Repositories of knowledge in digital spaces: accessibility, sustainability, semantic interoperability:
11th German ISKO conference: Konstanz, 20-22 February 2008

Freely faceted classification for a Web-based bibliographic archive

The BioAcoustic Reference Database

Claudio Gnoli, Gabriele Merli*, Gianni Pavan**, Elisabetta Bernuzzi**, Marco Priano***

* *University of Pavia. Department of Mathematics, via Ferrata 1, I-27100 Pavia, Italy, <gnoli@aib.it>*
** *University of Pavia. Interdisciplinary Center for Bioacoustics and Environmental Research, via Taramelli 24, I-27100 Pavia, Italy*

Abstract

The Integrative Level Classification (ILC) research project is experimenting with a knowledge organization system based on phenomena rather than disciplines. Each phenomenon has a constant notation, which can be combined with that of any other phenomenon in a freely faceted structure. Citation order can express differential focality of the facets. Very specific subjects can have long classmarks, although their complexity is reduced by various devices. Freely faceted classification is being tested by indexing a corpus of about 3300 papers in the interdisciplinary domain of bioacoustics. The subjects of these papers often include phenomena from a wide variety of integrative levels (mechanical waves, animals, behaviour, vessels, fishing, law, ...) as well as information about the methods of study, as predicted in the León Manifesto. The archive is recorded in a MySQL database, and can be fed and searched through PHP Web interfaces. Indexer's work is made easier by mechanisms that suggest possible classes on the basis of matching title words with terms in the ILC schedules, and synthesize automatically the verbal caption corresponding to the classmark being edited. Users can search the archive by selecting and combining values in each facet. Search refinement should be improved, especially for the cases where no record, or too many records, match the faceted query. However, experience is being gained progressively, showing that freely faceted classification by phenomena, theories, and methods is feasible and working.

1: The ILC project

The Integrative Level Classification (ILC) research project [ISKO Italia 2004] aims at building and testing a knowledge organization system based on an innovative approach.

ILC schedules, though being structured in a way similar to classical bibliographic classifications like Dewey, UDC and Bliss, do not take as main classes canonical disciplines, such as chemistry, zoology, or sociology, but directly phenomena of the real world, such as molecules, animals, or societies.[1]

This agrees with the theses advocated in the León Manifesto, published after the last ISKO

1 During the conference and in subsequent personal discussion, Ingetraut Dahlberg commented that her Information Coding Classification (ICC) also abandons disciplines, as it takes as main classes ten general object areas arranged by levels [Dahlberg 1982]. Each of these areas is then divided by aspect categories. Thus, ICC seems to be in a middle position between the classifications by disciplines like Dewey and Bliss, and the classifications by phenomena advocated by Foskett, Austin, Beghtol, Gnoli, Szostak, and others. Phenomena schedules have not been developed in ICC yet, but they would fall under the Object category 2: e.g. 26 "chemistry" has a facet 262 for the objects of chemistry, under which all chemical substances could potentially be listed.

Spain conference [ISKO Italia 2007]. The Manifesto claims that, in order to serve the interdisciplinary needs of contemporary research, knowledge organization systems should give priority to phenomena, and allow indexers to express by separate notation the additional dimensions of theories and methods under which the phenomena are studied [Szostak 2007; 2008].

ILC general schedule follows a sequence of increasing integrative levels, starting with the most abstract and fundamental phenomena like structures, particles and atoms, going on through material, organic, and mental phenomena, until social and cultural phenomena. Lower level phenomena are expressed by letters of low ordinal value (`d` "particles", `f` "molecules", `i` "rocks"), while more complex and evolved phenomena have higher ordinal values (`u` "economies", `w` "artifacts", `x` "art works").

