Thesaurus Maintenance

Methodological Outline

Martin Doerr[1], Maria Daskalaki[1], Chryssoula Bekiari[1]
Helen Katsiadakis[2], Helen Goulis[2], Christos Terzis[2]
[1]*Institute of Computer Science, Foundation for Research and Technology Hellas Heraklion-Crete, Greece*
[2] *Academy of Athens*

Introduction

It is difficult to speak about method. This is especially true when it comes to the epistemological methods of organizing and transferring existing knowledge into another language, computer language. The challenge of digitizing existing knowledge is one of the most exciting tasks, because not only does it determine the sort of knowledge we bequeath to the next generation, but it also defines the learning process, since every act of rewriting knowledge and information is also a transformation of the existing knowledge into another form with a different structure and organization.

In DARIAH we have the opportunity to contribute to the building of this new language that will be used to express a significant amount of humanities related knowledge. This task, to contribute to the transformation of knowledge into another form, is both a great opportunity and at the same time a very difficult task, as we have to define the fundamental concepts that will be used as the foundation for such a project. This is the reason why we propose a particular interest independent approach.

As Barry Smith observes, the problem is that "different groups of data- and knowledge based system designers have their own idiosyncratic terms and concepts by means of which they build frameworks for information representation".[1] The consequence is that "different databases may use identical labels but with different meanings or "the same meaning may be expressed via different names".[2]

[1] B. Smith, "Ontology", in: *Blackwell Guide to the Philosophy of Computing and Information*, Oxford: Blackwell, 2003, 159-166, p.159.
[2] Ibid.

Although doubts have been expressed about the success of a project that will attempt to create a classification system, by means of which different classification systems used by different groups could be made compatible to each other, there are cases that stand proof that such a project is feasible. As will become apparent below, the faceted classification gives us the intellectual tools to define the terms in a way that will not force us to reconstruct, at least the upper part of the classification with each extension, and thus provide a fundamental stability, at least for basic queries. Furthermore, the faceted classification leads us to the acceptance of methodological principles, which can protect us from the logical errors and the idiosyncratic decisions, in particular in the upper level terms, that are responsible for major incompatibilities in the classification. It is a kind of classification that is not only based on a small number of concepts under which we classify the terms, but it allows the continued expansion into new scientific fields without requiring redesigning - refinements of definitions notwithstanding. Our aim is therefore to build a backbone thesaurus, which will be capable of accommodating the specific terms and concepts of thesauri restricted only to a specific scientific domain. We use ontology-driven faceted analysis to define the top-level concepts that will be used as the backbone for organizing knowledge in thesauri.

Defining our Object

In designing and organizing our backbone Thesaurus, we have first to define the object we deal with, which, in our case, is the natural language that consists of words. But words can have multiple meanings depending on context. It is not in our scope in building the Thesaurus to list the various meanings of a word, as this is the purpose of dictionaries. Our aim is to organize terms under categories that express their substantive identity. D. Soergel explains the difference between dictionaries and thesauri as follows: "A **dictionary** is a listing of words and phrases giving information such as spelling, morphology and part of speech, senses, definitions, usage, origin, and equivalents in other languages (bilingual or multilingual dictionary). A **thesaurus** is a structure that manages the complexities of terminology in language and provides conceptual relationships, ideally through an embedded classification/ontology. A thesaurus may specify descriptors authorized for indexing

and searching. These descriptors then form a **controlled vocabulary (authority list, index language)**".[3]

Terms are also words defined as parts of an expert language. In this case, the words have a specific meaning and express in an unambiguous way the scientific opinions of the experts' group and their discipline. These scientific terms are gathered and defined *in disciplinary dictionaries.* Terms can be classified under concepts, which are classes or sets of items grouped together on the basis of some implicit or explicit criterion or rule. The criterion can be unconscious or even innate.

One of the main tasks in designing backbone thesauri is to select and identify the concepts that enable the data to be searched, so that all items relevant to research questions can be found. Our purpose is thus to design a backbone thesaurus, which can integrate in a principled manner other discipline-oriented thesauri or controlled vocabularies, in order to facilitate the interdisciplinary research and interoperability between different scientific fields.

Background of the faceted classification

The problem we are faced with is that every discipline has its own classification concepts and, as a consequence, there are many small thematic vocabularies, discipline or simply application specific. But discipline-oriented vocabularies are hardly hierarchically or semantically structured in a principled way. There have been some efforts to design larger vocabularies (Library subject headings LCSH, SWD), which however are not semantically structured and thus the interactivity between the disciplines cannot be achieved. But even classifications that are organized in a principled and hierarchical way, such as Dewey Classification, adopt arbitrary and biased criteria in order to classify their terms. Finally, the Herein project, an attempt to merge three thesauri into one, led to a complete failure. In our view, the major reason for the failure of Herein-project is that it tried to maintain the complete merged vocabulary on the basis of a central authority. Our intention is, on the contrary, to achieve a strict agreement only on high-level concepts ensuring a wide autonomy of the particular vocabularies of each field.

