Authority control and subject indexing languages

Stefano Tartaglia
AIB - Gruppo di ricerca sull'indicizzazione per soggetto

Analogously with natural languages, artificial languages with which the mediation between a collection of documents and its potential users takes place are also made up, as Elaine Svenonius has often explained, of four necessary components:

> vocabulary, that is, all the elementary expressions used to define individual entities, attributes and relationships
> semantics, that regards the meaning for which a determined expression is included in the vocabulary and used in the language, that is, the relationships which are derived a priori from that meaning and which define it
> syntax, that regards the composition, by means of the structure of individual elements taken from the vocabulary, of more complex expressions
> pragmatics, which concerns the conditions and modalities for language application.

Authority control, or with the term vocabulary control, which Elaine Svenonius considers to be equivalent, involves the first component, vocabulary, and comes about in the second, semantics[1].
This analytical approach to linguistic systems of "cataloguing" mediation (meaning the adjective in a general, non just a library science, sense) provides the most correct theoretical reference for a series of considerations, from which the properties and functions of authority control should result easier to recognize and consequently once clarified the relationships between authority control and the different cataloguing languages, to mark the subject indexing languages.

1. Authority control can not be identified, nor is it advisable that it be identified, with a particular cataloguing language, since syntax and pragmatic components are extraneous to authority control, which vice versa more than the vocabulary and semantic components characterize and diversify the single languages. This independence in relation to single cataloguing languages, or to particular types of language, constitutes a fundamental property and strong element of authority control. Its efficiency must be preserved and widely extended, with the punctual exclusion of the procedures of vocabulary control of those elements and aspects which, being distinctive of single languages, result reciprocally incompatible and thus do not lead to a process of unitary control. The origins of authority control are easy to locate in the paragraphs of the rules for bibliographic cataloguing by author and title, but it is now time to completely abandon the idea that authority control is a phase in the application of a particular cataloguing language, in order to project and realize a system of an unique and general authority control, which satisfies not only the needs of systems of truly bibliographic mediation, but also of systems of, among others, archival and museum mediation. The more it is general and international, the more vocabulary control is efficient, since it permits the reduction of the waste of resources produced by the repetition of the same operations of control in different places, times and contexts. That is something particularly appreciated in an epoch in which everything is aimed at economizing even on intellectual work. Especially, however, the more it is unique and general, the more authority control is efficient, since it furnishes the unifying element between different cataloguing systems, indispensable for the realization of instruments, logicians

---

[1] *The intellectual foundation of information organization* / Elaine Svenonius. - Cambridge (Mass.) ; London : MIT Press, c2000. - (Digital libraries and electronic publishing). - In particular p. 53-58.

and informatics, which facilitate, and make coherent to the utmost, the research for information in documentary spheres which are traditionally separated, as for example, that of collections in libraries, archives, museums and those in digital formats accessible only by telematic means[2].

2. Cataloguing languages are different, first of all, because of diverse pragmatics, that is because they are destined for use in distinct circumstances with distinct functions and modalities. But, as mentioned above, authority control does not concern pragmatics. This means that the conditions and relationships which depend upon the material, formal and substantial characteristics of the single objects to describe and index, the various contexts in which such objects are described and indexed, the particular purposes for their description and indexing, and which determine the use of specific cataloguing languages, at any rate do not determine a diversification of authority control, either regarding the process or the product of such a process, which will be an univocal normalized expression inserted in a web of relationships.