2: Characteristics of freely faceted classification

The ILC scheme can be used as a freely faceted classification [Austin 1976; Gnoli & Hong 2006], in which notation for each phenomenon can be freely combined with any other phenomenon to synthesize compound classmarks. For example, `wni` "vessels" and `mqvtn` "whales" can be combined by facet indicator `60` to yield `wni60mqvtn` "vessels damaged by whales". As each concept has a constant notation, by searching this in a digital archive one can retrieve all classmarks where that concept is combined with any other: searching for `mqvtn` "whales" will yield

`mqvtn2a`	whales in Atlantic ocean
`t8mqvtn`	institutions dealing with whales
`wa4mqvtn`	food consisting of whales
`wni60mqvtn`	vessels damaged by whales
`xg8mqvtn`	paintings of whales

Thus, ILC facets work more like role operators, or phase relationships, than like the facets of disciplines in classical bibliographic classifications [Gnoli 2007*a*]. They can express standard classification categories like Property, Agent, Part, and Process, as well as epistemic dimensions of subjects like Theories, Methods, Viewpoints, Applications, and Disciplines. The latter facets are of the kind advocated in the León Manifesto. In particular, the representation of Theories in ILC is discussed by Szostak & Gnoli [2008], while that of Methods is addressed within the present paper.

Like in the standard theory of faceted classification, facets should be written in a prescribed citation order, starting with those of higher ordinal value: "food" `wa` "in Japan" `29q` "consisting of whales" `4mqvtn` will be expressed by `wa4mqvtn29q`, because `4` has a higher ordinal value than

29. However, in case a facet is more focal in the document being indexed, it can be promoted to a leading position: a paper focusing on Japanese food, and dealing only secondarily with the fact that it consists of whales, can be classified as wa29q4mqvtn. Expressing focality by the sequence of facets aims at producing more helpful sequences in browsable lists; indeed, documents sharing the focal facets will be grouped together, while the other facets will only serve as secondary specifications. For each document, the indexer should identify a set of primary facets, a set of secondary facets, and so on, then cite facets in the standard order only within each set.

Such sophisticated kind of knowledge organization system has the potential to produce very accurate indexes. On the other hand, learning and mastering it can result in a non-trivial task for indexers. Just because the system allows for more freedom, it can also be more demanding, as the indexer has to interpret and translate the subject into the appropriate compound notation, rather than just identify a ready classmark in an enumerative list. In order to make this work easier, some usable interface should be made available [section 4].

Another possible problem is that users are confused by very complex notations, like those designed to express syntactical relationships between concepts for retrieval purposes. Indeed, users are known to prefer shorter and simpler notations, e.g. made of all digits (like in Dewey) or all letters (like in Bliss), which can be easily remembered and copied. This is relevant mainly for library shelfmarks, as in the digital environment both search and display can be done through verbal captions. For these reasons, Austin [1976] concluded that freely faceted classification best suits machine retrieval, while shelfmarks in libraries would be better served by the so-called mark-and-park systems. However, accepting this would be a pity, as one should use two different systems for shelving and for retrieving documents, and this would work against the integration and interoperability of knowledge resources.

To cope with these problems, ILC notation is being progressively refined, and now consists only of letters, digits, and brackets. The alternation between letters and digits can help in visually scanning a long classmark, in the same way as a blank space is used every three characters in Dewey and Bliss (though in digital systems the blank space is usually removed). Furthermore, the overall lenght of classmarks is reduced by a mechanism called *extra-defined foci*: in the default values of a facet, some letters are not written (but can be retrieved from the database) as they are obvious; e.g., "signals by whales" is written q9vtn instead of q9mqvtn, as signals q are always produced by some animal or human being[2] mq.

A property of freely faceted classification, that has been noticed while testing it, is that items having more facets are more likely to be retrieved. This happens because they can be included in the results of search by one or by another of their facets, while simple concepts are retrieved only when

2 Other meanings of *signal*, like in cells or information systems, are represented by different classes.

searched directly as a whole. This means that specialized documents, which usually have more facets, tend to be retrieved more often than general documents; which is paradoxical, as general documents could contain answers to a larger range of information needs. This, more generally, happens with any subject indexing system where elements can be retrieved separately, including keywords, folksonomies, synthetic subject heading systems, and synthetic classifications with an expressive notation. To avoid biases in retrieval, an indexing policy should be defined accurately, so that documents with the same degree of specificity be indexed with a proportional number of facets.

