Classification for collections mapping and query expansion

Claudio Gnoli[1], Laura Pusterla[1], Anna Bendiscioli[1], Cristina Recinella[1]

Science and Technology Library, University of Pavia, Italy
bst@unipv.it

Abstract. Dewey Decimal Classification has been used to organize materials owned by the three scientific libraries at the University of Pavia, and to allow integrated browsing in their union catalogue through SciGator, a home built web-based user interface. Classification acts as a bridge between collections located in different places and shelved according to different local schemes. Furthermore, cross-discipline relationships recorded in the system allow for expanded queries that increase recall. Advantages and possible improvements of such a system are discussed.

Keywords: browsing · Dewey Decimal Classification · knowledge organization · mapping · OPAC

1 Introduction

Among the different types of knowledge organization systems (KOS), classification schemes offer additional functionalities with their ability to organize and present subjects in a systematic way, by means of their meaningful notation. This makes them especially suitable to act as user-friendly guides to the variety of information, allowing users to browse available subjects even in large collections.

We have used a classification scheme to organize materials owned by the three scientific libraries at the University of Pavia and to allow integrated browsing in their union catalogue. The Dewey Decimal Classification (DDC) has been chosen as it is the most widespread classification scheme in Italian libraries, which increases interoperability of the local catalogue with other information resources and tools. Among these is the Italian version of the online DDC schedules WebDewey, to which our libraries have recently subscribed.

2 Science Library Collections in Pavia

Scientific libraries at the University of Pavia were quite scattered until 2009, when they have been reorganized into 8 libraries, including the Science Library (Biblioteca delle Scienze, BDS), the Science and Technology Library (Biblioteca della Scienza e della

Tecnica, BST) and the Medical Library (Biblioteca di Area Medica, BAM). On the other hand, each of these libraries is still divided physically into several sections, each with a different tradition of shelving based on local schemes. While shelfmarks are being progressively converted to a DDC-based system, librarians still have to manage many old shelfmarks belonging to old schemes. The subdivisions often depend on historical or accidental factors, such as the actual position of departments in the town, rather than on subjects themselves; for example, most physics and chemistry books belong to BDS, but most engineering and mathematics ones belong to BST. Books on related subjects, or even on the same subjects, are often shelved in different places — a potential source of confusion for users.

In this situation, a standard classification scheme can work as a virtual bridge between different local schemes and locations. It can also be a useful organizational tool in the eventuality of a further unification between small sections of the same library (books will find the right position in the new shelf among other books of the same subject).

A known limitation of such enumerative classification schemes as DDC is that they force every item to be shelved under a specific discipline, thus losing information on related subjects also touched upon in the same work. For example, many books owned by our libraries deal with mathematical subjects applied to physics, or with physical subjects applied to engineering, or engineering subjects applied to building, or building subjects applied to architecture… Navigation across subjects and disciplines is only possible if appropriate links exist in the catalogue between related subjects [1], which is hardly the case with most Italian catalogues [2].

3 Relationships between Classes

In order to overcome these difficulties, we have developed SciGator, a freely accessible web interface [`http://www-dimat.unipv.it/biblio/deweye.php`] written in PHP, which allows users to browse and navigate between subjects available in the three scientific libraries before launching their search in the catalogue. SciGator data are stored in a MySQL table including fields for DDC notation, informal Italian caption, informal English caption, scope notes, related DDC classes, and equivalent non-DDC classes from local schemes.

While the fields for related DDC classes allow to manage navigation between different disciplinary hierarchies, thus linking e.g. `532` fluid mechanics with `627` hydraulic engineering, the fields for equivalent classes allow to manage mapping between local schemes. DDC thus can work as a standard common language for distributed resources [3-4], besides providing a language-independent notation that can organize resources with titles in different languages. In some cases mappings can be quite precise, like it happens with the local scheme of the Mathematics section, as this was originally based on the

Mathematics Subject Classification which defines mathematical subjects in a quite precise way, for which DDC has good correspondence. In other cases, however, mapping is only approximated, due to various inconsistencies in classing books in the past at various degrees of detail in different sections. Still, using a standard classification scheme as a common reference brings some order in the complex of available materials.

In our approach, a class can be related to more than one other class. E.g., for DDC 532 fluid mechanics we have recorded "see also" relationships to 530.42 liquid state physics and to 627 hydraulic engineering, as well as mappings to ZA.4 fluid mechanics at the Mathematics department, ID4 fluid mechanics at the Engineering department, etc.