Inserting this consideration into the specific theme of this paper, it is impossible not to observe that the conviction that authority control is distinguishable, and must be distinguished, without ulterior specifications is still widespread, and that it is commonly held regarding "authors". This conviction, which is certainly involuntary, and thus not justified, in the normative activity of the IFLA, which on this subject published two distinctive guidelines, one which was formally presented as relating to the "authority records" (formerly "authority entries")[3], and one which was formally presented as relating to the "subject authority entries"[4] Moreover it is also easy to establish that such a persistent conviction has determined, although with non identical consequences, the same organization of two of the most important national systems of vocabulary control, that of the Library of Congress and that of the Bibliothèque Nationale de France[5]. In the search screens the Library of Congress authorities are divided into subject authority headings, name authority headings, title authority headings and name/title authority headings; nevertheless, searching for the same expression either as name, title or name/title or as subject the authority record is retrieved. This demonstrates that the authority file is unique, that the control operations relating to the same expression are carried out only once, and that it is simply considered necessary to adjust the appearance of the control system to the conviction held above, making the "subjects" seem separate from the rest of the entries (and here we are speaking about access to the authority file, not about access to the Library of Congress catalog, where separation of the indexes is obviously appropriate). The *Notices d'autorité* of the BNF are analogically presented separated into *personnes physiques*, *collectivités*, *titres uniformes* and RAMEAU (*Répertoire d'autorité-matière encyclopédique et alphabétique unifié*) effectively divided (frBN000 for persons and corporate bodies, frBN002 for titles and frBN001 for subject entries). This means that at the BNF all the management and control operations for the same expression are duplicated, with a waste of resources which does not correspond to an increase of visible efficiency in the system, and which would be sad if it were motivated only by the obligation to conserve some, slightly justifiable formal differences between expressions which are equivalent (the most evident case is that of the dates of birth and death, not normally stated in the terms given in the *fichier d'autorité personnes physiques* - which proposes, for example, the preferred form Dante Alighieri - but are present in the RAMEAU *vedettes* - where the preferred form is Dante Alighieri (1265-1321).

---

[2] P. G. Weston, indicating this cataloguing connection between different collections of documents, uses the quite appropriate expression "interoperability among heterogeneous systems"; in *Il catalogo elettronico : dalla biblioteca cartacea alla biblioteca digitale* / Paul Gabriele Weston. - Rome : Carocci, 2002. - p. 28.

[3] *Guidelines for authority records and references*. - 2nd ed. / revised by the IFLA Working Group on GARE Revision. - Munich : Saur, 2001. - (UBCIM publications ; n.s., 23). - Title of the previous edition: *Guidelines for authority and reference entries*.

[4] *Guidelines for subject authority and reference entries* / Working Group on Guidelines for Subject Authority Files of the Section on Classification and Indexing of the IFLA Division of Bibliographic Control. - Munich [etc.] : Saur, 1993. - (UBCIM publications ; n.s. 12).

[5] Respectively accessible at URL http://authorities.loc.gov/ and http://noticesautorites.bnf.fr:8095/.

These solution differences and uncertainties in the authority control systems, from which can be derived only problems and waste, beginning with the error of considering the "subjects" as semantically distinguishable entities, on the same level as persons, corporate bodies and works. It is therefore indispensable to clarify that it is impossible to identify an entity as a subject if not in relation to the to pragmatics, that is, to the concrete circumstances which have implied in such a case the use of a subject indexing language in the cataloguing record of a document. However since authority control does not regard those circumstances, it is not theoretically correct, nor practically useful, to perceive and practice a subject authority control as a process in itself. In the FRBR model[6], which in reality is also reflected in the *Guidelines for authority records and references*, are the classes of entities, as noted above, *work, expression, manifestation, item, person, corporate body, concept, object, event* and *place*. These entities "represents the key objects of interest to users of bibliographic data"[7], and each one having a formal expression as the primary necessary attribute which designates it (title, identifier, name or term), can be subjected, and it is potentially opportune that they all be subjected, to the authority control procedures. The "subject" is not an entity in the FRBR model, it is a relationship (the relationship "has as a subject"), which does not rigidly correspond to single classes of entities, but which can involve vice versa entities belonging to each of the classes proposed in the model.