On the other hand, when search is performed by selecting options from a list of available values for each facet, it can easily happen that no document corresponds exactly to that particular combination of facets. Such "zero match problem" [Tudhope & Binding 2008] could be solved by allowing for more fuzzy retrieval, where results would include not only the documents matching the search combination exactly, but also those differing from it by one facet or another, or by one hierarchical degree in the same facet.

3: The BioAcoustic Reference Database

The ILC scheme, currently consisting of some 5000 classes, is still a draft being continually developed and corrected. This happens especially through its application to sample bibliographies of special domains, for which hierarchies are developed more deeply according to the needs. The earliest of these bibliographies dealt with the geography and ethnography of a mountainous area in Northern Italy [Gnoli & Merli 2005], and with faceted classification itself [Gnoli & Hong 2006].

A more substantial and relevant corpus of documents is now being classified deeply with ILC. This consists of a collection of about 3300 bibliographical references in the domain of bioacoustics, that is the study of acoustic signals uttered by animals through both recording them in the field and processing the records in a laboratory. Papers published after year 2000 mostly deal with whale and dolphin sounds, while for the previous period classical bioacoustic works on other animals are also included.

This seems to be suitable domain to test the theses of the León Manifesto, in that it is typically interdisciplinary, including concepts not only biological, but also physical (acoustics), ethological (associated behaviour), technological (microphones, signal analysis devices), political (fishery management, conservation acts), military (impact of submarines and of explosions), social (impact of and on tourism), etc. Animal sounds are studied sometimes as the main object of research themselves, sometimes only as a mean to detect indirectly the presence of a population of

cetaceans, or to compare them with sounds uttered by non-living sources. Indexing such a domain by a disciplinary KOS, like Dewey or UDC, would have some limitations, like being forced to decide whether a paper about the impact of whale conservation acts on tourism in rural Scotland should be filed under ecology, or economics, or law [Gnoli 2007*a*].

The bibliographic archive has been christened the BioAcoustic Reference Database; its acronym *BARD* remembers of a Celtic musician, thus evoking something acoustic, and is depicted in the icon of a dolphin playing a harp. The archive is fed by the staff of the Interdisciplinary Center for Bioacoustics and Environmental Research (CIBRA) at the University of Pavia, as they collect printed or digital copies of scientific papers in the domain. As no publicly accessible library yet exists at CIBRA, BARD is not a library catalogue, but only a bibliographic service providing information about the existence of papers and the fact that they are owned and used at CIBRA.

In 2006 we moved the archive from a local MS Access file to a MySQL database hosted at the Mathematics Department of the same university, making it accessible for free on the Internet through a PHP interface within the ILC project website <http://www.iskoi.org/ilc/bard/>. Although not all papers are fully classified yet, the archive can already be searched, also to demonstrate the use of freely faceted classification to anyone interested.

4: The indexer interface

Although most papers in the database come from the original local version of the archive, more of them can now be input through a Web interface, only accessible to the CIBRA staff. At the same time, the ILC staff works on indexing papers already in the database. This can be done by a separate indexing interface, showing in one page the title, the authors, and other bibliographical data of a single paper, and providing an input box to type the corresponding classmark [figure 1].

In order to build the classmark, the indexer can quickly search the classification schedule for all classes having a required word in their caption field, or in their synonyms field, or in their class description field, or in their related discipline field (e.g. the word "ornithology" can lead to the class `mqvto` "birds", as a link between these concepts is recorded in the discipline field). After examining them, the indexer can navigate the scheme to look for a broader, a narrower, or a related class, and eventually choose one to become part of the classmark.

