible at the address http://www.math.unipd.it/~biblio/msc-cdd/index.html . In view of the revision of the 510 section of DDC, *Mathematics*, we are updating such a draft along the proposal presented by Giles Martin, Assistant Editor of the Dewey Decimal Classification, which is visible at the Web address www.oclc.org/dewey/updates/discussion/doc/request_for_comment.htm .

Buses in the classification space

Historically, Subject Classifications for documents owe their main monohierarchical, that is tree-like, structure to a consolidated habit in shelving and retrieving material documents in libraries, where a fixed space is definitively, or almost permanently, divided in sections and subsections, and relocating is expensive, given that a material object cannot be simultaneously in different material places. Such a habit consists of proceeding in choices from general to particular topic, and from large to smaller and smaller not overlapping divisions of space. Even in the digital world this habit is maintained and file systems display tree-like organizations.

However, the spaces and objects of knowledge and human activity bring up structures that are far more complex and dynamic than a simple fixed tree. At the crossroad of Artificial Intelligence, Computational Linguistics and Database Theory, these structures can be represented with good effectiveness in the frame of reference of Formal Ontology [G98], by means of formalisms as Conceptual Graphs (CG) [S84; for an application of a variant of CG, see GMV99], Description Logics (DL) [JoLC99; CDLNR98; see *The DL Website* at the address: www.ida.liu.se/labs/iislab/people/patla/DL/index.html], and the Unified Modeling Language (UML) [BJR98], which comes from the field of software engineering and is proposed as an approach for modeling ontologies and encoding the knowledge content of Web pages [C01].

Metadata formats for document representation are being defined progressively along this way; the draft for the *Academic Metadata Format*, which is being defined in the scope of the Open Archives Initiative (visible at the Web address: openlib.org/amf/doc/ebisu.html) is a clear example of such a trend. Further information on this topic can be found in [H01].

As we face with Subject Classifications, such representations (which were conceivable even in times when formal languages for expressing them were lacking) have yet to be cut down to get compliance with the tree-like forms in which Subject Classifications constrain their operability. Although this reduction comports unavoidably serious information losses, Subject Classifications have been provided with more or less effective devices to remedy for this gap.

From the pioneering work of Ranganathan since 1933 with Colon Classification, through the elaborations of the British Classification Research Group in the '50s and '60s, the addition of Auxiliary Tables to the Dewey Decimal Classification since its 18th edition, published in 1971, the development of the Preserved Context Indexing System (PRECIS) in the '70s, and the publication in 1986 of the standard ISO 2788 (BS 5723) *Guidelines for the establishment and development of monolingual thesauri*, a compositional approach to subject analysis, named *facet analysis*, has been progressively established [F96]. Within facet analysis, complex concepts are decomposed into specified combinations of atomic elements, which belong to homogenous, mutually exclusive classes, the *facets* [AG87].

Turning back to Subject Classifications, an organization of the classification space (named *pre-coordination*) which permits complex objects to be recovered via suitably compound addresses, and a more or less rich and organized apparatus of cross-references between places, are useful means especially if objects may be located in one place only.
If a Subject Classification is used in settings that allow the simultaneous employment of different classification codes for the same object, mechanisms and directions for *post-coordination* are provided in order to partially recover complex meaning by listing addresses together in suitable ways, either in databases that offer information or in queries that ask for it.
Without pretending to cut the fuzziness of natural language off, as thesauri apparently do, an analysis of descriptive texts, together with the recognition of structural dynamics in time, is still the basis for identifying the objects one or more classifications refer to, in the framework of the knowledge organization of people in charge of using those classifications. Textual identities or similarities in the descriptions inside one classification (detected from the same version – synchronically – or from different versions in time – diachronically –), or across different classifications (in the same language), give important cues for such identifications, although a word or phrase may have different meanings, even inside the same classification.
So the first step in the process of getting objects out of their envelope, that is the classification space, is to recognize the envelope as a structure which develops in time through a course of succeeding versions of the classification, moving across the addresses which mark the paths and the places in such a space. The output of this step consists of sequences of descriptions for *buses* in the classification space-time. Each bus during its trip passes through one or more places; the addresses of such places, with the indication of the period of passage, set up the schedule for that bus.
To work this guideline out, we have started an analysis of the whole Mathematics Subject Classification, along its evolution since 1959, as available for online searches in the *MathSci* database. The result is a relational database, for which we are going to define a Web presentation to be realized with an adaptation of the C programs already developed.

