68th IFLA Council and General Conference
August 18-24, 2002

Code Number:	**097-144-E**
Division Number:	**IV**
Professional Group:	**Classification and Indexing with Cataloguing - Workshop**
Joint Meeting with:	**CILIP Cataloguing and Indexing Group**
Meeting Number:	**144**
Simultaneous Interpretation:	**-**

Libraries and archives: integration of two professions in creating a framework for a thesaurus/classification in Italian universities

Giovanna Granata, Gianni Penzo Doria & Zanetta Pistelli

Universita degli Studi di Padova
Padua, Italy

The attention that archivists have devoted to current archives is in no way comparable with that they have given to historic ones. The litmus test is not so much insufficient attention to classification in archival studies as the absence of editorial criteria.

The framework for classifying, the true act of organizing by subject, must be edited according to scientific criteria, so as to avoid reducing it to a banal listing of offices or organizations which exchange administrative correspondence, and above all a clear distinction between the logical and physical organization, not only of the documents but also of the organizations producing them must be drawn. It must also take account of the fact that classification is a technical activity which, in the hands of unqualified practitioners can create conflicts and lack of certainty in its proper application.

In order to reach a degree of standardization, Italian universities collaborating on project "Titulus 97" (http://www.unipd.it/ammi/archivio/tit_idx.htm) have, since 1997, been working to develop a common classification system, accompanied by an index, so as to obtain a flexible and efficient means of orientation in the classification scheme: the "terms" identified from the documents have been arranged after the lead term and manipulated by rotating the significant terms. This rotation does increase the possibilities of access, but it is accompanied by a loss of intelligibility. It was immediately obvious that the least ambiguous and most controlled terms must be used and yet we were convinced that collaboration between the two professions, librarians and archivists, was a significant bonus, even granted the distinction between their roles and methods.

So we have worked on the side of syntax which as far as terminology is concerned presents many problems and reveals a great variation in the construction of terms. In fact, the terms which occur have a specific character, typical of bureaucratic terminology, full of double meanings and colloquial terms to denote procedures and actions. This language is already very complicated in itself, since it relates to complex objects (standards, structures, resources and reports from outside) within a complex organization. Above all, it is used to describe not bibliographic material but administrative documents and the difference is significant. If it is true that the object about which, according to the rule, the actual document carries information may be compared to the title of a bibliographic document it has also to be conceded that with such a document one can rely upon the tried and tested bibliographic practice that, apart from obscure imaginary titles or those that are clearly misleading, the title contains, in summary form, the central theme of the document. Contrariwise, administrative documents, far from having an adequate formula and standardization, often convey their content by specific and minutely detailed references with details which are clustered or fit into a central core in a manner similar to that which can be seen in the texts of laws or the syntax present in a stream of subordinate clauses.

Thus, one has to reformulate the terms and transform the statements to a true condensed form for indexing, with a rigorous citation order and a clear and consistent controlled vocabulary.

In a totally contrary manner to the complexity of these condensed formulae another problem occurs. When creating an index from the schedule of a classification, terms are used in a distinctive manner, different from common practice, and peculiar to the condensed format of the subject matter. This is shownin countless words - for example, "regulations" - consisting only of substantives which clearly group themselves with documents relevant to different, connected situations, irrelevant of the individual significance of each. In these instances, the point of reference is useful to us, forming a title and convenient class. For example, for terms such as "picture frames" or "damages" only references to the titles and the relevant class clarify what is happening, for the first the acquisition and trimmings of picture frames and for the second the disputes subsequent to damage caused to people or to the rights of the university.

Once the subject is reconstructed, it has to be expressed.

By what syntax and according to what principles?

The very first point of reference has to be the "Subject List" of the Biblioteca Nazionale di Firenze, which always has been submitted to many reviews the chief of which is an agent typical of the paper age, in which the efficiency of the recovery of information is left to the significance of word order. This occupies the foremost position in the string and imposes a citation order based on the assumed importance of terms, to ensure direct access. This makes it difficult to express subjects that present different ideas combined together for which, however, it becomes necessary to put in place more than one string, none of which actually comprises the total significance of the subject.

Our choice then is based on an indexing method devised by GRIS (Gruppo de Ricerca sull'Indicazzazione per Soggetto) a method which not only permits the expression of strings appropriate to cover the total meaning of the subject, but still maintains the citation order according to a coherent and rigorous intelligibility. In its development GRIS resorted to studying the theoretical foundations of subject analysis developed principally in England by the CRG (Classification Research Group) and culminating in the PRECIS system. In the Biblioteca Centrale di Firenze a project to revise the "Subject List" in the light of the GRIS methodology is in process.