At his request, he can also be helped by automatically displaying a set of suggested classes. This set is generated by matching the words forming the title of the document with the content of the fields just mentioned. Matching classes are then suggested to the indexer as possible components of the subject: each of them can be cut-and-pasted to become part of the classmark.

Titles of the papers indexed in BARD tend to be long, and each of them can occur in many ILC captions or related fields: as a result, a large number of unrelated classes can be displayed. Matching has then to be refined in order to improve precision. A first rough measure – ignoring all words shorter than 4 letters as well as the word "*with*" – is already implemented, while others can be studied.

As the indexer has produced a draft classmark, he can look at the resulting verbal equivalent, in order to check whether the notation he has entered has the intended meaning and is complete. The verbal equivalent is produced automatically by a PHP script, which parses the classmark into its phases and facets, takes the caption for each from the ILC main schedule, and connects the captions with prepositions corresponding to the appropriate facets. For example, a paper entitled "Seasonal and diurnal trends of chorusing humpback whales wintering in waters off western Maui" [Au et al. 2000] gets the classmark `q9vtncmh5h2p22pdzh15Y11Y`, which is translated by the system into "signals, by humpback whale, acoustic, in Pacific Ocean, near Hawaii, in some season, at some day time". When seeing this subject statement, the indexer can either be satisfied and move to the next paper, or edit it again and look at the modified verbal equivalent. The synthesized subject statements will also be displayed in the user interface.

