Indexing KOSs in BARTOC by a disciplinary and a phenomenon-based classification: preliminary considerations

Andreas Ledl
Basel University Library, Basel, Switzerland

Claudio Gnoli
Science and Technology Library, University of Pavia, Italy

Abstract: The paper discusses the Basel Register of Thesauri, Ontologies & Classifications (BARTOC) and its suitability as a tool for testing knowledge organization systems (KOS), and in particular how two different classification schemes perform when applied to the same items. It examines the recently launched project on using Integrative Levels Classification (ILC) for classification of top-ranked KOSs in BARTOC. The knowledge organization accomplished with ILC is compared to that produced by the application of Dewey Decimal Classification (DDC). This represents a case study for evaluating phenomenon-based classification in comparison to a disciplinary classification. The comparative study also contrasts a faceted classification (ILC) with an enumerative scheme (DDC). Some technical aspects, such as importing ILC into Drupal CMS and creating URIs for terms to use them as Linked Open Data, are addressed exactly like some intellectual aspects of this subject indexing endeavour.

Keywords: BARTOC, Integrative Levels Classification; Dewey Decimal Classification; knowledge organization systems; terminology registry; subject indexing

1. Introduction

The Basel Register of Thesauri, Ontologies & Classifications (BARTOC[1]) is a terminology registry for knowledge organization systems (KOS) and a directory of terminology registries. While it used to be a *basic* terminology registry, containing only metadata of KOSs, the work on transforming it to a *full* terminology registry, 'including also the members (e.g. concepts, terms, relationships) of the vocabularies', is well advanced (Golub et al., 2014: 1903). By integrating PoolParty Semantic Suite[2] it has recently become possible to browse the content of first 'skosified[3]' vocabularies. Moreover, when creating a SKOS (Simple Knowledge Organization System) vocabulary, it can directly be published in BARTOC and be shared with the community according to semantic web standards.

1 Bartoc can be accessed at https://bartoc.org.
2 PoolParty Suite can be accessed at https://www.poolparty.biz/.
3 Vocabularies presented using SKOS standard (XML/RDF format). https://www.w3.org/2004/02/skos/.

BARTOC's main goals are, firstly, to describe KOSs in form and content, and secondly to provide access to these KOSs (Ledl & Voss, 2016). Therefore, to specify the subject coverage of KOSs, it is itself utilising controlled vocabularies, amongst others, EuroVoc, the multilingual thesaurus of the European Union, and the Dewey Decimal Classification (DDC).

Although EuroVoc is mainly intended to cover the European parliamentary activities, it was selected because of its broad range of 21 domains. Moreover, it is regularly updated, maintained by a trustworthy institution and is available in 25 languages, which is essential for BARTOC multilingual search.

At the start of the project in 2013, DDC, the most widely used library classification system in the world, seemed like a natural choice, as BARTOC wanted to address an international audience. This huge reputation qualified DDC to make the search interface more easily accessible to wide-ranging groups of users, especially because of the captions available in various languages. DDC also gives a good overview about the different fields of BARTOC content. For reasons of clarity and usability as well as of public availability of data, only the first three levels are used (e.g. *300* Social sciences, *370* Education, *378* Higher education).

Since BARTOC is running on the content management system Drupal, KOSs can be uploaded in different formats (e.g. CSV, JSON, SKOS, XML) to manage and use them for tagging and browsing, respectively. Moreover, BARTOC content is not only aimed at viewing by humans but is also machine readable (RDFa). This means that in addition to the metadata of the classification, each class can be displayed in RDFa and be assigned with RDF predicates. By its unique identifier[4], data can be distilled and expressed in Turtle, RDF/XML, JSON-LD or N Triples later on.[5]

All this provides the opportunity of applying BARTOC as a playground or laboratory to test, e.g. universal classifications, because its content covers the whole of knowledge, has international relevance, but at the same time contains a manageable number of vocabularies (around 2,700 at the moment).