It was certainly not indispensable that FRBR remind cataloguers that the name of a person, the corporate body name or the title of a work can be inserted in a cataloguing record in order to realize different functions (typically, author or title access and subject access). However, FRBR goes further since it suggests the possibility that even the terms which represent these entities (*concept, object, event and place*) for which only the subject function is indicated can be linked to a cataloguing record with a diverse function. If in fact "typically the user will formulate a search query using one or more attributes of the entity for which he or she is searching, and it is through the attribute that the user finds the entity sought"[8], and if, for example, the second and most important attribute of the work, after the title, is the "form", that is "the class to which the work belongs"[9], how is it possible to allow an efficient search, using this attribute, of a work concretely expressed in a manifestation if not designating this attribute with a controlled term, which cannot but be the same used, in the context of other cataloguing records, for subject access? Putting it very simply, if Walter Scott's *Ivanhoe* is an historic novel, how is it possible to retrieve, by literary genre, the various manifestations of this novel if not directly or indirectly linking the relative cataloguing records to the term "Historical novel", thus inserting into the search mechanisms of the editions of a work a term presumably also used in the subject indexes, but which in this case does not express the subject of the work? It is true that above all in our catalogs until now, we did not bother to activate retrieval requirements for "genre" too much, but this has been only a convenient default, seeing as how Cutter already, over a century before the FRBR, had indicated the search by literary genre as one of the requisites of the catalog. If we then move from the bibliographic context to other cataloguing contexts, the use of potentially expressive subject descriptors in order to indicate something else becomes quite prevalent. "Etruscan ceramics" for example, is a term for which the function of a subject is typical in a bibliography, but probably is to be excluded in the catalog of a museum collection, in which "Etruscan ceramics" will indicate what the indexed object is, not its theme. That of the "form of the work" in FRBR is the most relevant, but not the only case of the potential use of an entity *concept, object, event or place* for the purpose of indicating not the subject, but another characteristic of the indexed and described document.

[6] *Functional requirements for bibliographic records* / IFLA Study Group on the Functional Requirements for Bibliographic Records. - Munich : Saur, 1998. - (UBCIM publications ; n.s., 19).

[7] *Functional requirements for bibliographic records*. - p. 12.

[8] *Functional requirements for bibliographic records*. - p. 56. Compare on p. 30: "Each of the entities defined in the model has associated with it a set of characteristics or attributes. The attributes of the entity serve as the means by which users formulate queries and interpret responses when seeking information about a particular entity".

[9] *Functional requirements for bibliographic records*. - p. 33.

Absolutely nothing in the form and the meaning of a term can predetermine the function that the term will play in the single cataloguing record and consequently in a particular index, above all if from the bibliographic context we widen the field to other cataloguing contexts. On the other hand however, to be efficient, unique and general, authority control must leave that function out of this consideration. This further confirms that there is no theoretical or practical justification for continuing to distinguish a subject authority control in the general activity of vocabulary control. It is not by chance that IFLA, having produced the two guidelines on authority records quoted above, has nevertheless produced an inevitably unique guideline for the codification of authority records in a readable format for the machine, that is to say *Unimarc. Authorities* in which a specific field for all subject entries alone as such do not exist[10].

3. Semantics is pertinent in authority control, in the three forms described by Elaine Svenonius: referential, relational, category. Referential semantics concerns the uniqueness of meaning (unambiguousness) of each of the elementary expressions included in the vocabulary: it is the task of authority control to guarantee this uniqueness, supplying each strictly indispensable and sufficient expression of all the formal elements to impede any misunderstanding in use and interpretation.

The potential polysemy of an expression is a frequent phenomenon in natural languages, which has no effect on communication, since normally pragmatic circumstances clarify the meaning for which the polysemic expression is effectively used. But a general authority control, although it is functional in most cataloguing contexts, cannot delegate the referential semantics to the clarifying properties of a single context, nor be conditioned by those properties, and must rather satisfy completely and alone the needs of unambiguousness of the expressions of the vocabulary in all the cataloguing contexts. In this regard, the case that immediately comes to mind is that of those proper names, often called "geographic", which, in common use, ca indicate both a territory (that is, in a narrow sense, a portion of the earth's surface: "the climate in Italy"), and the population of that territory ("the economic relations between Italy and France in the 16th Century"), as well as the corporate body that holds sovereignty over that territory ("the international relations of Italy in the period after the second World War"). In some contexts these proper names do not result poysemic: in the context of the author cataloguing "Italy" is not effectively polysemic, since the meanings of the territory and of the population are not pertinent to the context; while, certainly but not alone in other contexts, of the subject indexing, those proper names are very polysemic, to the point of sometimes making syntactic relationships ambiguous and the comprehensive meaning of the entries in which they are inserted doubtful, as well as often making the search and selection of the cataloguing information fastidiously heavy. To be efficient, general authority control must resolve the problems of real polysemy even when not common to all the cataloguing context, and must guarantee the unambiguousness of the expressions in any context in which they are used.