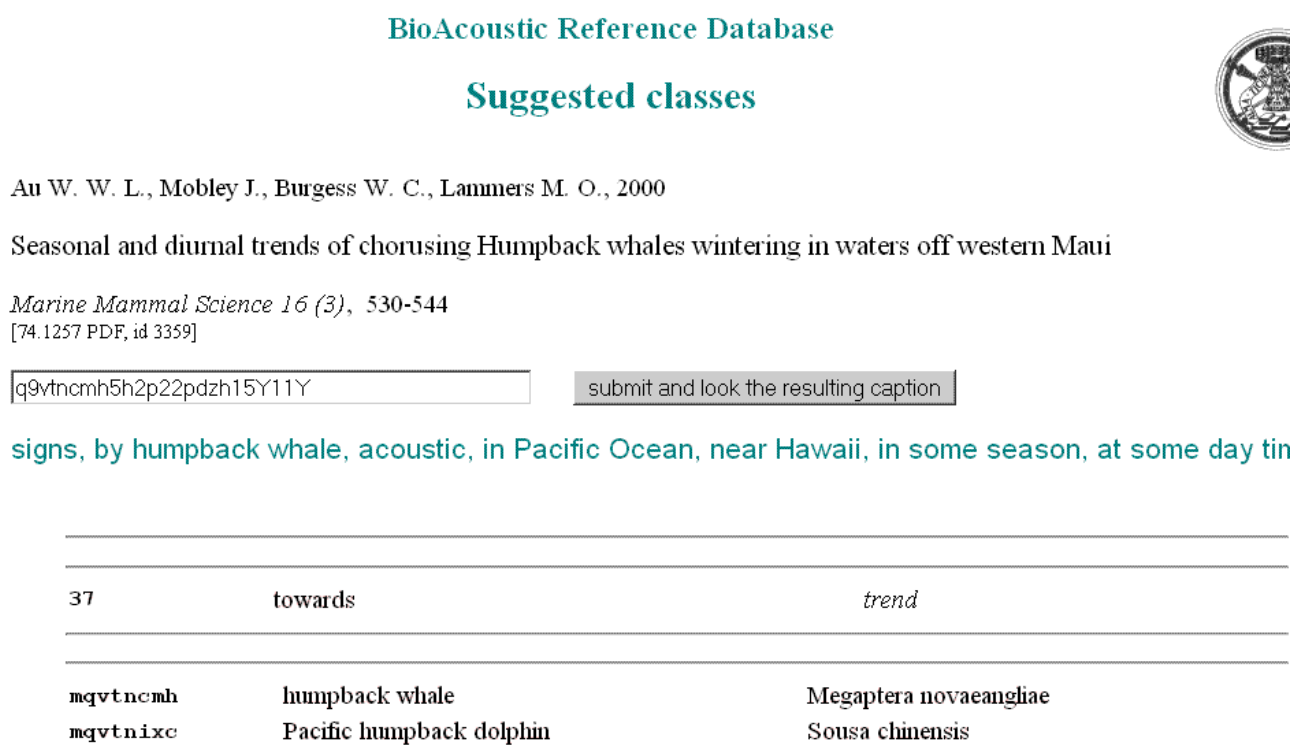

Figure 1: The indexer interface

Despite these facilities, the usability of the indexer interface could still be improved, by enabling click-and-select or drag-and-drop features for the suggested classmarks, and by providing

automatic generation of the default citation order for facets, which could then be edited by the indexer to express focality.

As mentioned in section 1, the León Manifesto advocates for classification by phenomena, theories, and methods: a book about evolutionary field ornithology could be indexed as "birds *[phenomenon]*, according to Darwinism *[theory]*, studied by observation *[method]*". While theories often play a crucial role in determining subjects in the human sciences [Szostak & Gnoli 2008], in the domain of bioacoustics we found that this is much less the case: indeed, almost the totality of papers share the standard theoretical approach of contemporary biology, being comparative and evolutionary. On the other hand, we found the facet of method in a significant fraction of papers. This is expressed in ILC notation by the facet indicator `03`, and cited after both the phenomenon class with all its facets, and the theory facet (if present). Values in the method facet are taken mainly from Szostak's [2007, section 3.1] list of basic methods and their subclasses.

In this way, all papers applying a particular method of study can be searched, independently from the phenomenon to which the method is applied. By selecting `03fn` "sound recording", one gets the following class headings:

`o9vtnpnp5j03fn` "instincts, of finless porpoises, echolocation, as studied by sound recording"
`q9vo5h03fn` "signals, by birds, acoustic, as studied by sound recording"
`q9vtni5h18os03fv03fn` "signals, by delphinidae, acoustic, during behavioural sequences, as studied by filming, as studied by sound recording"

As shown by these examples, classmarks of specialized scientific papers can result quite long, as well as very informative. In some cases this can also produce an undesirable complexity. A document included in the archive and entitled "Guidelines on the applications of the environment protection and biodiversity conservation act to interactions between offshore operations and larger cetaceans" could get a multi-nested classmark such as `tn8ve(4qvtn(902o68v(3)25c))4d`. Although formally correct, this classmark is hard to be interpreted both by human users and by the computer. In theory, it could be managed by syntactical rules translated into a parsing algorithm; but this would require to write complex scripts, and to wait some seconds for information processing while interacting remotely with the database. To avoid such nesting problems, classmarks can be cut into more manageable chunks, like `tn8V4d ve4V mqvtn902o68v(3)25c`, where the blank spaces represent phase relationships, and `V` represents a special deictic class meaning "the following".

Our experience suggests that this kind of system can be used in free or faceted versions

according to the needs. For example, one can use fully faceted classmarks in indexing specialized literature, while listing only the most relevant phases, and omit facet indicators, in indexing simpler items, like often found in websites or in generalistic document collections [Gnoli 2007*b*].

5: The user interface

To retrieve references from the archive, a faceted search interface is provided [figure 2]. Besides the classical search boxes for authors, title, date, and journal (planned), users are offered with a subject search menu divided into facets. For each of the available facets (zoological taxon, function, kind of signal, related topic, region, and tool/method), values can be selected from a list of the most used foci, using a radio button (drop-down menu has also been considered). The user is thus able to specify a particular combination of facets she is interested in. The default value for each facet is set to "anyone", in order to limit the zero match problem discussed in section 2. Author, title, or date values can also be specified in combination with the subject facets.

