Elena Cardillo (Fondazione Bruno Kessler, Povo)
Antonietta Folino (University of Calabria, Arcavacata di Rende)
Maria Taverniti (University of Calabria, Arcavacata di Rende)
Roberto Guarasci (University of Calabria, Arcavacata di Rende)

GoldThes: a faceted thesaurus for goldsmith handcraftsmanship in a regional context

Abstract
In industrialized countries, the rapidity of development has not taken into account the fact that it would be opportune to formalize “niche” knowledge. There is an ever-growing need for a written formalization of tacit knowledge. The aim of this activity consists in building specialized knowledge collections, which are useful to help and to guarantee the continuity of small and medium-sized business during successions, and to guarantee staff training related to specific practices. To reach this aim, this work presents the construction of a faceted thesaurus (GoldThes) for the domain of goldsmith handcraftsmanship, which tries to classify and organize a domain extracted from a regional context. The use of this kind of system demonstrated how classifying and organizing information into multi-dimensional hierarchies makes it more accessible than using a single taxonomy, that is a unique hierarchical dimension.

1: Introduction
In this paper we want to present a work which aims at the formalization of the tacit knowledge used to manage the passage from one generation to the next in firms, and to create unsupervised learning paths. To reach this aim we created a knowledge organization system (KOS), in particular a faceted thesaurus that could be used to develop a model of local innovation in the field of goldsmith handcraftsmanship.

The exigency of creating a KOS for specialized fields such as handicraft, where the expert’s know-how is lacking in formalization, grows from the need of protecting this expertise, making it formal, explicit and shareable. The traditional practice hand down of the master craftsmen is further threatened by other factors, such as: a state of missed and unorganized knowledge; the lack of technical knowledge on the domain and on the local specificity of techniques; and limited access to data and information. So it is essential to define procedures and methods for managing this knowledge, taking into account the specialness of some manufacturing techniques, of materials, and of some utensils.

The choice of adopting the facet technique for constructing a thesaurus in our specific domain consists in the possibility of using it as a user interface for searching, which would allow access to a knowledge repository filled in with the domain knowledge.

2: Background: faceted classifications and thesaurus
Faceted thesaurus is a special kind of thesaurus based on faceted classification, where terms are organized according to a series of facets, i.e. a sort of attributes describing properties of objects. The choice about the number and the type of facets depends on the domain knowledge, on the categorizing objectives and on the final user, but the principle underlying this kind of classification is always the same: multidimensionality in terms organization.
This organisation makes faceted classification a natural way to organise things. After choosing categories, each facet is populated with the corresponding terms that could receive a further organisation in the same facet, by the use of sub-facets that allow a multi-level description. The use of facets, in particular in a searching context, permits to describe things by combining the different aspects used to describe them. The flexibility of this kind of classification makes it possible to categorise new concepts, whose corresponding terms could be inserted in the faceted organisation without changing the existing one.

Many contributions in literature are focused on faceted classification, most of which give an overview of its usefulness and explanation of advantages and disadvantages of this kind of classification with respect to the hierarchical one. It is the case of the works of Gnoli (2007) and Slavić (2007).

During the last decades, there was a trend in preferring the use of faceted thesauri instead of the hierarchical ones to allow the access at knowledge bases on the Web. The advantages provided by faceted thesauri consist in a more efficient research activity and in an easier possibility of updating. A hierarchical representation of terms does not allow to structure knowledge dynamically, as it imposes a fixed organization where complex topics are difficult to express and new concepts are difficult to insert, because it is not possible to previously determine the increase of knowledge and the creation of new concepts. The continuous evolution of knowledge requires a more flexible system where the relationships between terms are logic and coherent and the different aspects of information can be easily combined.

One of the most important thesauri built on faceted principles is the AAT (Art & Architecture Thesaurus (AAT) <www.getty.edu/research/conducting_research/vocabularies/aat/>, hosted by the Getty Institute, which contains terms about art, architecture, decorative arts, material culture and archival materials. Another example of faceted thesaurus is represented by EARTH <uta.iia.cnr.it/earth.htm> that contains geological terms about the environment organized into three macro categories: entities, attributes and dynamic aspects. Then we can mention the Thésaurus du réseau européen du patrimoine (HEREIN) <www.european-heritage.net/sdx/herein/thesaurus/consult.xsp> specifically realized by several European partners for research activities on a document corpus about the national politics of the European Countries. Concerning multilingual thesauri, a good example is Agrovoc <www.fao.org/aims/ag_intro.htm>, a multilingual thesaurus that covers the domains of agriculture, forestry, fisheries, food and other similar fields.

3: Our approach
The project activities have been carried out in two macro-phases. The first phase has been dedicated to the creation of a knowledge base for the domain by means of an in-depth knowledge acquisition process (Cardillo et al. 2008). The second phase focuses more on the construction of the faceted thesaurus for the interrogation and consultation of the created knowledge base.