In most cases, these relationships are symmetrical, that is 530.42 liquid state physics also links back to 532 fluid mechanics. However, for some approximated mappings, such as between 532 and ID4, in order to reduce noise in retrieval the links only work in one direction: users browsing physics classes in DDC are warned that an (approximately) equivalent local class exists, as this is displayed, but such class is not included in expanded search (see below). Practical experience with such situations could lead to develop a more accurate model distinguishing between several types of associative relationships.

4 The SciGator Interface

As users access SciGator homepage, they are presented with the first two hierarchical degrees in the scheme, only for those DDC classes for which our libraries actually own some books (Fig. 1). For example, religion classes (200) are not displayed as our scientific libraries do not own any significant number of documents concerning religion.

Users can select a second-order class, e.g. 530 physics, which is then expanded to show all its subclasses and related classes (Fig. 2). This allows for classical browsing of the hierarchical classification tree. Navigation can go down to special classes with several additional digits, or go up back to more general classes by two degrees every click; the interface is designed so that navigation only requires a limited number of clicks, without forcing the user to go down or up step by step (as it happens in WebDewey) which in our experience would make exploration too long and uncomfortable. Also, displaying an appropriate number of adjacent and subordinated classes is important to make the scope and context of a class more immediately clear [5].

Related classes are displayed in two ways:

1. their notation only is displayed on the right of each class (Fig. 2);
2. both their notation and caption are displayed in the bottom section of the page (Fig. 3).

Fig. 1. SciGator homepage

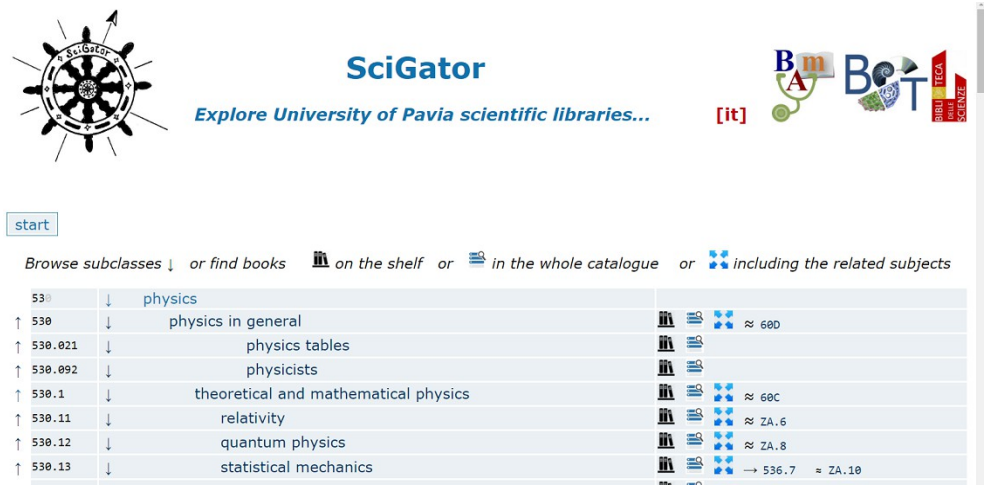

Fig. 2. Display of expanded class `530` in SciGator

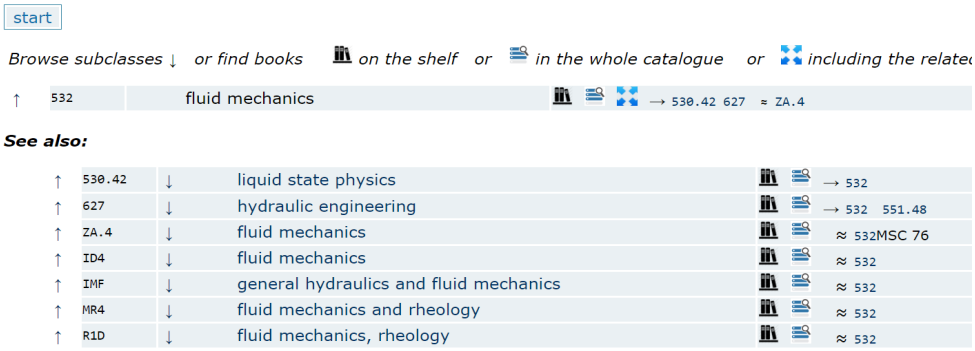

Fig. 3. Display of classes related to `532` in SciGator

Icons on the right of each class launch searches in the university catalogue (Fig. 4), with a default selection of resources owned by at least one of the three scientific libraries. There are three icons which allow for increasingly comprehensive searches. Their function is illustrated in a short help text on the bottom of the same page:

(A) "browse this shelf" (icon with black book spines) retrieves all and only the documents having a shelfmark that begins with the corresponding notation (notice that default truncation is applied, so that its subclasses will also be included. This is a standard application of the hierarchical structure of classification schemes with an expressive notation, like Dewey);

(B) "browse the catalogue" (icon with a lens and a blue list of records) retrieves all the documents having the corresponding notation as their shelfmark *or* as subject metadata. This is useful to cover documents not yet shelved by Dewey, as they are shelved by old local schemes, as well as documents shelved under a different Dewey class though also indexed by the present class;

(C) "expand in the catalogue" (icon with four blue divergent arrows) retrieves all the documents in B plus documents shelved or indexed by related classes, including both associated Dewey classes and equivalent classes in local schemes. This icon is only shown for those classes which actually have some related or equivalent class in the system. Fig. 4 shows results of an expanded search for class `532` fluid mechanics which also include books shelved under `627` hydraulic engineering and under `ZA.4` fluid mechanics at the Mathematics department.

Fig. 4. Results of expanded query for class 532

5 Query Expansion Covering Related Classes

Functionality C is the most innovative feature of SciGator. It leverages the association and equivalence relationships recorded in the MySQL table to provide a wider coverage of subject search, thus increasing recall.

Clearly, every move from A to B or from B to C will produce a greater number of results, potentially also increasing noise, due to the well-known inverse proportionality between recall and precision. This is why the three icons are presented one after the other. Indeed, for those classes where the A-type query already yields a good number of results, say 10 or more, most users will be satisfied with it without need of any greater coverage.

Results of query A can be expected to be very precise, provided shelfmarks have been assigned with enough accuracy.

The "zero-match problem" [6] cannot occur in SciGator because only classes for which some material is actually owned are displayed in the browsable menu. Still, for some classes users may be unsatisfied with such little number of results as 4 or 5, or may need a more complete coverage for their bibliographic purposes. This should lead them to shift to icon B or even to icon C, thus obtaining more results.

On the other hand, results from expanded queries can be less precise and lead to an opposite problem of information overload, especially for 3-digit Dewey classes which correspond to very general concepts or in disciplines where the libraries own many documents. To face these limitations, three strategies have been adopted:

1. the interface is designed in ways encouraging people to use the icons in the sequence A, B, C, and is provided with warnings and explanations about what each of them will produce. Although most users are known to pay little attention to instructions, in time they can acquire more experience with the tool and become aware of how it works;
2. while associative relationships between classes form a complex virtual network, only one arch of it is considered for each node, as recommended by Tudhope et al. [7-8]. In other words, in the case of such relationship chains as between 530.42 liquid state physics and 532 fluid mechanics, and between 532 fluid mechanics and 627 hydraulic engineering, a search for 530.42 will be expanded to only include 532 but not 627. Reciprocity of relationships may also be relevant to these purposes: although relationships between a pair of classes are usually recorded in both directions, in some cases it may be advisable to limit them to only one direction. This is especially the case with equivalent classes in local schemes, which may point to a roughly corresponding Dewey class while the inverse may not be the case, as mentioned above;
3. results are sorted by descending date of publication. Most recent documents are thus displayed first and may provide a quick relevant answer to the user needs, without having to examine the totality of results when their list is very long, thus exceeding futility point [9]. Ideally, sorting could also be based on relevance by listing results of A before results of B, then results of C, in a similar way as with the "double query method" [10]; however this would require a level of technical integration between SciGator and the union catalogue that is not currently available.

6 Future Development

SciGator still is a developing tool, and improvements may be needed in various details concerning both the Web interface and the scripts. Until now, testing has been performed more by librarians than by library users. One advantage of adopting a home-made tool is that it can be continuously tuned through feedback from the front desk to the cataloguing office and the web page developer. Such kind of integration between the different library services is recommended since the times of Ranganathan. The authors of this paper are involved in several of these services at the same time and have frequent first-person exchanges with colleagues, which allows for quick decisions and corrections.

