Simpler search in a complex world: Browsing ethnographic videos by freely faceted classmarks

Claudio Gnoli[1], Douglas Tudhope[2], Patrícia de Almeida[1], Romana Barbui[1], Ceri Binding[2] and Véronique Blot[1]

[1] *ILC project, University of Pavia, Via Ferrata 1, Pavia, 27100, Italy*
[2] *ILC project, University of South Wales, HRG, Llantwit Rd., Pontypridd, CF37 IDL, Wales, UK*

Abstract

Nobel Prize Giorgio Parisi recommends that the complexity of knowledge is reflected in all its levels and in an interdisciplinary way in schools and universities. On the other hand, complexity is often considered to be a hindrance in information retrieval. One method to both represent complexity and enable flexible information retrieval is faceted classification. In particular, the Integrative Levels Classification (ILC) provides ways to index phenomena and relationships among them by combining facets freely in an expressive notation. We have selected a sample of representative YouTube videos that document traditional feasts in Western Europe, and indexed them by ILC faceted compounds. Some of these are intrinsically complex events involving a variety of phenomena at different integrative levels, including animals, regions, artifacts, music, celebrations, religion, etc. We describe "Traditional Europe" (TradEU), a web-based demonstrator by which users can select topics from a set of 5 facets and launch a search in the video index. English equivalents of ILC classmarks for each video are displayed dynamically. Expressive faceted classification is leveraged by PHP scripts. Overall, the demonstrator shows a possible compromise between simplicity of use and intrinsic complexity of indexed contents.

Keywords

Complexity, ethnography, faceted classification, Integrative Levels Classification, ILC, web-based user interfaces.

1. Complexity and knowledge organization

In 2021, the Nobel Prize for Physics was awarded to Italian physicist Giorgio Parisi for his research on complex systems (see [1]). Parisi recently had a public conversation with philosopher Mauro Ceruti concerning "complexity and the organization of knowledge": they discuss "how the ways knowledge is organized in our schools and universities promote the possibility of acknowledging a knowing mode not reducing any knowledge object to a single level of description, a single level of reality, a single level of observation."

Indeed, the opposite traditional approach is limited by "this idea that knowledge can be divided into several special disciplinary fields, each qualified for a part of knowledge, and that one discipline does not interfere with another" [2].

To account for such variety of knowledge levels and of their interactions, an interdisciplinary approach to knowledge organization is needed, including the development of new knowledge organization systems (KOS) that are not constrained by the grid of traditional disciplines [3]. Also, in order to represent complexity and to address it in a variety of intellectual tasks, we need KOSs that are sophisticated enough.

[1]ISKO UK conference, July 24–25, 2023, Glasgow, Scotland
EMAIL: claudio.gnoli@unipv.it; douglas.tudhope@southwales.ac.uk, mebpatricia@gmail.com, ceri.binding@southwales.ac.uk
ORCID: 0000-0002-4721-7448 (Gnoli); 0000-0002-5222-0430 (Tudhope); 0000-0002-6376-9613 (Binding)


Nowadays, information is increasingly shared online through linked data, where only binary relationships can be expressed through RDF triples. However, this format may miss part of the complexity of such relationship networks. In an example by Soergel, we may need to express the benefit of *a certain dosage of* a chemical to *a certain combination of* condition, severity, age and sex: this requires a KOS with a powerful syntax, as after all "the world is not built of triples" [4].

One highly sophisticated type of KOS is faceted classification, which is known to enable accurate indexing of complex subjects [5]: it thus seems an appropriate tool for representing the variety and complexity of contemporary knowledge, being

> a way to represent and to organize, in a predictable manner, the increasingly complex content in documents. The representation of different aspects of a subject through notation, and their coordination through the consistent application of an order of combination, or citation order, provided a reliable way to determine where complex topics would be located in the linear sequence [6, p. 11].

However, the complexity of faceted KOSs also means that their potential is not always fully implemented in retrieval interfaces. Some believe that they are too demanding for both indexers and users to be applied to significant databases [e.g. 7]. In this paper, we address the problem of reconciling complexity of meanings with simplicity of use.