The relational semantics concerns the relationship of meaning between all the expressions included in the vocabulary. Among these relationships one is unquestionably considered pertinent to any form of authority control, and that is the equivalence relationship: if in various cataloguing contexts two or more expressions can be used with the same meaning, that is if they indicate the same entity, these two or more expressions are correlated in the authority file, in such a way to guarantee the individualization of that entity by means of any of these expressions. Traditionally the principle of uniformity was applied to the equivalence relationship, for which one of the expressions was designated as preferred, and became the only one effectively present in the cataloguing records. The informatics management of authority control can offer alternatives to this solution, which allow us to realize a unique and general authority control without imposing the use of the same expression in all contexts, but it must be kept in mind that uniformity has its value as an element of coherence and

---

[10] *Unimarc manual. Authorities format.* - 2nd rev. and enlarged ed. - Munich : Saur, 2001. - (UBCIM publications : n.s., 22). - Tit. of the 1st ed.: *Unimarc. Authorities*.

of predictability in catalogs, and that thus it is opportune to give it up only in the presence of real and fundamental cataloguing needs.

The control of equivalence relationships, however, does not exhaust the needs of semantic correlation of any cataloguing language. Author and title bibliographic cataloguing imposes control of some cases of associative relationships in point (between denominations of corporate bodies, distinct bibliographic identities corresponding to the same physical person and works[11]) and of hierarchical relationships (between a corporate body and one of its organs, a work and one of its parts). The opportunity of controlling this type of relationship can only increase with the application of the FRBR model. Extended to all categories of entities, the control of associative and hierarchical relationships represents a fundamental need for the subject indexing languages, but it is presumable that it is indispensable or useful for any other cataloguing language too, and certainly advantageous for the interconnection between different languages. It is thus appropriate carried out on the level of general authority control. Obviously on this level of control only hierarchical and associative relationships of universal validity may be expressed, that is those explainable by the typical and essential definition of the entity submitted to authority control, and which thus cannot be contradicted in any particular context. Any entity can then be involved in semantic relationships that are not universally valid, but inherent to the context of application of a determined language. Such relationships must not be expressed on the level of general authority control, but in specific and sectorial vocabulary and classification control instruments, for the elaboration of which the existence of a general semantic control instrument does not necessarily constitute an obstacle, rather a facilitation.

Substantially a semantic relationship structure is a classification structure, and thus needs a primary general criteria of subdivision, upon which ulterior articulations can be based. General bibliographic classifications are based upon a primary articulation by subject, which implies the possibility that the same concept, if pertinent to more than one subject, belongs to several classes. This does not correspond to the needs of characteristic semantic structuring of authority control, since, as has been repeated several times in this paper, such a process must leave out of consideration the single context, even when disciplinary, in which the single entities are concretely quoted. With a view to authority control it is therefore necessary to adopt another general criterion of classification which is independent of single contexts, but is valid and useful in all of them and has a secure application. More than a century of studies and of practice in indexing has shown that the criterion of categorical analysis which consists, as regards semantics in the individuation of a limited number of general and universal semantic categories answers these requisites, so that each entity can belong to only one category and in the assigning of each entity, for the single attributes which constitute the typical definition, to one of the individuated categories[12]. The individuation of semantic categories is necessary to avoid incongruities and contradictions in the identification and expression of semantic relationships, but it is also more generally necessary for the normalization and the informatics management of authority control. It is certainly not a case that upon the citation of general semantic categories is based all of the above mentioned *Guidelines for subject authority and reference entries*, nor is it a case that every field of *Unimarc. Authorities*, excluding those of the notes and equipment information, correspond to a semantic category. In the exercise of this general categorial semantics the authority control is however not alone. Even in the FRBR model, in order to control other types of cataloguing relationships, some general semantic categories have been individuated (*work* etc., *person, corporate body, object, event and place*) principally usable even in authority control. This shows a convergence of needs and solutions which induce one to

---

[11] *Guidelines for authority records and references.* - p. 17-19.