[3] D. Soergel, "Thesauri for Knowledge-based assistance in searching digital libraries", Proceedings of the 2nd European Conference on Digital Libraries, Heraklion, Crete, September 20, 1998, p. 16.

In light of the above, building a thesaurus that can be sustainable and applicable across different scientific fields, thus enabling interdisciplinarity sounds urgent. The examples of AAT, English Heritage and Merimee allow us to consider our goal, to build a backbone thesaurus, feasible.

Having in mind that: a) for resource discovery, recall is more important than precision, b) the terms for discovery can be much coarser than for documentation, c) the more generic the concept the higher the recall and, finally, d) the "known Item Search" works better by keywords and factual associations, we came to the conclusion that it is in our best interest to agree on and define the higher terms. In other words, we will attempt to build a backbone thesaurus based on a restricted number of top-level concepts (less than 100) that will enable compatibility among terms from different scientific fields. It will thus become possible to map small vocabularies to a set of top terms without imposing terms on experts, who want to preserve the meaning of the specific terms of their scientific field.

Our intention is to build a backbone thesaurus that can function as an indexing language and an "interlingua", from which smaller vocabularies borrow the upper concepts, without eliminating the context focus of small vocabularies. This can be achieved through the homogenization of hierarchies of small vocabularies and the creation of a new contextual relationship or term collection. The homogenization of hierarchies will be achieved on the basis of fundamental ontological distinctions as normalizing principles, while, at the same time, through the introduction of a new contextual relationship – next to "related terms" – the specificity of the experts' language will be preserved.

As the example of the GETTY'S AAT demonstrates, the best way to achieve interoperability between small vocabularies, feasibility and validity of the classification without losing the meaning of the terms derived from different scientific fields is through the faceted classification.

The invention of the "facets"

Facets are fundamental concepts, which appear as characteristic syntactic constituents for term composition, such as: "Persian X 19th century X rugs". As Taylor declares, a set of facets comprises "clearly defined, mutually exclusive and

collectively exhaustive aspects, properties or characteristics of a class or specific subject".[4] Facets can thus provide a global subdivision of concepts through the reduction of the composite terms to more primitives ones. The facets are primitive ideas, which build multidimensional conceptual structures and are fundamental for the classification of terms derived from different fields. In this sense, facets do not introduce an index but a taxonomy.

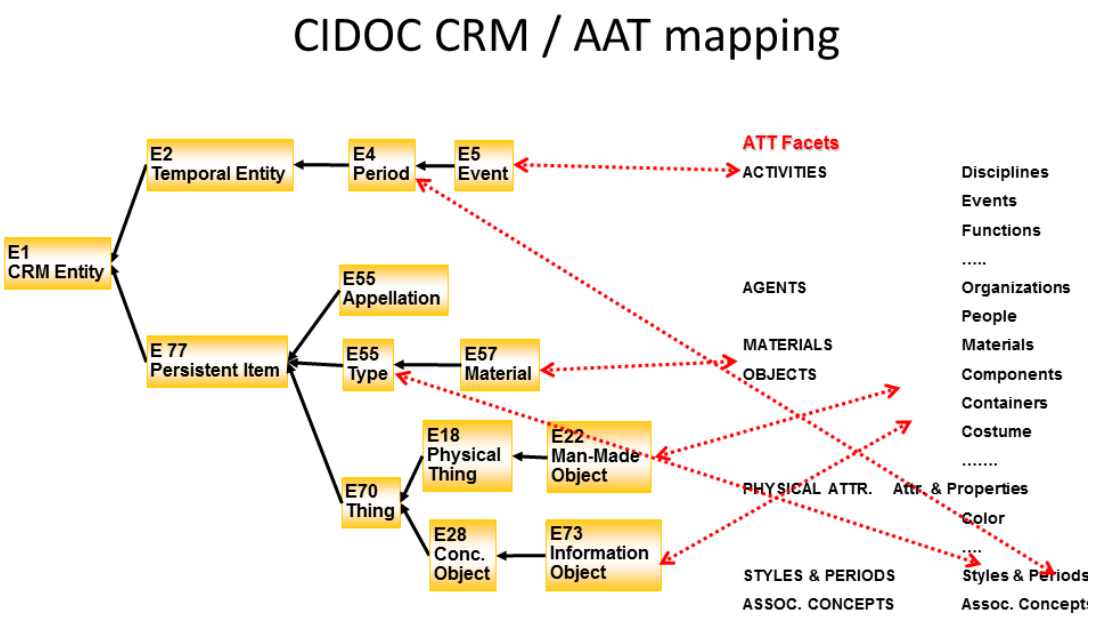
In the time span from 1925 to 1965 S. R. Ranganathan developed a system of library classification, which is called colon classification system. It constitutes basically the invention of the faceted classification. It uses five primary categories, or facets, to further specify the sorting of a publication. According to Ranganathan these categories or facets are: Personality, Matter, Energy, Space and Time. Collectively, they are called PMEST.