3.1: Knowledge acquisition
The knowledge acquisition process, realized from domain experts, consists in the extraction of all tacit knowledge to build a knowledge model and implement the
conceptual model, and is based on the use of a series of elicitation techniques. These
techniques are usually used in different phases of the knowledge acquisition process,
since each of them permits to capture a specific typology of knowledge and to achieve
specific aims. The most common techniques are interviews, followed by the direct
observation of expert performances, which allows for the extraction of procedural
knowledge, mostly connected to manual skills (think aloud problem solving, self-report,
and shadowing). Other techniques, such as card sorting and repertory grid, are useful
for understanding how experts conceptualise knowledge related to their own domain of
reference (Milton 2007).

In the case of goldsmith handcraftsmanship, we used the following elicitation
techniques: interviews; self reporting; shadowing; repertory grid and card sorting.

In order to simplify the creation of a conceptual model from the acquired knowledge,
we used the toolkit PCPack5 <www.epistemics.co.uk>, a knowledge acquisition tool
that allows conceptualizing knowledge from different sources. In particular, the
Protocol tool permitted us to structure the reference domain into significant classes, i.e.
instruments, techniques, materials, etc., proceeding with a normalization of these
elements, in view of their inclusion in the thesaurus to be created. The various elements
merge automatically into the knowledge base and into an ontology browser which
allows for their visualization and use in the remaining software modeling tools. The tool
allows to produce a final product represented by a knowledge model in a XML format.

In order to build the faceted thesaurus, it was necessary to perform a semi-automatic
extraction of the candidate terms from written sources such as ANSI and ISO standards,
the interviews, technical reports and a glossary of general jewelry terms.

The terminology extraction tool used for the detection of the candidate terms is Text-
2-Knowledge (T2K). Once the linguistic analysis of the T2K texts is performed, the
program semi-automatically acquires several ontologies as advanced support for
document management. It then provides, as final output, a term-based vocabulary
whose added value is represented by the terms’ semantic and conceptual information
regarding the vocabulary itself. The extracted terms became the labels/descriptors of our
thesaurus, and were organized into a hierarchical hyponym/hyperonym relation
depending on the internal linguistic structure of the terms; that is, by sharing the same
lexical head (e.g. incassatura), modifiers, and so on (e.g. incassatura invisibile, incassatura a binario, incassatura a lastra, incassatura a castone ribattuto). The candidate terms detected by T2K can be either single or
multi-word terms, and represent the terminology index of the domain analysed. In
general, the terms belonging to the vocabulary provided by T2K can be organized in
different lexical relations.

The computational analysis system adopted by T2K includes a specific tool for the
analysis of Italian (Bartolini et al. 2005). In particular, T2K integrates systems for
automatic linguistic analysis of texts, using statistical algorithms to identify and cluster
conceptual structures; annotation tools, or knowledge mark-up of the text; and finally
structured data supporting the conceptual-terminological indexing of documents
(Montemagni 1996).

A parallel manual extraction has been effected side by side with the automatic one, in
order to deal with the peculiarity of the processed texts and also to function as a control
group for the software extraction. Both extractions have been carried out separately for
each one of the corpora. Nevertheless, the evaluation of the results obtained using the two different procedures has demonstrated that a good amount of terms, certainly representative of the specific knowledge, were not automatically extracted since, given the corpus dimensions, their occurrence was inferior with respect to the predefined threshold value. Consequently, we decided to generalise the extraction, even though this process required a more extensive activity of manual cleaning of the terminological lists.

<table>
<thead>
<tr>
<th>Total doc.s</th>
<th>Total terms extracted</th>
<th>Total single nouns</th>
<th>Total compound nouns</th>
<th>Total manually extracted terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>6893</td>
<td>2504</td>
<td>4389</td>
<td>1623</td>
</tr>
</tbody>
</table>

Table 1: Terms extraction statistics

3.2: Thesaurus construction
With regard to the thesaurus construction, we complied with the standards of the field: ISO 2788 (1986), relative to the construction of monolingual thesauri, and by the more recent ANSI NISO Z39.19 (2005) related to the various typologies of controlled vocabularies.

The approach followed to realise a faceted classification included, as first writing of the thesaurus, a systematic presentation, where relations between terms are pointed out by graphical expedients. While in subsequent alphabetic presentation, realised using the MultiThes software <www.multithes.com>, specific for the construction of thesauri, the relationships established between each term and the other terms contained in the controlled vocabulary have been highlighted. Nevertheless, we can also find the alphabetical order in the systematic presentation, where it is used in order to arrange the terms of each facets or subfacets. Anyway, the two presentations will be considered complementary, especially with regard to consultation. Using the tool MultiThes we created both a systematic and an alphabetical presentation of the faceted thesaurus.