2. DDC vs. Integrative Levels Classification (ILC)

DDC was first published in 1876 and has been updated regularly until its current 23rd edition. It is a typical representative of traditional classifications, based on a hierarchical tree structure of disciplines in decimal notation. Its subdivision of knowledge is ultimately inspired by Francis Bacon's tripartition into the disciplines of memory (history), those of imagination (arts and literature) and those of reason (philosophy and the sciences). These can still be

4 BARTOC URI: http://bartoc.org/ILC/1/{notation-caption}; http://bartoc.org/DDC/23/{notation-caption}).

5 https://www.w3.org/2012/pyRdfa/Overview.html#distill_by_uri

identified in the reversed order of main classes, which are expanded to 9 plus a class for generality in order to match the notational base of Arabic numerals. The nine main disciplines have been chosen to reflect the state of knowledge in the late 19th century United States of America which explains, somewhat unusual organization of sciences on the top level of the classification. For instance, the subsumption of psychology under philosophy and the prevalence of Christianity over other faiths in the religion class, or those of European languages over other languages in classes of linguistics and literature. This has not prevented its continuous updating by successive committees of expert editors to reflect more recent and international developments in published knowledge. Thus, class captions are being reformulated and integrated to reflect more recent terminology, and new concepts are being introduced as additional subdivisions. Still, the overall structure of the classification remains basically the same, which has consequences especially on the resulting sorting of classified items (e.g. in displaying psychology documents amidst philosophy documents). Also, concepts, such as e.g. 'Europe', get repeated with different notations when they get structurally subordinated in different disciplines (Broughton, 2004: 18; Slavic, 2007). Thus, *914* only means 'geography of Europe', which is different and far away from *325.34* 'European colonization', *327.4* 'international relations of Europe', *509.4* 'European science', *709.4* 'European arts', *940* 'history of Europe', etc. These, so called, distributed relatives, managed in this way are a well-known feature in enumerative disciplinary classifications, of which DDC is an example.

Dispersion of concepts under many different notations can be mitigated by a faceted classification structure, in which a concept is assigned a stable notation that can then be combined with other concepts by syntactical devices. However, in DDC this technique is applied to a very limited extent only, making it a basically *enumerative* rather than a *faceted* classification. DDC Common Subdivisions, such as *-094* 'Europe', have recently been described as 'facets' but are such only in a very broad sense (Gnoli, 2017) as they can only be appended at the end of classes following instructions in the schedules, rather than freely combined in any position. Real facet analysis has been implemented in DDC only in a small number of revised classes, such as *780* Music.

On the other hand, the Integrative Levels Classification (ILC) is an experimental innovative classification that has a structure similar to classical faceted classifications, but with the important difference: classes represent phenomena and their subdivisions instead of disciplines and subdisciplines. Phenomena are listed in a series of 26 knowledge levels with an increasing organizational degree (Kleineberg, 2017): from abstract forms through particles, atoms, molecules and organisms, to minds, civil society, economies and cultures. For example, the concept of 'European Union' in ILC is listed under the main class *t* 'governments' which has no particular disciplinary implication. As a result, it can be combined freely with any other concept by means of such free facet relationships as 'having quality', 'having part' or 'affected by', to give compounds

where EU specifies or is specified by something, is part of something or has some part, is affected by or affects something, etc. (e.g. 'EU, having part UK'). Another way of combining ILC classes is by simply listing them as a set of themes without specifying the kinds of the relationships between them (e.g. 'EU; UK'); this is the syntax that will be adopted to index BARTOC items, for purposes of simplicity both informational and computational.

Distributed relatives (far away compounds including the same concept 'EU') will always be present, but will only depend on whether the concept of, e.g. 'European Union' is the base theme in the combination, such as in 'EU; economic crisis' that will be listed adjacent to other items having EU as their base theme, or just a particular theme specifying it, such as 'law; EU' that will be listed adjacent to other items about law, though still retrieved in a search for the concept 'EU'. On the other hand, distribution of concepts will not depend anymore on the disciplinary organization of the main classes to which a concept has to belong, which should allow more natural grouping according to the phenomena they represent.