Relevant items are extracted by matching the appropriate fragment of notation with the classmark field in the database. Default truncation is applied, as in most cases the searched notation will appear only as one facet in a compound classmark. Truncation also makes the subclasses of the searched class included: by searching for "cetaceans" one also obtains more specific papers on "dolphins".

Resulting records are presented sorted by classmark. This produces a meaningful order which is helpful for browsing, especially when items are many. Both classmarks and the equivalent captions are shown in blue-green colour in the first line of each item, followed by the descriptive information in black. The captions are synthesized automatically by the same PHP script described above for the indexer interface. These result into quasi-natural phrases easy to be interpreted, except for a minority of cases for which the script must be still improved.

Records extracted in this way can amount to between zero and several hundreds. Indeed some classes, like "cetaceans" in the zoological taxon facet or "Pacific" in the region facet, occur very frequently, and are not enough to select a manageable number of items. Possible ways to face this problem include providing a more detailed list of cetacean species by a drop-down menu in the search interface, and automatically offering ways to refine search in case the count of results is greater than a certain *futility point*, presumably around 20. Refinement could be done by specifying more facets, or by choosing a more specific subclass within a facet. Some of these things have already been realized for one of the former test bibliographies [Gnoli & Merli 2005].

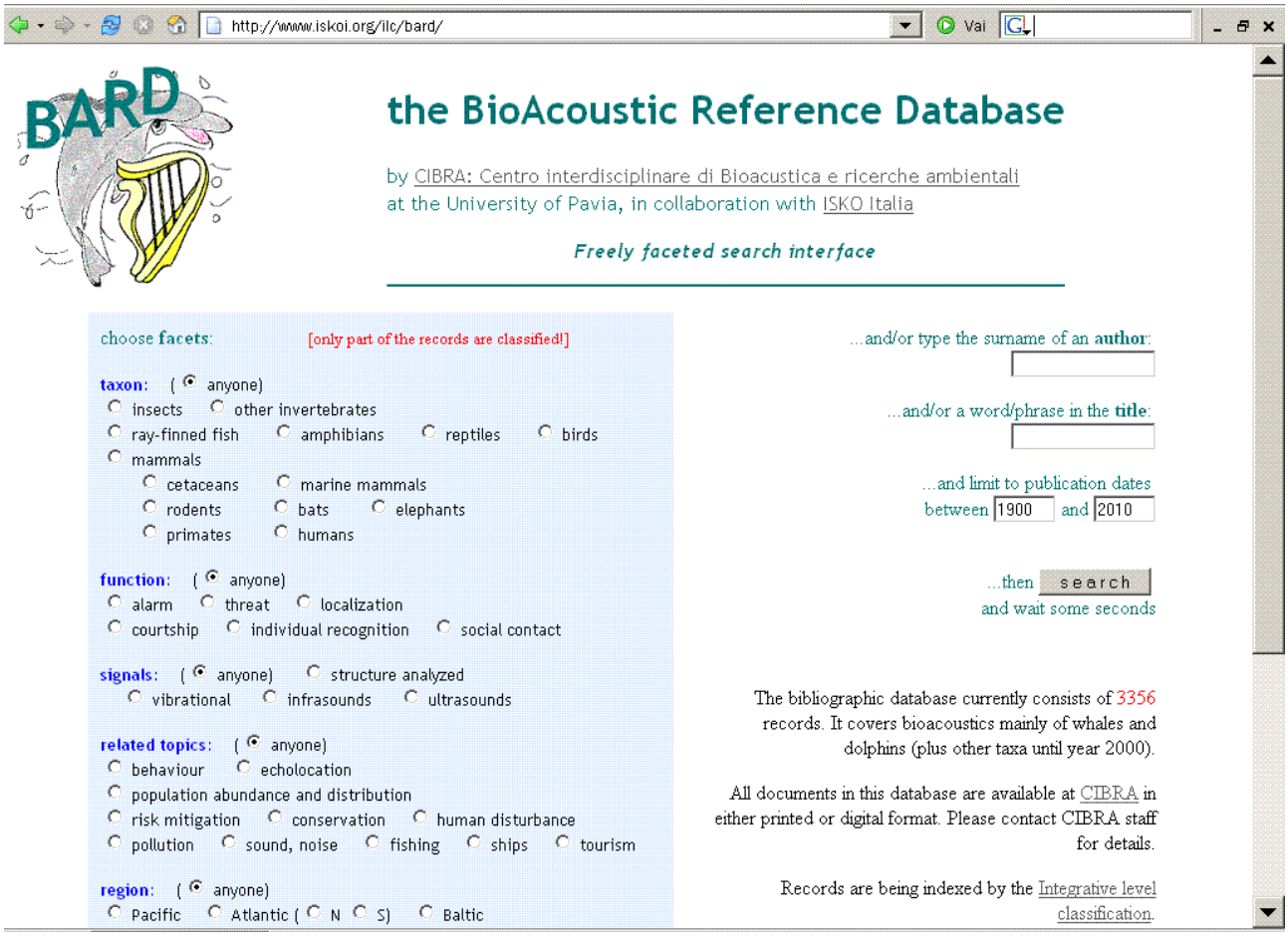

Figure 2: The user interface

6: Conclusions

The first obvious thing that remains to be done is completing the classification of all items with fully faceted classmarks (some items have been classified provisionally by broad classmarks). In doing this, a consistent indexing policy has to be developed from practice, so that all papers are classified with a degree of detail proportional to the specificity of their content. This may imply that some classmarks already assigned are revised in order to improve consistency.

Automatic caption generation is to be fixed for some complex cases. Although for this archive this is quantitatively not a major problem, its solutions are expected to teach lessons relevant for the more general purpose of automatic management and application of freely faceted classification.