In the meantime there have been new, independent efforts to define facets in order to cover different needs emerging from different scientific areas. Such efforts were carried out by the AAT and the CIDOC CRM. So AAT and CIDOC CRM respectively have built the following facets according to the needs they have to deal with:

- AAT: Agents, Objects, Materials, Styles and Periods, Activities, Physical Attributes, Associated Concepts.
- CIDOC CRM: Actor, Physical Thing, Conceptual Object, Material, Type, Language, Period, Place, Time-Span, Dimension.

In order to show the similarities and differences between those two classification systems we can map them in the following way:

[4] A.G. Taylor, *Introduction to Cataloguing and Classification*, Westport CT: Libraries Unlimited, 2006.



At this point it is important to notice, that the above classification systems – and possibly other faceted classification systems that we do not mention here – although they are different and cover different needs, have striking conceptual similarities: in most cases, the facets that are developed in each system more or less coincide. This is an *a posteriori* indication that the high level concepts, which are used in different faceted classification systems, are not artificially imposed on the classification system, but they appear to be deduced from the way through which we generally perceive our reality. In other words, the facets acquire their validity because they express the most general categories of our perception. This is further supported by the fact that they exhibit analogies to linguistic constructs such as verb ("Energy"), subject ("Agent"), object, etc. So, if we classify according to facets and we follow the principles corresponding to this kind of classification, an intersubjective validity of our classification should be ensured.

How to design an effective faceted classification system

The first step in designing the backbone thesaurus is to define the functional restrictions of the indexing language. In other words, we have to define the purpose and the domain of discourse of our classification system. The purpose of our

classification is to facilitate a successful search of the existing knowledge. Therefore, we have to find the *"bonds"* between the terms and not to demarcate the terms. The domain of the discourse is crucial for every classification system, since the meaning of the terms very often varies depending on the context within which it appears; for example, the term "dependence" has different meanings in the contexts of computer science, medicine, psychology or social relations. In our case, the domain of discourse is humanities.

The second step in designing the backbone thesaurus is to detect the intensional properties of the concepts to be classified. The intensional properties are the *necessary*[5] and *sufficient*[6] conditions for belonging to a category. An example of a concept that exhibits intensional properties (necessary and sufficient conditions) is the *bachelor* defined as 'unmarried man'. Not being married is an essential property of a bachelor, because one cannot be a bachelor unless one is an unmarried man (necessary condition) and any unmarried man is a bachelor (sufficient condition).

The intensional properties are thus the essential characteristics, which express the 'substance' of a concept and provide an unambiguous recognition of an item as belonging to a category. Therefore the intensional properties cannot be replaced without loss of meaning. The recognition of the intensional properties must be transparent and based on information accessible to everyone. So, if the intensional properties are implicit or not commonly accessible, the term is defined through confining or referring to commonly known phenomena. (e.g. a necessary condition for defining human being could be the DNA. But DNA is not accessible to everyone, so we have to refer to other morphological characteristics, which are accessible to everyone in order to define our term. In the case of human beings, these characteristics are morphological). Knowing the intensional properties of a concept one can deduce the potential properties of an item. Potential properties are

[5] If we say that "x is a necessary condition for y," we mean that if we don't have x, then we won't have y. Or put it differently, without x, you won't have y. To say that x is a necessary condition for y does not mean that x guarantees y. For example air is a necessary condition for being alive, but it does not, by itself, i.e. alone, suffice for human life. Something can be a necessary condition without being sufficient.

[6] If we say that "x is a sufficient condition for y," then we mean that if we have x, we know that y must follow. In other words, x guarantees y. For example, rain pouring from the sky is a sufficient condition for the ground to be wet. Something can be a sufficient condition without being necessary.

consequences of the nature of a thing. They may be confined to a category or not. They may appear at some instances at some time. Detecting the intensional properties of a concept is a very important task, which takes us to the third step, in order to design an effective backbone thesaurus.