One component that is going to be developed further is the integration of SciGator with signs and instruction at the library shelves. Fabbrizzi [5] recommends that shelves provide users with information about the context of every Dewey class within the whole classification scheme, and that this be connected appropriately with the catalogue. In Pavia, we have provided shelves with some basic illustration of the classification scheme and with signs reporting both notation and the corresponding caption for all classes of reasonable generality. This is now being further improved by providing dynamic links to SciGator itself, through production of QR codes linking to the URL of a specific Dewey class in SciGator, thus suggesting its position in the general scheme of knowledge adopted in the library. Such development may be especially useful as a large proportion of university students nowadays is provided with a smartphone accessing university wifi connection to the Internet.

Having adopted an open web interface with explicit dynamic addresses clearly is an advantage in implementing this, as a QR code can represent a dynamic URL corresponding to a single DDC class. For now, the project does not involve publication of any linked open data. Ideally this would be possible for both bibliographic metadata in the union catalogue and DDC classes; locally-selected relationships between classes, including see-also relationships of various types, could be represented in SKOS or OWL. In practice, however, this is inhibited by current policies and priorities at the level of the union catalogue development, and by copyright restrictions for DDC. The `dewey.info` service that made DDC classes available as linked open data has unfortunately been discontinued by OCLC in June 2015 without any further explanation.

Although precise data on SciGator use are not available, some estimations can be made based on both log files and everyday experience at the library desks. In log files it is unfortunately not easy to identify the exact proportion of sessions coming from automatic crawlers of search engines. Estimates can be done by only considering IP addresses belonging from the local area network of the university, although this misses such genuine remote users as students accessing the system from home. In a three months period, from March to May 2016, we had 260 accesses from IP addresses of the university local network. We estimate that a relevant part of them are by library cataloguers using SciGator as a reference in the process of shelving (another important application of this tool), or staff

using SciGator for quick reference at the front desk. Direct access by library users is rare for now, as both the tool and the methodology of subject search are not very popular among them yet. This situation should be improved by strengthening both information literacy course offer (which our libraries already provide, though focusing on different services) and visibility of the link to SciGator in the context of the big university libraries website, on which we only have partial control.

We believe that our experience can show how KOSs, and classification schemes in particular, have the potential to provide more powerful search tools than is currently the case in most information services, provided there is enough investment it them, in terms of time for indexing, of interface programming, and of service promotion among users.

References

1. Gnoli, C., De Santis, R., Pusterla, L.: Commerce, see also Rhetoric: Cross-Discipline Relationships as Authority Data for Enhanced Retrieval. In: Slavic, A., Cordeiro, M.I. (eds.) Classification & Authority Control: Expanding Resource Discovery, pp. 151-162. Ergon, Würzburg (2015)
2. Casson, E., Fabbrizzi, A., Slavic, A.: Subject Search in Italian OPACs: an Opportunity in Waiting? In: Landry, P. et al. (eds.), Subject Access: Preparing for the Future, pp. 37-50. De Gruyter, Berlin (2011)
3. Si, L.E., O'Brien, A., Probets, S.: Integration of Distributed Terminology Resources to Facilitate Subject Cross-Browsing for Library Portal Systems. Aslib Proc. **62**, 415-427 (2010)
4. Boer, V. de: Connecting Collections across National Borders. http://www.victordeboer.com/digital-humanities/sound-and-vision/connecting-collections-across-national-borders/ (2016)
5. Fabbrizzi, A.: An Atlas of Classification: Signage between Open Shelves, the Web and the Catalogue. JLIS.it **5**, 2, 101-122 (2014), http://leo.cineca.it/index.php/jlis/article/view/10068
6. Tudhope, D., Binding, C.: Faceted Thesauri. Axiomathes, **18**, 2, 211-222 (2008)
7. Tudhope, D., Alani, H., Jones, C.: Augmenting Thesaurus Relationships: Possibilities for Retrieval. J. Digital Info. **1**, 8 (2001), https://journals.tdl.org/jodi/index.php/jodi/article/view/181/160
8. Tudhope, D., Binding, C., Blocks, D., Cunliffe, D.: Query Expansion via Conceptual Distance in Thesaurus Indexed Collections. J. Doc., **62**, 4, 509-533 (2006)
9. Zach, L.: When is "Enough" Enough? Modeling the Information-Seeking and Stopping Behavior of Senior Arts Administrators. J. Am. Soc. Info. Sci. Techn. **56**, 1, 23-35 (2005)
10. Gnoli, C., Cheti, A.: Sorting Documents by Base Theme with Synthetic Classification: the Double Query Method. In: Slavic, A. et al. (eds.), Classification & Visualization: interfaces to knowledge, pp. 225-232. Ergon: Würzburg (2013)