Other improvements pertain more to the usability of the search tool, like improving the selection menu, and allowing subject search by typing words and matching them with the synonym set of each class. Also, the indexer interface could evolve into a real assistant tool, making the

indexing process quicker and easier, hence sustainable for larger amounts of documents.

Despite these improvements yet to be done, the experience we have already gathered with the BioAcoustic Reference Database shows that freely faceted classification by phenomena, theories, and methods is feasible [Szostak 2007], and that it can be used in a reasonably effective way to index real literature. As our experience with it will grow, more details of both the general ILC scheme and the BARD indexing policy will be gradually improved.

Once the system will be more consolidated, it would be possible to submit it to effectiveness measures in terms of precision and recall,[3] and to compare these with those of different systems. Studies of this kind have been done already in the past decades for technological domains, and have become famous as the Cranfield tests: in their results there was apparently not much difference between the effectiveness of faceted systems and that of free keyword systems [Cleverdon 1991]. However, Austin [1971] commented that effectiveness probably also depends on the domain indexed: while the expression of relationships between concepts can be not strictly necessary in technological domains, it becomes much more crucial in the human sciences and in the interdisciplinary domains: the relationships between concepts like "history", "printing", and "bibliographies" can have several senses, making a set of search results relevant or not according to the needs of different users. Tests with our system could reveal something on this line, as different concept combinations like "vessels damaged by whales" vs. "whales damaged by vessels" will have different relevance.

If these hypotheses are true, one can conclude that freely faceted classification can be especially useful for some indexing situations, including (1) specialized rather than generic documents (*deep classification*), (2) social and human sciences rather than natural and technical ones (*soft* vs. *hard sciences* in Austin's terms), (3) interdisciplinary contexts. These, after all, are not minor areas: in particular, the need for interdisciplinarity seems to be rapidly growing in present-day research [ISKO Italy 2007], making freely faceted classification something worth to put more efforts in.