[12] *Guida all' indicizzazione per soggettd* Associazione italiana biblioteche, GRIS Gruppo di ricerca sull'indicizzazione per soggetto. - Repr. with corrections. - Rome : AIB, 2001. - p. 60-63. An examination of the functions and requirements of categorization is also in: *Per un nuovo Soggettario : studio di fattibilità sul rinnovamento del Soggettario per i cataloghi delle biblioteche italiane* / commissioned by the BNCF from IFNET Florence ; a study undertaken by the Gruppo di progetto per il rinnovamento del Soggettario. - Milan : Editrice bibliografica, 2002. - p. 328-332.

hypothesize that the full actuation of the FRBR model make a unique and general authority control realizable, but that even a unique and general authority control is indispensable for the full actuation of the FRBR model.

4. Authority control is applied to the single units of vocabulary regarding a priori their characteristics and relationships, those which are constantly valid and for which the single unit is admitted in the comprehensively understood linguistic system. That is, authority control concerns the paradigmatic, not the syntagmatic, side of linguistic communication, and does not thus a posteriori regard the relationships which depend upon connotations of objects to describe and index, and which exist between the unit of vocabulary only when and since both are present together in a determined statement. Thus the syntax which is the representation of these relationships through the ordering in sequence of correlated units, and which is one of the most distinctive elements between diverse languages cannot be object of a real authority control, but can be object, when necessary, of only a correction verification and of substantial uniformity in the area of the application of a particular cataloguing language.

Understanding the syntax of authority control involves a series of inconveniences, which do not appear to be compensated for by any particular advantage. Above all, it impedes the realization of a unique and general vocabulary control which is a connecting factor between different cataloguing languages. This is because syntax is the element for which less compatibility is shown not only between typologically distinct languages, but also between kindred languages which could easily share the same vocabulary. The cases of subject indexing languages are seen as emblematic. The differences between various languages which express a theme in pre-coordinated form, and of those with post coordinated languages, are not so many differences in terminology, but are rather differences in the grade and form of expression of the syntactic relationships, which sometimes makes it impossible to establish the equivalencies not only between index entries expressed by languages which draw from different natural languages, as the MACS project[13] has shown, but even between entries expressed by languages which draw from the same natural language.

As an inevitably ulterior consequence, the authority control pretext of syntactic constructions adds a notable heaviness to the entire process to this contraction of its own area of efficiency, with a very useless increase. It does not determine a major expressiveness, since it is virtually uncontainable of the whole of the forms submitted to control, which contradicts the first requisite of controlled vocabulary which is to be such even regarding its growth. The constant expansion of the vocabulary is a need of cataloguing languages, but it has the condition that the gathering of new forms is always motivated by the need to express new unitary concepts as a general limit, not new concept correlations which can easily be expressed through the combination of pre-existing forms. To create a new authority record for the string Hospitals - Administration - Data processing - Evaluation - Computer programs (*Library of Congress authorities*, sh 85062292), when authority records already exist for each of the single terms that compose it, it is not in any way useful to the expressiveness of the language, but obliges us, at least in theory, to the fulfillment for this entry of all the procedures of control and of semantic correlation. Being in this case particularly articulated and heavy, such procedures were not in the least way effected (the entry in question has no link with other entries), which confirms the complete uselessness of this authority record, the creation and measurement of which has nevertheless absorbed resources. And here is one of the motives, certainly the most concrete and practical, for which the authority control of syntactically constructed entries must be avoided. The creation of authority records for this type of entry determines a waste of resources, which are subtracted from the more qualifying aspects of authority control, making the realization more difficult, if not impossible. Moreover, the problem is noticed even in the Library of Congress and the Bibliothèque Nationale de France since both institutions, in

---

[13] *Multilingual ACcess to Subjects*, a project of cooperation among four European national libraries for the construction and management of multilingual subject authority archives in English, French and German; an analysis of this project is in: *Per un nuovo Soggettario*. - p. 121-124.

order to avoid the collapse of their vocabulary control systems, are activating a review of the procedures aimed at limiting the application of authority control to only elementary expressions, at least to the extent that they are compatible with the use of indexing languages which remain substantially enumerative. Inverting the preceding procedure at the Library of Congress a new one has been established to avoid the creation of new phrase headings (for example, to Access for physically handicapped to recreation areas has been preferred the form Recreation areas - Access for physically handicapped) and the progressive transformation of many pre-existing phrase headings in entries with subdivisions is in course. At the same time the list of free-floating subdivisions has been notably amplified, with the attribution of the qualification of free-floating to many subdivisions which were not previously free-floating, which has offered the eventuality that for a new entry constructed combining pre-existing expressions it will be quite rare that a new authority record will be produced. Analogous interventions are also taking place in RAMEAU, although it is a more recent language conception (born after 1980) and thus less enumerative that LCSH[14].