Phenomenon-based classification is an alternative approach to classification explored by various authors, especially in recent years (Gnoli, 2016). Although its possible merits and problems have been discussed in literature on a theoretical plane, few data are available allowing for a direct comparison of performance between phenomena and disciplinary classifications. One preliminary experience in an academic library with a limited sample of books in nature conservation classified by both DDC and ILC is described in Szostak et al. (2016: 104-106).

KOSs in BARTOC, that are already indexed by DDC (and EuroVoc), provide a case for evaluating phenomenon-based classification and comparing it to disciplinary classification. The resulting organization of knowledge is produced by the two systems applied to the same items and, therefore, allows a more accurate analysis. Based on our description above, DDC and ILC should mainly differ in two features:

(1) DDC is disciplinary while ILC is phenomenon-based
(2) DDC is enumerative while ILC is faceted

In BARTOC, difference 2 is partially neutralized by the practice of assigning several DDC classes to the same item, thus producing a sort of free classification similar to what has been described for ILC, despite the potentially greater capabilities for concept combination that are available in ILC. This puts the two systems on the same plane as for difference 2, allowing to focus the comparison on difference 1, that is disciplinary vs. phenomenon-based.

Further in this paper we will address both technical and intellectual aspects of this project. Depending on our findings, this could be the starting point for extending the study on other universal classifications, e.g. UDC.

3. Applying ILC in BARTOC

ILC is maintained as a MySQL database including information on notation, English captions, synonyms, corresponding disciplines, facets, their foci, sources of foci, semantic factoring, etc. (Gnoli et al., 2011). Only the basic database fields for notation, captions and synonyms have been used for this project, in view of their application for searching and browsing in BARTOC. The relevant fields have been exported in a CSV file for later import into BARTOC. For the sake of notational stability, we adopted edition 1 of ILC, although in edition 2 currently under development various changes are introduced in the scheme.

In a first stage, only ILC data of basic classes have been kept, while records of facets and their foci have been left out, because complex faceted notation was not expected to be needed for indexing the broad topics covered by the items listed in BARTOC. The resulting dataset consisted of 6,456 basic class records with their captions and synonyms.

However, it soon became obvious that some general concepts needed to index the BARTOC KOSs were only available in ILC in the form of facets. For example, 'diseases' can be expressed in ILC only as a facet of organisms, and 'law' can be expressed only as a facet of governments. This problem was addressed by entering these individual facets in the table manually.

ILC, which had been divided into several parts of less than 1,000 items each, was imported in Drupal as flat lists in CSV format with the help of a Drupal module called Taxonomy CSV import/export. To model the hierarchical structure between classes, manual work had to be done. By defining parent term relationships and term weight in the Taxonomy module, the original order could be restored and ILC was ready to use.

Indexing by ILC has started with the 'top-rated' KOSs, that is the KOSs that have received most votes by users in the BARTOC rating system. This choice was aimed at quickly obtaining a set of relevant KOSs indexed by both DDC and ILC. It has to be taken into account, that with two different indexers for DDC (Ledl) and ILC (Gnoli), the human factor plays a certain role. Besides subjective view, the fact that DDC tags had already been there when ILC tags were added could have influenced the indexing process.

As mentioned in the previous section, ILC classes are used in free combinations, such as 'industry: artefacts' or 'schools: Europe', without the relation syntax being specified. A similar solution was already applied in BARTOC for DDC, as more than one DDC class can be assigned to the same KOS.

Reflecting both the experimental character and the importance of the project, ILC, in contrast to EuroVoc and DDC (which appear as exposed filters on the homepage), has its own Tab and View (https://bartoc.org/en/ilc) in BARTOC.