Following the rules of the GRIS guide[1], to which we revert for a more analytical presentation of objectives and methodology, we have consequently proceeded to identify the role of each concept in the descriptions: action, agent, object, etc., and we have then arranged the terms in the string according to the citation order assigned to each role and according to the criterion of contextual dependence. We give a practical example of the methodology used below. Let us consider a statement of the subject present in the index of a classification schedule, such as: "Mod. 101, predisposition, educational personnel". It involves an apparently simple string, in which one can identify an action (predisposition), an object (mod. 101), and a beneficiary (educational personnel). In reality, the significance of "mod. 101" is not at all clear; it involves a form necessary for the declaration of income for teachers. At this point, one can identify two concepts, each in turn decomposable in subsequent concepts: the first is the "Predisposition of mod. 101", the second "the declaration of teachers' income" On the basis of the analysis of roles, the citation order will be as follows:

Teaching personnel - Revenue - Declarations - Formulary : Mod. 101 - Predisposition

On account of the "denomination" of the origin the advantages gained are better intelligibility of concepts, conveyed in the index to the classification schedule in a synthetic and colloquial form; clarification of the roles of semantics and syntax; access to all the terms without rotation or inversion.

A problem which we posed for ourselves, especially with very complex statements, is the rigid analysis in the identification of roles and in the expression of equivalent terms which determine a compartmentalization more extensive than the strings according to customary natural language. For example:

Doctorates of research - Co-operation - Board of examiners - Nomination - Communication with organizers of the doctorate

Without doubt the strings reconstituted in this way because a prearranged tool is used and controlled by archivist colleagues, can create greater problems than practical advantages for easy application: natural language, even if it is subjected to lexical contortions so as to achieve the most likely retrieval of information, is much more direct than a controlled and greatly formalized language.

In sum, using a coherent syntax for the construction and comprehensibility of strings seems to be a major advantage to us. Indeed, if the classification schedule, and consequently its reference index, is destined to be a useful instrument for all Italian universities which collaborate in project "Titulus 97", totally standardized, controllable, updateable according to clear and predictable criteria, tools are essential for successful co-operation.

Side by side with the work of creating strings of subjects there is the task of controlling terminology: what wordlist should be used? Actually, in the "denominations" used like the index to a classification schedule, despite the legal and administrative nature of the lexicon it lacks the elements of control, giving way to the proliferation of a variety of synonyms or quasi-synonyms, often used in an incorrect way, for example, personnel, subordinates, workers, etc. Sometimes the exact meaning of the sense of the "denominations" was handicapped through the use of homonyms, for example "University diplomas" which in Italy means a sort of course lower than Masters' level, as well as a document certifying the conclusion of a course of study: masters' degree, doctorate. Finally, another problem arose from the use of made up expressions typical of bureaucratic language, for example "periods of work in the public service"

[1] Associazione italiana biblioteche. GRIS-Gruppo de ricerca sull'indicazzazione per soggetto. *Guida all'indicazione per soggetto*. Roma: A.I.B., 1997.

or "payments for litigation" which present difficulties of syntactical control on the terminological level: from this point of view, the sparseness of choice plays its part in the choice of the most simple terms and of co-ordinating them at the time when subjects are formulated.

Because of the complexity of problems, the definition of vocabulary cannot simply lead to a uniform list but one has to recognize from the outset the need to use a thesaurus to structure the terms in respect of all their reciprocal inter-relationships. The co-operative context of Titulus 97 brought another stimulus to this decision. On the one hand the co-operation of different organizations made it even more essential to create clear and exact tools for vocabulary control; on the other, it guaranteed a strong support for updating which is one of the principL difficulties inherent in choosing a thesaurus.

For the construction of the thesaurus we had to consider starting from scratch, at least for the actual terminology, or making use of existing tools. In this regard, certain well known examples, emanating from the educational field (Eurovoc)[2] or from the legal (TESEO)[3] were looked at. However, the choices of a disciplinary nature did not work very well for the semantic fields covered, owing to the level of specificity needed. A more useful model was the TRT (Thesaurus regionale toscano)[4] drawn up on faceted principles, using general and abstract facets, based on the criteria proposed by the same GRIS which guided us through the problems of syntactical control.