In the presence of synonyms we have chosen the preferred term according to the following sources: the ISO norms referring to the specific field, the “disciplinari di produzione” of some of Italian regions, and the terminological database realised and provided by the SSLMIT at Forli (Italy). Moreover, in order to obtain a better classification of terms, we have referred also to the already mentioned AAT thesaurus, hence containing some of the terms belonging to our vocabulary.

Facets have been chosen from those defined by the Classification Research Group (CRG), which revised and extended the five facets of Ranganathan: Personality, Matter, Energy, Space and Time (Broughton 2008).
By means of this faceted system we put into relation also terms referred to the specific domain of Calabrian goldsmith with those used in the more general context of Italian goldsmith. In fact, this faceted system allowed us to consider the regional terms as a local specification of each facet included in the thesaurus. So, each facet, such as techniques, materials, products, tools, includes a sub facet named place that contains all terms extracted from the transcriptions of the expert’s interviews and that could be categorized according to that facet. It is important to specify that the criterion followed for this choice is purely linguistic and statistic.

The GoldThes faceted thesaurus will be finally associated to the XML knowledge model generated by PCPack and transformed into a consultable product. By doing this the access to the knowledge base will take place through a closed terminological list, so that information may be accessed through different points. By adopting such a system it is possible to maintain the terminological individuality of each goldsmith. In fact, the specificity of the domain and the relative lexicon showed that often there are

<table>
<thead>
<tr>
<th>Thing/entity</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts designating the principal objects of study in a topic or in a discipline</td>
<td>Ex. Embossment</td>
</tr>
<tr>
<td>Kind</td>
<td>Production of techniques</td>
</tr>
<tr>
<td>Concepts indicating a relationship genre-specie with concepts defined in the category thing/entity</td>
<td>Ex. Problematic nature</td>
</tr>
<tr>
<td>Part</td>
<td>Properties of techniques</td>
</tr>
<tr>
<td>Concepts indicating parts of the concepts included in the category thing/entity</td>
<td></td>
</tr>
<tr>
<td>Property</td>
<td>Materials used in goldsmith, with the exception of those ones used in activities</td>
</tr>
<tr>
<td>concepts designating properties or attributes of the concepts included in the category thing/entity</td>
<td>Ex. Metals</td>
</tr>
<tr>
<td>Material</td>
<td>Actions not performed by human agents</td>
</tr>
<tr>
<td>concepts related to materials or substances of any type</td>
<td>Ex. Metal Oxidation</td>
</tr>
<tr>
<td>Process</td>
<td>Activities performed by human agents</td>
</tr>
<tr>
<td>concepts related to actions that happen in a spontaneous way, not performed by human agents</td>
<td>Ex. Alloys preparation</td>
</tr>
<tr>
<td>Operation</td>
<td>Concepts used in activities in intermediate phases of the productive process</td>
</tr>
<tr>
<td>concepts related to actions performed by human agents on an object</td>
<td>Ex. Models in wax</td>
</tr>
<tr>
<td>Patient</td>
<td>Products of activities</td>
</tr>
<tr>
<td>concepts indicating objects used in intermediate phases of the productive process</td>
<td>Ex. Earrings</td>
</tr>
<tr>
<td>Product</td>
<td>By-product</td>
</tr>
<tr>
<td>concepts indicating products of activities</td>
<td>Intermediate products of activities</td>
</tr>
<tr>
<td>Agent/Tools</td>
<td>Tools, People, Substances</td>
</tr>
<tr>
<td>concepts used to perform actions</td>
<td>Ex. Burin</td>
</tr>
<tr>
<td>Space</td>
<td>Places</td>
</tr>
<tr>
<td>concepts related to places</td>
<td>Ex. Laboratories</td>
</tr>
<tr>
<td>Time</td>
<td>Concepts related to time</td>
</tr>
<tr>
<td>Concepts related to time</td>
<td>Styles, Periods, etc.</td>
</tr>
<tr>
<td></td>
<td>Ex. Byzantine style</td>
</tr>
</tbody>
</table>

Table 2: CRG categories and corresponding concepts of our domain
terminological incongruities, therefore to the same object, carrying out the same function, different terms are attributed or, in the event in which a different function is recognized, two different objects are attributed with the same denomination.

4: Conclusions and future works
In conclusion, the obtained result is a methodological structuring that can certainly be applied to the different contexts in which it is necessary, for various exigencies, to formalize tacit and consuetudinary knowledge in dimensionally modest domains, but with high added value. All this can be seen both in the light of capitalizing productive techniques and the creation of training programs able to guarantee business inheritance. According to this point of view, we are evaluating the feasibility of the creation of a multilingual thesaurus, at least for English and French, which would provide symmetric and equal structures for all the linguistic versions carried out. To improve the results of this work, another important prospective is the possibility of extending our GoldThes into a domain ontology, to allow the semantic enrichment of the information search.

References
Trigari M., 1992, Come costruire un thesaurus, Panini.

Web documents have been accessed 18 September 2009.