One of the main benefits of defining the intensional properties is that these reveal hierarchical relationships that can lead us to broader categories, which can be used in controlled vocabularies for the classification of the terms. For example, defining the intensional properties of the term "bachelor" reveals the broader category under which bachelor can be classified: that is the concept "man", since any bachelor must be a man. Finding the intentional properties of a concept we can thus construe hierarchical relationships that express the substantive characteristics of a concept in a dynamical way; in a way namely that allows the deduction of new properties (the potential properties) from the already known and accessible.

Forming broader categories

As already mentioned, the third step in building the backbone thesaurus is to find the broader categories, which enable an "open world". For that reason, it is important to take into account the fact that, when we generalize a concept into a broader category, it must be ensured that this concept possesses more general intensional (and potential) properties, **possibly** together with other concepts under that category. On the other hand, the items (terms, classes), which ***are not*** included in a specific broader category, are not characterized by the intensional properties of that category. However, it must be possible to identify things not belonging to this specific category (e.g. what is not a "Research Object"?). That does not mean that we define those terms as opposed to the terms (antonymity, complements), whose intensional properties we can identify. In an open world "having not a property" does not imply anything. Complete decomposition is a kind of negation! If for example we define a female human as not male man then how can we define the *transsexuals?*

Insofar as the potential properties are derived from an intension, a good broader category "confines" many potential properties that can only apply to a specific intension. Each refinement of intension may confine or guarantee another set

of potential properties: For example “Material Object” can have weight, elasticity; or a “living individual” consumes energy.

In designing classification systems it is not uncommon to confuse these two levels of intensional and potential properties. But the potential properties are in a great degree incidental and cannot thus be the criterion for classifying terms. When a concept is defined solely on the basis of potential relationships, not only the essential properties of the concept do not come into view, but also no further independent properties can be derived. Furthermore, classification according to potential properties often results in misleading conclusions. For example, if we define “human” as “driver” i.e. by his incidental property to drive a car, then all people who do not drive are not human.

As mentioned above, the broader categories must be defined on the basis of the intensional properties of the concepts. So, the broader categories can express the substance of the terms and enable the deduction of potential properties. Forming broader categories from common substance/nature, which enables potential properties, does not divide the world into disjoint classes. A particular thing may fit into multiple intensions. The higher categories derived from the intensional properties of the concepts cannot be justified in a logically exhaustive and strict way, but can be reached intuitively, by common sense and by reducing the more complex terms and concepts to primitive ones. Building such categories from bottom-up, we eventually reach the context-independent level, and finally we identify those elementary concepts that we already posses and through which we perceive and conceptualize our reality. These elementary categories are the *facets.*

Briefly, the *golden rules* of hierarchy building are the following:

- Levels of hierarchies are never absolute. Even Facets may have generalizations.
- Levels of hierarchy are never complete.
- Generalizations are never unique.
- Sets of sibling concepts are never complete. Anything that does not fit goes into the next level category, until a better specialization is found.
- Don’t complement levels by “other objects” or “elephants and non-elephants”
- Particulars (gazetteers, person lists) are NOT terminologies (but other KOS*)*

It is evident from the above that the faceted classification is not an artificial classification of the terms or a "top to bottom" classification, but is generated from the analysis/decomposition of one term in its elementary characteristics. Thus, the hierarchies (IsA relationship) resulted from this kind of classification are not arbitrary, but reveal the substantive properties of a concept and bring to light the "multiple links" between the terms. Accordingly, the faceted classification, on the one hand, reveals the complexity of a term while, on the other, reduces the term to its fundamental components. Faceted classification does not divide the world into closed spheres of meanings according to specific characteristics, but enables its enrichment and expansion with new hierarchies and facets without disrupting or disorganizing the existing ones.

Benefits of the faceted classification

The benefits of *faceted classification* can be summarized as follows:

- A term can be classified in multiple hierarchies (e.g. doll toys/visual works).
- It is independent of the context, within which each term appears, although the context is crucial for the classification of a term in facets.
- It is based only on a restricted number of fundamental concepts.
- It makes feasible the collaborative development of an upper level.
- It can be expanded and thus enables compatibility between different classification systems from different domains without imposing terms on the experts.
- It does not presuppose knowledge regarding the exact context of the terms.
- It helps us to discover concepts that are needed in searching or that enhance the logic of the concept hierarchy (e.g. train/bus station, harbour, airport=>traffic station)

Consequently the faceted classification system:

- ➢ Allows the designing of a consistent, stable and highly expressive set of fundamental concepts that will enable humanities experts to find adequate generalizations.
- ➢ Ensures interoperability between the thesauri already developed in specific scientific fields of the humanities within the DARIAH project.
- ➢ Facilitates users with their research inquiries.
- ➢ Helps us to avoid the methodological errors that will lead to inconsistencies and incompatibilities between the terms.
- ➢ Help us to achieve the greatest economy of effort in the process of organizing terms.