According to those criteria, the grouping of descriptors in facets was effected by the TRT on the basis of the same sorts of concept, for example activity, agent, etc. independent of their specific context of use. As a result, it provides as objective a method of organization as possible which is well adapted to complex terminology and is also easily kept up to date. Additionally, although it was developed in a library environment, it was created in a context which approximates to our terminology: a legal and administrative type of library which supports the activities of the organization "Regione Toscana". It was with this aim that the facets proposed in the abstract by GRIS were more analytically determined to handle the terminology in question. These are: Legal Acts, Activity, Conditions, Disciplines, Forms, Objects, Organizations, Peoples and groups, Process, Place, Agents, Structures, Time, Theories and movements.

Once the choice of the general structure of the TRT had been made as a point of reference, the next task was to ascertain hospitality with reference to the terminology of the "denominations".

Here, we tried to determine the descriptors first, using as an authoritative guide for lexical problems the choice already made by the TRT, the standard ISO 2788:1986.[5] The principles were rigidly applied, always taking into consideration the inherent problems of bureaucratic language; we were paticularly concerned with compound terms which, as we have seen, abound in the index of a classification schedule and we have sometimes chosen to use "accepted" terms without further reducing them to simple expressions, so as not to lose their specificity.

When making a choice between the singular or plural form we adopted the criterion of "countability" following the lines offered by the GRIS guide and followed by the TRT: we used the singular form for terms involving an activity and plural for "count nouns"; in the latter case, however, because of the

[2] Commission of the European Communities. *Thesaurus Eurovoc*. 3rd ed. - Brussels: CECA-CE-CEEA, 1995.

[3] Italy. Senato della Repubblica. *TESEO: TEsauro Senato per l'Organizzazione dei documenti parliamentari.* 3rd ed. Rome: Senato della Repubblica, 1998.

[4] Regione Toscana. *Thesaurus regionale toscano.* Florence: Edizioni Regione Toscana, 1996. A more up to date version may be consulted at http://www.regione.toscana.it/ius/ns-thesaurus

[5] ISO. *International standard ISO 2788: documentation: guidelines for the establishment and development of monolongual thesauri*. 2nd ed. n.p.: ISO, 1986.

semantic uncertainty between the use of singular and plural we decided to retain both forms, singular for activities and plural for the result of those activities: laws, documents, etc.

For the treatment of proper nouns, that is to say the names of organizations, foundations, consortia, projects, administrative assessments, we opted for an authority list outwith the thesaurus rather than including each within the hierarchy of the thesaurus, given the great variety of descriptors for the universities involved in the project Titulus 97. On the other hand, we put into the thesaural hierarchy the term "common": each university has a different name for its own undertakings, but they all, really, perform the same function.

Inserting terms into the facets of TRT gave good results showing the hospitality of its structure, but the expansion of terminology for university activities in some instances created certain problems given the extreme specificity spelt out in the facets such as "Peoples and groups", "Organizations", "Activity". This lack of homogeneity does not, however, bring into question the general structure of the thesaurus in the way that one might imagine the development of a dictionary could, thanks to the updating of the vocabulary.

As far as the management of relationships is concerned, we resorted to a program called "Beat". Worked out by Josep Sau of the Informatics Centre of the University of Barcelona, it is available free of charge for non-commercial use at the address http://www.willpower.demon.co.uk.thessoft.htm#BEAT from where you can download it. " Beat" sustains the creation of all the relationships recognized by the ISO standard mentioned above: equivalence, which makes the reciprocal from the rejected term but retains it in the thesaurus as a key to access from the accepted one; hierarchical relationships, which connect each accepted term with its superordinate one in the structure of the thesaurus; associative, which signals for every term those which are semantically connected by the thesaurus.

Finally, as far as results are concerned, "Beat" provides different types of output: systematic, alphabetical or permuted.

To use the thesaurus, beyond the hierarchical display, essential for creating thesaural relationships, we thought we ought to offer an alphabetical arrangement in which every term appears with the totality of its relations.

To avoid rotating terms in the strings prepared by archivist colleagues in the first phase of creating the index of "denominations" , we associated each term in the thesaurus's alphabetical list of subjects as it occurred, so as to allow users to access each subject, expressed exhaustively and entirely intelligibly, from each of its elements.

To guarantee reciprocal visibility for terms and subjects an ad hoc program was devised which connects the one with the other.

In the final analysis, for every descriptor in the alphabetical output, the strings are also indicated in their hierarchical position., The strings, in their turn, relate to the title and to the class in which they occurred, so that the thesaurus and the list of titles are not independent of one another, but it is possible to move from one to the other maintaining either the syntax or the semantics.

This allows users, as well as having a guide to the introduction of new strings by analogy with those already existing, without duplication, to effect control from relevant descriptors.