Identifying and describing objects out of the bus space

Even if any synchronic slice of the bus space structure is tree-like, the whole structure may not be tree-like, as nodes or subtrees can migrate from one branch to another; besides the main hierarchical structure, cross-references and explicitly stated pre-coordination and post-coordination mechanisms, taken dynamically as well, give substantial contributions to the definition of the classification space-time.
The further step of the identification process is the extraction of conceptual elements from the descriptions. A full relational analysis should then be performed by means of a suitable representation language.
Facilities for textual analysis give effective help at this stage. In this direction, we produced some H-volumes presenting the phrases extracted from the descriptions in pairs of classifications as DDC21 and MSC2000: phrases are circularly permuted on significant words, i.e. form a KWIC list. This redundant but properly paginated presentation allows one to explore rapidly the lexical similarities among categories and to obtain suggestions about their affinities of contents. Such kind of preliminary lexical support shall be worked out for investigating the connections among other groups of classification schemes.
Furthermore, some improvements obtainable adding to phrases discrimination of homonyms, synonyms and secondary terms shall be investigated.

Object descriptions in the metadata machine

Mathematics Subject Classification is one of the classification systems provided for by the Dublin Core (DC) metadata format, and is used inside DC metadata for the search engine developed in the European Union project *European Libraries and Electronic Resources in Mathematical Science* (EULER) – Web address: www.emis.de/projects/EULER/ .

The main objective of EULER was the realization of a "one-stop shop" for research on mathematics information resources such as books, pre-prints, Web pages, abstracts, collections of articles and reviews, periodicals, technical reports and theses. The result is a Web meta-interface for parallel simultaneous queries to a heterogeneous collection of databases.

A similar strategy could be exploited for connecting classifications: descriptions for objects identified out of different classification schemes could be conveyed into the metadata managed by the search engine.

REFERENCES

AG87 J. Aitchison, A. Gilchrist, "Thesaurus construction: a practical manual", ASLIB, 1987

BJR98 G. Booch, L. Jacobson, J. Rumbaugh, "The Unified Modeling Language User Guide", Addison-Wesley, 1998

C01 S. Cranefield, *Networked Knowledge Representation and Exchange using UML and RDF*, "Journal of Digital Information", 1(8), 2001. http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Cranefield/

CCMS96 L.M. Chan, J.P. Comaromi, J.S. Mitchell, M.P. Satija, "Dewey Decimal Classification: a practical guide. 2nd ed., revised for DDC 21", OCLC Online Computer Library Center, 1996

CDLNR98 D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, R. Rosati, *Description logic framework for information integration*, "Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR'98)", Morgan Kaufman, 1998. p. 2-13

D01 M. Doerr, *Semantic Problems of Thesaurus Mapping*, "Journal of Digital Information", 1(8), 2001. http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/

F96 A.C. Foskett, "The Subject Approach to Information", 5th ed., Library Association Publishing, 1996

G98b N. Guarino, *Formal Ontology and Information Systems*, "Formal Ontology in Information Systems: proceedings of FOIS'98", N. Guarino (ed), IOS Press, 1998. p 3-15

GMV99 N. Guarino, C. Masolo, G. Vetere, *Ontoseek: Content-based Access to the Web*, "IEEE Intelligent Systems", 14(3), 1999. p. 70-80

H01 J. Hunter, *MetaNet - A Metadata Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains*, "Journal of Digital Information", 1(8), 2001. http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/

JoLC99 "Journal of Logic and Computation" – Vol. 9, No. 3: "Special Issue on Description Logics"

S84 J. Sowa, "Conceptual Structures: Information Processing in Minds and Machines", Addison-Wesley, 1984