2. Expressing complexity by ILC facets and SKOS

The Integrative Levels Classification (ILC) is a developing general faceted KOS listing classes of phenomena, as opposed to disciplines (http://www.iskoi.org/ilc/). This is meant as a solution to the restrictions imposed by the disciplinary approach to the interdisciplinarity of knowledge. ILC features a rich syntax by which various kinds of relationships between themes can be expressed, including branching networks that may be represented in brackets. Syntactic relationships are expressed by facets belonging to ten fundamental categories expressed by numerals *9* to *0* (respectively: quality, quantity, part, property, change/process, opposition, agent, context/place, sequence/time, perspective).

Recent experience in ILC development and testing has led to the consistent identification and representation of different kinds of facets, based on the extension of their domain and range [8]. These are now being recorded in the developing version of the system (ILC3). Adopting Derek Austin's [9] phrase, the system is said to be *freely faceted* as it allows the expression of relationships between any two phenomena in different classes. For example, *wlge* "Easter" and *xb* "sculptures" belong to two different main classes (customs and creative arts respectively) but can be combined as *wlge50xb* to express sculptures used during Easter feasts. *50* is a free facet connecting any two classes in a process relationship. On the other hand, *w5c* "procession" is a process facet bound in class *w* "customs", from which compound *wlge5c* "Easter, procession" may be built. Parallel facets also exist, which allow one to use the hierarchy of a class or facet as the foci of another facet: *w92 [uU]* "customs, typical of *region*" takes its foci from the subclasses of *uU* "contemporary political divisions", so that *wlge92ff* means "Easter, typical of Lombardy".

While most of the elements of the ILC can be modeled within the SKOS format, the ILC's facet relationships and its ability to express coordinated compound descriptions extend beyond the SKOS standard. Binding et al. [10] describe the recent conversion of ILC second edition to SKOS (https://bartoc-skosmos.unibas.ch/ilc2/) and reflect on various theoretical and technical issues arising from the process. The method involved the use of *rdf:Property* sub-hierarchies to represent ILC facet indicators. This was complemented by defining extensions of the associative relationship (sub-properties of *skos:related*) to model the ILC's fundamental facet relationships. This offers a low-cost migration path to the world of the Semantic Web, with a limited extension to incorporate RDF properties, allowing the representation of faceted compound descriptors in the ILC (and potentially other faceted classification schemes). A series of SPARQL queries demonstrate how compound faceted expressions in the SKOS representation could be queried, and future options for query builders to support search interfaces are discussed, building on the ILC fundamental categories [10]. This current paper reports on an alternative approach (not employing associative relationships or SPARQL) towards

developing search functionality capable of taking advantage of the useful complexity offered by ILC faceted compound descriptors.

3. Demonstrating freely faceted classification by friendly interfaces

While believing that complexity is inherent in any KOS aiming at a careful representation and indexing of knowledge contents, we are also aware of the need for ILC applications that are simpler to understand and use, so that the system potential can be demonstrated and made known more widely. These should be publicly available on the Internet, so that interested parties can find and test them. We hope to show that it is possible to design faceted search interfaces that are relatively easy to understand and use, while still keeping the amount of detail and richness of relations in a freely faceted classification.

For example, a limited number of knowledge items belonging to a widely-known domain can be indexed and made searchable by ILC on the Web. We have considered indexing plots from a selection of movies, e.g. all the movies of a widely-known director, to show the application of ILC to fictional contents familiar to many users. Indeed, retrieval of movies by traditional disciplinary classifications, such as the Dewey Decimal Classification or Library of Congress Classification, is problematic [11] and cannot be done by their subject content, thus losing interdisciplinary connections with non-fictional knowledge that are potentially useful [12].

A first online faceted application of ILC used its first edition (ILC1) in the BioAcoustic Reference Database, allowing for selection of foci in different facets and browsing of the resulting bibliographical references sorted by classmark [13]. Attempts in this direction with different KOSs had been made already by Pollitt et al. [14], who showed "how drop-down menus revealing individual facet hierarchies could be used to frame complex searches" [6, p. 14], while Tudhope et al. [e.g. 15] combined facet selection with functionalities for query recommendation and expansion.