For the reasons mentioned above, we propose the construction of a backbone thesaurus based on faceted classification!

For illustration: Our experimental facets

1. Facet: Materials

2. Facet: Material objects

   Hierarchies: i) monuments

   ii) artefacts/objects

3. Facet: Conceptual objects

   Hierarchies: i) symbolic objects=>information object

   ii) propositional objects=>information object

   =>methods=>processes,

techniques

4. Facet: Natural Processes (“CRM Temporal Entity”)

   Hierarchies: i) natural disasters

   ii) natural genesis

5. Facet: Epochs (“CRM Temporal Entity”)

6. Facet: Activities (“CRM Temporal Entity”)

Hierarchies:

i) Disciplines => a) production of material objects and installations
b) conception and comprehension of phenomena
c) provision of knowledge and expertise (know-how)
d) production of aesthetic phenomena

ii) Events=> a) social events
b) conflicts
c) political, social and financial phenomena
d) administration

iii) Functions

iv) ….Other Activities: this is not a hierarchy!

For illustration: Scope notes

Facet: Activities

Scope note: The “Activities” facet comprises types of intentional actions that result in the preservation, creation, production, modification or destruction of an entity (living beings, conceptual/material objects, groups, social, intellectual, physical etc. phenomena).

Hierarchies

i) Disciplines: This hierarchy comprises types of branches of professional or potentially professional occupations socially and/or legally acceptable under the criteria of sector self-subsistence, practice efficiency, adoption of common methods and transferability of knowledge and expertise. Each sector includes types of unified activities that express some sort of professional or potentially professional specialization.

ii) Events: This hierarchy comprises types of intentional activities carried out by at least one actor causing or changing phenomena or states of affairs on the social, political, financial, cultural and intellectual level.

iii) Functions: This hierarchy comprises types of activities that are structural parts of a relatively stable complex system of permanent and self-contained procedures that repeat themselves within this system and thus contribute to its preservation. Although functions are part of a wider system, each function is completely distinct from the rest. As structural parts of a complex system, functions are types of actions that play a certain role within a system and aim at a specific goal, which they must accomplish.

In this respect the purpose that a certain function has to achieve cannot be different from that for which the function is performed. In other words, the purpose of a function is one of its identity criteria.

Consequently, the notion of the function univocally relates the actions performed and the target achieved by these actions in such a way that, if some other target is achieved due to external factors, we speak of a different function or activity.

Bibliography

Franz Baader, Ian Horrocks and Ulrike Sattler, "Description Logics", in: Steffen Staab, Rudi Studer (Eds), *Handbook on Ontologies,* Berlin Heidelberg: Springer, 2009, 21-43.

D. Beneventano, Fr. Guerra, St. Magnani, M. Vincini, "A Web Service Based Framework for the Semantics Mapping amongst Product Classification Schemes", in: *Journal of Electronic Commerce Research,* Vol. 5, No. 2, 2004, 114-127.

M. Doerr, D. Iorizzo, "The Dream of Global Knowledge Network – A New Approach", in: *Journal on Computing and Cultural Heritage*, 1 (1), 2008, 1-23.

Thomas R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing", in: *Int. J. Human-Computer Studies (*1995), 43, 907-928.

S.R. Ranganathan, *Colon Classification*, N. Delhi: Ess Ess Publications, 2012.

B. Smith, "Ontology", in: *Blackwell Guide to the Philosophy of Computing and Information*, Oxford: Blackwell, 2003, 159-166.

D. Soergel, “Thesauri for Knowledge-based assistance in searching digital libraries”, Proceedings of the 2nd European Conference on Digital Libraries, Heraklion, Crete, September 20, 1998.

J. F. Sowa, *Knowledge Representation. Logical, Philosophical and Computational Foundations*, Pacific Grove CA: Brooks Cole Publishing, 2000.

E. Svenonius, “Design of Controlled Vocabularies”, in: *Encyclopedia of Library and Information Science*, DOI:10.1081/E-ELIS 120009038, New York 2003, 822-838.

A.G. Taylor, *Introduction to Cataloging and Classification*, Westport CT: Libraries Unlimited, 2006.