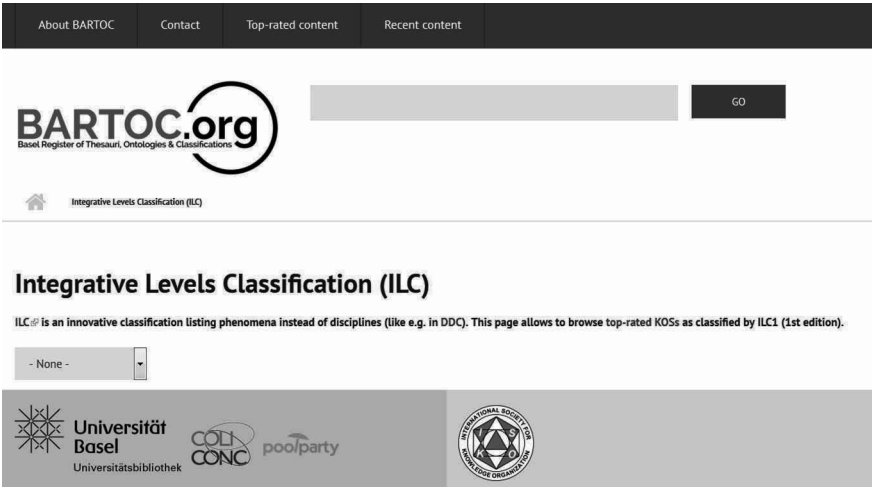

Figure 1: *ILC View in BARTOC*

As shown in Figure 1, this is used to explain, briefly, what ILC and the page is about, and to provide browse functionality for searching BARTOC vocabularies along the ILC phenomena tree by means of the Simple Hierarchical Select module.

Figure 2 on next page illustrates how BARTOC content can be specified according to ILC tags, since they appear, when available, as a navigation facet next to keyword search results to further refine the list of hits.

4. Preliminary results

Upon the completion of the classification of the first set of BARTOC KOSs, a preliminary analysis, concerning the resulting intellectual organization of items, was possible. Combinations of classes seem to be adequate to express the subject matters covered by KOSs both with DDC and with ILC. The degree of accuracy (co-extension) obtained by the combined classmarks looks similar in the two cases. In evaluating this, one has to take into account that ILC is still a developing system, that has not yet reached the degree of refinement and hierarchy depth already available in DDC; however, as the indexed subjects are usually quite broad, and ILC covers the whole spectrum of knowledge, good approximations can be obtained in most cases. In principle, more detail can be obtained by developing subclasses where needed following the existing principles of ILC structure.

Clearly, what is different is the meaning of classes, as these express disciplines and subdisciplines in the case of DDC, but phenomena and their types or facets in the case of ILC. Users can perceive this especially in the display of scheme trees for browsing: while with DDC one has to start with 10

Figure 2: *ILC navigation facet*

disciplinary macro-classes, such as philosophy, religion, social sciences etc., with ILC one is presented with 26 classes of phenomena sorted by integrative levels, such as molecules, rocks, cells, organisms, populations, civil society, etc. One can expect the effect to be cognitive, as in the latter case the classification scheme will guide users to the exploration of the universe of subjects by classes of phenomena rather than those of disciplines.

The two schemes may also group items in different ways, which affects the set of KOSs users find under an individual class. Indeed, KOSs dealing with connected phenomena though approached by different disciplines will be grouped by ILC but not by DDC; while the opposite will happen with disciplinary approaches. An evaluation of this aspect of system performance will require a more detailed analysis of a set of, at least, several hundreds of KOSs classified by both systems, allowing for both quantitative measures and the identification of individual meaningful examples. Such work is planned during the following stages of the project.

Some examples of KOSs classified with both DDC and ILC follow:

Events Name Authority List
DDC: *328* 'the legislative process'
ILC: *t6 tUE* 'law; EU'

Treaties Name Authority List
DDC: *341* 'law of nations'
ILC: *tUE u t6* 'EU; economies; law'

Thesaurus Europäischer Bildungssysteme
DDC: *379 940* 'public policy issues in education; history of Europe'
ILC: *se tUE* 'schools; EU'

CERL Thesaurus
DDC: *094 911 940* 'printed books; historical geography; history of Europe'
ILC: *js w r55n yu* 'landforms; settlements; noun; humanities'

5. Discussion

Our experience shows that BARTOC, thanks to its coverage of a wide corpus of items dealing with potentially any field of knowledge, is a suitable platform to test the application of different kinds of classification schemes. It produces experimental data that can be useful for the advancement of classification research, including facet analysis and comparison of disciplinary vs. phenomenon-based approach.