We are now working on a web-based demonstrator using the full power of faceted syntax that is being implemented in the developing ILC3. While the hierarchical trees of individual facets are indeed a good guide to special domains, in a freely-faceted classification any concept from the whole scheme can potentially occur in a combination (in our real case below, "Easter" gets combined with "birds" and "sculptures"): this makes it possible to only show a selection of the most typical classes in the focused domain – in our case, frequently-occurring facets of traditional feasts – leaving the remainder of the whole classification scheme as a further potential source. In general, indexing cultural objects [16] needs to account for their richness and complexity.

4. Technical solutions for the Web-based demonstrator

Under the title "Traditional Europe" (TradEU), our demonstrator (https://www.iskoi.org/ilc/ tradeu/) indexes a selection of a hundred publicly available YouTube videos that document various folk events in Western Europe, such as feasts and celebrations on the occasion of Carnival, Easter, May Day etc. in places where such traditions are kept alive and carry relevant ethnographic knowledge. Part of these are identified and selected by Lo Stivale che Balla ("The Dancing Boot"), an association for research and education in the traditional dances of Italy.

By accessing a relatively intuitive web page (Fig. 1), users can click and select the most common foci from the five facets Custom, Region, Symbol, Rite and Instrument, e.g. Custom = "Easter" and Region = "Italy". As observed above, these are just a selection of the many possible classes and facets: displaying full hierarchies could improve the browsing functionalities, but would also make the user choice more complex.

While users interact with English captions, HTML and Javascript code associates them with ILC notation fragments (*wlge* and *92f* respectively) and adds truncation symbols after each of them. Truncation enables the leveraging of the structure of an expressive faceted notation in two ways:

- matching facets with more specific foci, such as *-92ff-* "typical of Lombardy" as this belongs to the more general focus *-92f-* "typical of Italy";

- matching combinations featuring any other facets before the selected one, such as *wlge95eiq92f* "Easter, typical of Romanians, typical of Italy", as facet *-95-* "typical of ethnicity" is cited before facet *-92-* "typical of region" in the standard citation order. (A more likely subject would be *wlge95eiq27f* "Easter, typical of Romanians, in Italy".)

Figure 1: TradEU homepage

As one clicks on the "search" button, the selected notation fragments are passed to the results page as URL variables. The results page runs a query for such variables in the MySQL database where data about the indexed documents are stored, including a classmark and the YouTube identifier for each video. The videos that match the query are then directly embedded and displayed in an HTML table, each with the corresponding classmark, in our case *wlge92ff5d53a* and *wlge92fq75vo6ee5c5e5h 50xf50xb53a53c53d* – the latter having a degree of complexity beyond the average – together with its automatic translation into English.

Playable videos provide an immediate sense of the content without any further abstract step. Classmarks obviously look tough to interpret, but are displayed in a less evident font, only to suggest how the system works. What a user will notice first is their meaning, which is shown by automatically translating each facet's notation into the corresponding English caption taken from the ILC schedules. The translation is performed by a PHP script using string functions that leverage the expressive structure of ILC notation, by parsing each classmark at each sequence of numerals (i.e. facet indicator) and looking for the focus corresponding to the subsequent letters, depending on the facet type as expressed in the numerals (indicators ending with *-0* are free facets, those consisting of a single digit or ending by *-9* are bound facets, the remainder are parallel facets). In the examples, the resulting translations are "Easter, typical of Lombardy, itinerant begging, by voice" and "Easter, typical of Campania, birds, invoking mother goddess, procession, offering, ritual meal, through paintings, through sculptures, by voice, by idiophones, by membranophones". As it can be seen, each facet is cited after a comma and consists in a term possibly preceded by a preposition; many bound facets, such as "procession", have no preposition as their meaning is clear as it is.

5. Discussion: a complex world

Automatic translation of classmarks is made possible by the consistency of ILC expressive notation, which follows a set of predictable rules. As the algorithm is being developed to cope with all possible cases of faceted combinations, it will also be possible to apply it in future applications, such as an online

classmark builder that is planned to help users understand the mechanisms of ILC without getting lost in exceedingly technical details. At present, the algorithm manages most facets of all the three mentioned types, while it does not treat any bracketed branching (which almost never occurs in the sample documents), as this is simply ignored without displaying any caption. We expect that this will be solved by developing additional code.