3 This was appropriately suggested by Hans-Peter Ohly during the conference.

Acknowledgements

The Integrative Level Classification research project, lead by Claudio Gnoli within the Italian chapter of ISKO, also involves currently Mela Bosch, Enzo Cesanelli, Philippe Cousson, Hong Mei, Gabriele Merli, Roberto Poli, and Rick Szostak; and previously, Viviana Doldi, Marcella Patania, and Lorena Zuccolo.

The Interdisciplinary Center for Bioacoustics and Environmental Research at the University of Pavia is lead by Gianni Pavan, and involves currently Claudio Fossati, Marco Priano, Elisabetta Bernuzzi, Amanda May Koltz, and previously Michele Manghi and other collaborators.

We are grateful to all people from the ILC team, the CIBRA team, the ISKO UK chapter, the Konstanz conference, and other occasions, who provided stimulating comments and discussion to this work and to the León Manifesto.

References

Au W.W.L., Mobley J., Burgess W.C., Lammers M.O. 2000, Seasonal and diurnal trends of chorusing humpback whales wintering in waters off western Maui, Marine mammal science, 16, n. 3, p. 530-544.

Austin D. 1971, Two steps forwards, in B.I. Palmer, Itself an education, 2nd ed., Library Association, London, p. 86-88.

Austin D. 1976, The CRG research into a freely faceted scheme, in A. Maltby ed., Classification in the 1970s: a second look, Bingley, London, p. 158-194.

Cleverdon C.W. 1991, The significance of the Cranfield tests on index languages, in Proceedings 14th Annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, p. 3-12.

Dahlberg I. 1982, Information Coding Classification: principles, structure and application possibilities, International classification, 9, n. 2, p. 87-93.

Gnoli C. 2007*a*, Progress in synthetic classification: towards unique definition of concepts, in Information access for the global community: proceedings international seminar on the Universal Decimal Classification: The Hague: June 4-5 2007, Extensions & corrections to the UDC, 29, p. 167-182, also in DLIST, <http://dlist.sir.arizona.edu/1945/>.

Gnoli C. 2007*b*, "Classic" vs. "freely" faceted classification, in Ranganathan revisited: facets for the future: ISKO UK meeting: London: November 5 2007, audio and slides at <http://www.iskouk.org/kokonov2007.htm>.

Gnoli C. & Hong M. 2006, Freely faceted classification for Web-based information retrieval, New review of hypermedia & multimedia, 12, n. 1, p. 63-81.

Gnoli C. & Merli G. 2005, Notazione e interfaccia di ricerca per una classificazione a livelli, AIDA informazioni, 23, n. 1-2, p. 57-72, <http://www.aidainformazioni.it/indici/tuttonline/2005-12.pdf>.

ISKO Italia 2004, Integrative Level Classification: research project, <http://www.iskoi.org/ilc/>.

ISKO Italia 2007, The León Manifesto, <http://www.iskoi.org/ilc/leon.htm>, republished in Knowledge organization, 34: 2007, n. 1, p. 6-8.

Szostak R. 2007, Interdisciplinarity and the classification of scholarly documents by phenomena, theories, and methods, in B. Rodríguez Bravo & M.L. Alvite Díez eds., La interdisciplinariedad y la transdisciplinariedad en la organización del conocimiento científico: actas VIII Congreso ISKO-España: León: 18-20 Abril 2007, Universidad de León. Secretariado de Publicaciones, p. 471-477.

Szostak R. 2008, Classification, interdisciplinarity, and the study of science, Journal of documentation, 64, forthcoming.

Szostak R. & Gnoli C. 2008, Classifying by phenomena, theories and methods: examples with focused social science theories, in Culture and identity in knowledge organization: proceedings Tenth international ISKO conference: Montréal: August 5-8 2008, Ergon, Würzburg, forthcoming.

Tudhope D. & Binding C. 2008, Faceted thesauri, Axiomathes, 18, special issue on Facet analysis, forthcoming.