In particular, phenomenon-based classification seems to produce a sorting of entries that is significantly different from those of traditional disciplinary classification, thus leading users to approaching a corpus of knowledge resources in a way less influenced by traditional disciplinary thought. As for the full expressive potential of a freely faceted phenomenon-based classification, in order to be evaluated and compared to that of an enumerative disciplinary one, one would need indexing of more specific subjects with a full specification of faceted relationships.

More substantive and accurate data are expected from the development of the current project. This can also involve evaluation of the introduction of ILC edition 2, with more developed schedules for some classes and reuse of concepts as free facets made even more explicit on the notational level; or of other systems such as UDC, which given its strong analytico-synthetic power, although still based on disciplines, can be considered as an intermediate between DDC and ILC.

References

Broughton, V. (2004). Essential classification. London: Facet.

Gnoli, C. (2016). Classifying phenomena. Part 1: dimensions. *Knowledge organization*, 43 (6), pp. 403-415.

Gnoli, C. (2017). Syntax of facets and sources of foci: a review of alternatives. In: *Faceted classification today: theory, technology and end users: proceedings of the International UDC Seminar 2017, London (UK), 14-15 September 2017.* Edited by A. Slavic, C. Gnoli. Würzburg: Ergon Verlag, pp. 243-256.

Gnoli, C. et al. (2011). Representing the structural elements of a freely faceted classification. In: *Classification and ontology: formal approaches and access to knowledge: proceedings of the International UDC Seminar, 19-20 September 2011, The Hague*. Edited by A. Slavic, E. Civallero. Würzburg: Ergon Verlag, pp. 193-206.

Golub, K. et al. (2014). Terminology registries for knowledge organization systems: functionality, use, and attributes. *Journal of the Association for Information Science and Technology*, 65 (9), pp. 1901-1916.

Kleineberg, M. (2017). Integrative levels. In: *ISKO Encyclopedia of Knowledge Organization*. Edited by B. Hjørland. Available at: http://www.isko.org/cyclo/integrative_levels.

Ledl, A.; Voss, J. (2016). Describing knowledge organization systems in BARTOC and JSKOS. In: *Proceedings of TKE 2016 - 12th International Conference on Terminology and Knowledge Engineering, June 2016, Copenhagen (Denmark), 22nd – 24th June 2016*. Edited by E. Thomsen et al. Copenhagen: Copenhagen Business School, pp. 168-178. Also available at: http://hdl.handle.net/10760/29366.

Slavic, A. (2007). On the nature and typology of documentary classifications and their use in a networked environment. *El profesional de la informacion*, 16 (6), pp. 580-589. Also available at: http://hdl.handle.net/10150/106049.

Szostak, R.; Gnoli, C.; Lopez-Huertas, M. J. (2016). Interdisciplinary knowledge organization. Cham: Springer.

About the authors

ANDREAS LEDL is subject librarian for psychology, philosophy and educational studies at the Basel University Library. He holds a graduate degree in educational science from University of Regensburg, a master degree in library and information science from Humboldt University of Berlin and a PhD from University of Flensburg. His research interests include knowledge organization, information literacy, subject indexing, open data, open access, and the Semantic Web. In recent years, he has focused on developing bibliographical databases / search interfaces with the content management system Drupal. He is the initiator, technical head and manager of the Basel Register of Thesauri, Ontologies & Classifications (BARTOC) and also founder and co-editor of the open access publication '027.7 Journal for Library Culture'.

CLAUDIO GNOLI has been an academic librarian since 1994, currently working at the Science and Technology Library, University of Pavia, Italy. He has taught various courses and lessons on classification and knowledge organization, including a recent invited lecture at the KO Research Group, University of Wisconsin Milwaukee. He is a specialist in classification and facet analysis on both the theoretical plane and its application to the development of classification systems based on disciplines (UDC editorial board) or based on phenomena (Integrative Levels Classification research project). He is co-author of Interdisciplinary Knowledge Organization (Springer, 2016); a chapter author in K. Golub's *Subject Access to Information* (Libraries Unlimited, 2014) and of numerous papers in academic journals; and web editor of the *ISKO Encyclopedia of Knowledge Organization*.