Our latter example represents the feast of Madonna delle Galline, an especially complex event as expressed in its very long classmark. Indeed, the celebration involves live chickens and other birds, setting of personal altars with various kinds of symbolic objects, use of various musical instruments and other cultural expressions, both religious and profane, all integrated in a single event. Complexity thus arises in the feast itself, while its indexation by ILC is just a way of representing useful parts of such complexity. Complex phenomena such as traditional events can be described in principle by adding as many details as wanted. Not all details are expected to be equally relevant for users in order to compare traditional feasts: for example, while it may be interesting that birds are employed as symbols, the exact species of birds employed in different feasts may be less important. A similar situation is encountered in indexing fictional novels, where such themes as moral values, personal relationships and certain details of settings may be relevant, while others like the main character's favoured food may not be [12]. Selection of the relevant concepts involves the identification of a base theme and several particular themes and their expression in an appropriate citation order [17, 18]. These considerations suggest the need for a consistent indexing policy and, more generally, for some guidelines on the subject analysis of documents [cf. 19], in particular those representing complex and rich events or stories like video documentaries, novels or paintings [20, 21].

Tudhope and Binding [22] discuss faceted search and observe that "[i]t may be hard to find exact matches for complex faceted queries and index descriptors. Semantic expansion allows partially matching results with semantically similar but not identical index terms". The "zero-match problem" especially occurs when all possible facets and foci are listed, but not all their combinations occur in the indexed documents. While our demonstrator currently offers a menu with only a selection of commonly occurring facets, making the zero-match problem less serious, one can conceive a more advanced interface allowing the user to browse the full hierarchies of all relevant facets as taken directly from the ILC schedules, and to integrate it with functionalities for recommendation of related classes, expansion of queries to skip a searched facet or include broader foci, and so on. Clearly, such functionalities are only made possible by using a classification system that fully implements the principles and syntax of facet analysis, as opposed to so-called "faceted" navigation interfaces that only use non-subject fields like date of publication, format, author names etc. [6].

Faceted classifications are powerful tools that allow the representation of complex networks of relationships among many concepts. Clearly, not all applications require such a degree of detail. ILC can also be used in a "tagging" mode by just listing several concepts without making their relationships explicit, as it is e.g. in BARTOC. However, the TradEU demonstrator has been conceived to test the potential of fully-faceted ILC classmarks.

In conclusion, our demonstrator shows how search can be made relatively easy while keeping the complexity of subject representation. Complexity is encoded in the freely-faceted classmarks, but users are not required to deal with it all. For example, ILC classmarks follow rules of citation order of facets, according to the inversion principle of facet analytical theory, in order to provide an optimal, consistent sequence of results: but knowing such rules is only up to the indexer, possibly helped by a classmark builder as we are planning, while users just have to select from a limited number of facets and to browse the results page. It seems that optimal compromises between complexity of knowledge contents and friendly access to them lies in an informed design of user interfaces.

6. References

[1] G. Parisi, Complex systems: a physicist's viewpoint, Physica A 263 (1999) 557-564.

[2] G. Parisi, M. Ceruti, La complessità e l'organizzazione dei saperi, Complexity Institute, 2013. URL: https://www.complexityinstitute.it/la-complessita-e-lorganizzazione-dei-saperi-un-dialogo-tra-giorgio-parisi-e-mauro-ceruti/.