When improperly extended to expressions constructed for syntax authority control results quite weakened and inefficient in its semantic component, which loses the necessary rigor both in individualization of relationships and in categorial analysis. To give a simple example, what is the entry immediately above, the BT of Hospitals - Administration? Hospitals? Administration? Health facilities - Administration? Health facilities - Organization? Hospitals - Organization? And what is the category to which the entry belongs? Is it the category of Hospitals given in the entry in the first position (but this is a merely formal criterion), or that of Administration, which is the focus of the corresponding syntax (administration of hospitals)? But above all, can such uncertain semantics ever satisfy the general needs of authority control? Or again, to ask a final question, does it make sense to corrupt to such a point authority control only to extend partially and occasionally syntax to an aspect of cataloguing languages, which is not pertinent to it?

In the articulation of preceding considerations, the sense of the relation between authority control and subject indexing languages is given, but almost hidden behind the closer examination of the single problems. It is well to reassume them here. The existence of subject indexing languages certainly does not call for or imply the existence of a particular authority control exclusively dedicated to subjects. Authority control, being an interconnecting element between different cataloguing languages, must concern all the categories of entities, and in such a way satisfy even the lexical needs of subject indexing languages. The use of subject indexing languages, as for other cataloguing languages, contributes to the individuation of new expressions to be submitted to authority control, expressions whose ordinary use can not predetermine their possible successive uses, and that thus, once completed the control procedures, can be used in any cataloguing context (bibliographic, archival, museums, etc.), either for indicating a subject or for indicating something else. Cataloguing languages, which include among others subject indexing languages, necessitate a verification of the expressions used which are not merely formal, but in addition more completely semantic. Such a verification, since it sticks to meanings and relationships of universal validity, is appropriate that it be effected at the level of general authority control. On the other hand, it is not opportune to extend authority control to syntactic constructions, because this reduces the rigor and efficiency of the control process, weighing it down until it becomes unfeasible, and impeding its function of unifying element between diverse cataloguing languages. Each cataloguing language, and thus each subject indexing language, possesses its own syntax and the verification of the correctness of its syntax can only be effected in the circumscribed context of the application of each language.

This conception of relationship between authority control and subject indexing languages is born from the meeting, already fertile, but which can become even more so, between technological

---

[14] The lines of the revision process of the two indexing languages result particularly evident in the respective updating newsletter: *Library of Congress subject headings weekly list* and *RAMEAU. Journal des créations et des modifications.*

progress and theoretic reflection. Informatics has made telematic interconnection between different documentary contexts, virtually simultaneous access to several contexts and deriving of data from one context to another not only possible but technologically easy. However a common logical linguistic instrument is lacking, and this is evident to everyone, but it is not the duty of informatics to furnish it. As concerns subject indexing languages the theoretic reflection has already individuated the principles which are functional in the elaboration of that instrument. Principles which in Italy have already been adapted, diffused and often clearly defined by the Gruppo di ricerca sull'indicizzazione per soggetto, and which can now find application in the project for renewing the *Soggettario*, promoted by the Biblioteca Nazionale Centrale di Firenze. Among the most important are: the principle of the subject as relationship property, a factor of coherence and pertinence in a text and no longer as a pre-existing entity supplied with a name; the principle of the separation between semantics and syntax, with the control of the first which can leave the proper formal structures out of consideration of the single languages, and thus be potentially unique and general, and the control of the second which is rather the actuation of those structures; the principle of the definable relationship, for which the semantic correlation of the vocabulary unity is exclusively founded on the essential definition of each unity which is universally valid.

Authority control being unique and general, and thus adequate for the situation created by technological progress, must make use of the most advanced theoretic reflection on subject indexing languages, and at the same time must satisfy the needs of subject indexing languages much more than it has until now.