[3] R. Szostak, C. Gnoli, M. López-Huertas, Interdisciplinary knowledge organization, Springer, Cham, Switzerland, 2016.
[4] D. Soergel, Knowledge models for a sustainable world, keynote presentation at 5th Conference of ISKO Spain-Portugal, Lisbon, 25-26 November 2021. URL: http://www.dsoergel.com/ Soergel-ISKO2021LisbonKeynoteKnowledgeModelsForASustainableWorld.pptx.
[5] V. Broughton, The need for a faceted classification as the basis of all methods of information retrieval, Aslib Proceedings 58.1-2 (2006) 49-72.
[6] V. Broughton, Facet analysis: the evolution of an idea, Cataloging and classification quarterly 61 (2023). URL: https://doi.org/10.1080/01639374.2023.2196291.
[7] J. Maniez, Are classifications still relevant in databases? in: Documentary languages in databases: Papers from the Rome Conference, December 3-4, 1990, volume 3 of Advances in knowledge organization, Indeks, Frankfurt am Main, 1990, pp. 120-129.
[8] C. Gnoli, Faceted classification as linked data: A logical analysis, Knowledge organization 48.3 (2021): 213-218.
[9] D. Austin, The CRG research into a freely faceted scheme, in: A. Maltby (Ed.), Classification in the 1970s: A second look, Bingley, London, 1976, pp. 158-194.
[10] C. Binding, C. Gnoli, D. Tudhope, Migrating a complex classification scheme to the Semantic Web: Expressing the Integrative Levels Classification using SKOS RDF, J. Documentation 77.4 (2021) 926-945. Available from URL: https://bit.ly/2ocaHC6.
[11] C. Higgins, 'I coulda had class': The difficulties of classifying film in Library of Congress Classification and Dewey Decimal Classification, Knowledge organization 49.2 (2022) 79-86.
[12] P. de Almeida, C. Gnoli, Fiction in a phenomenon-based classification, Cataloging and Classification Quarterly 59.5 (2021) 477-491.
[13] C. Gnoli, G. Merli, G. Pavan, E. Bernuzzi, M. Priano, Freely faceted classification for a Web-based bibliographic archive: The BioAcoustic Reference Database, in: J. Sieglerschmidt, H.-Peter Ohly (Eds.), Wissensspeicher in digitalen Räumen: Nachhaltigkeit, Verfügbarkeit, semantische Interoperabilität: Proceedings der 11. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissenorganisation, Konstanz, 20. bis 22. February 2008, Ergon, Würzburg, 2010, pp. 124-134. URL: http://hdl.handle.net/10150/106114.
[14] A.S. Pollitt, A.J. Tinker, P.A.J. Braekevelt, Improving access to online information using dynamic faceted classification, in: Online Information 98: Proceedings of 22nd Online Information meeting, London, 1998, Oxford: Learned Information Europe, 1998, pp. 17-21.
[15] D. Tudhope, C. Binding, D. Blocks, D. Cunliffe, Representation and retrieval in faceted systems, in: M.J. López-Huertas, F.J. Munoz-Fernandez (Eds.), Challenges in knowledge representation and organization for the 21st century: Proceedings of the Seventh international conference of ISKO, Granada, July 10-13, 2002, Ergon, Würzburg, 2002, pp. 191-196.
[16] M. Baca, P. Harpring, E. Lanzim, L. McRae, A. Baird Whiteside on behalf of The Visual Resources Association, Cataloging cultural objects: A guide to describing cultural works and their images, American Library Association, Chicago, 2006.
[17] A. Cheti, Testo e contesto nell'analisi concettuale dei documenti [Text and context in the conceptual analysis of documents], in: M. Guerrini (Ed.), Il linguaggio della biblioteca: scritti in onore di Diego Maltese, Editrice Bibliografica, Milano, 1996, pp. 833-855.
[18] A. Cheti, Il punto di vista del GRIS sulla "relazione di soggetto" in FRBR [GRIS's viewpoint on the "subject relationship" in FRBR], in: M. Guerrini (Ed.), Principi di catalogazione internazionali: una piattaforma europea? Atti del convegno internazionale, Roma, Bibliocom-51o Congresso AIB, 27 ottobre 2004, AIB, Roma, 2008, pp. 91-100. Also available at URL: http://www.aib.it/aib/congr/c51/chetint.htm.
[19] ISO 5963:1985: Documentation: Methods for examining documents, determining their subjects and selecting indexing terms, International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso:5963:ed-1:v1:en.
[20] S. Shatford Layne, Some issues in the indexing of images, J. Am. Soc. Information Science 45.8 (1994) 583-588.
[21] A. Salaba, Indexing of non-textual materials, keynote speech at conference Look beyond: Subject indexing of non-book resources, Rome, 6 February 2023. URL: https://www.aib.it/attivita/congressi/2023/104579-look-beyond-abstract/.

[22] D. Tudhope, C. Binding, Faceted thesauri, Axiomathes 18.2 (2008) 211-222. URL: https://doi.org/10.1007/s10516-008-9031-6.