

IFLA Satellite Pre-Conference of the Classification and Indexing Section  
 “Looking at the Past and Preparing for the Future”  
 Florence, Italy, 20-21 August 2009



## **Subject analysis and indexing: an “Italian version” of the analytico-synthetic model**

Pino Buizza \*

The tools and the tasks of indexing, of searching and organizing knowledge, are quickly changing. A crucial point is to maintain or to discover which are the essential features of a sound and efficient indexing system, the common characters of any shareable or at least communicable system.

The Italian research in the domain pays particular attention to theoretical aspects, in an autonomous way, not contrasting with other current systems. Indeed the future we are preparing is rooted in the shared past of common foundations: the analytico-synthetic model devised by S. R. Ranganathan for faceted classification, the outcomes of the Classification Research Group in the fifties and sixties of last century, leading to a verbal subject indexing system in PRECIS and to the ISO norms for subject analysis and thesaurus construction. We prefer to say “Italian version” of the analytico-synthetic model, instead of “Italian model”, because the model is the same, its inflexion is original. Not a local choice, but a contribution to the rethinking and rediscovering of well-known principles and practices.

To quote the title of our satellite meeting: the past we are looking at is not a shore we are leaving to ship towards a future promised land (a ‘lost’ past, whether nostalgically or gladly); it is a treasure we have inherited and wish to use, being sure of yielding fruits of it (a ‘living’ past).

The peculiar theoretical and methodological way, that the analytico-synthetic model of subject analysis and indexing have assumed in the geographic and linguistic Italian area, is well represented in two recent tools:

---

\* I’ll thank Alberto Cheti: the paper has been written in close collaboration with him and his contribution has been decisive.

- *Guida all'indicizzazione per soggetto* (Guidelines for subject indexing, 1996, 2001, in short *Guida GRIS*)<sup>1</sup>, drawn up by GRIS, Gruppo di ricerca sull'indicizzazione per soggetto (Research Group on Subject Indexing, of Associazione italiana biblioteche, the Italian Library Association),
- *Nuovo soggettario* (2006)<sup>2</sup> by the National Library of Florence, superseding and deeply renewing the old *Soggettario* (Subject headings, 1956)<sup>3</sup>, which is currently used by the majority of Italian libraries.

*Guida GRIS* develops the model by combining principles and methods from different settings into a general organic vision and into a consistent set of rules, applicable to any kind of general pre-coordinated indexing system.

*Nuovo soggettario* is an indexing system equally rooted in the analytico-synthetic model and consisting of (a) a set of rules, (b) a controlled and structured vocabulary (thesaurus), with full semantic relationships, and with syntactic notes enclosed, (c) the file of subject strings.

Only few notes about these tools, because the focus of the paper is on the underlying principles and the overall framework of Italian indexing, convinced that correct and strong bases are necessary to get high results.

## 1. The analytico-synthetic approach

Analytico-synthetic principles, criteria and rules are adopted in full in *Guida GRIS* and in *Nuovo soggettario*. The method is presented in short, step by step.

- *Conceptual analysis*. In the definition of the aboutness of the work, typical linguistic operations are adopted, such as deletion, generalization, selection and construction of concepts, in order to reach the base theme, the unifying intentional centre of all particular themes involved in the discourse. The logical functions of each concept included in the theme are analysed, following ISO 5963:1985. The output is the subject statement, a phrase in natural language.

---

<sup>1</sup> Associazione italiana biblioteche-GRIS-Gruppo di ricerca sull'indicizzazione per soggetto, *Guida all'indicizzazione per soggetto*. Roma, AIB, rist. 2001, <http://www.aib.it/aib/gris/gris.htm>.

<sup>2</sup> Biblioteca nazionale centrale di Firenze, *Nuovo soggettario. Guida al sistema italiano di indicizzazione per soggetto. Prototipo del Thesaurus*. Milano, Bibliografica, ©2006 (stampa 2007).

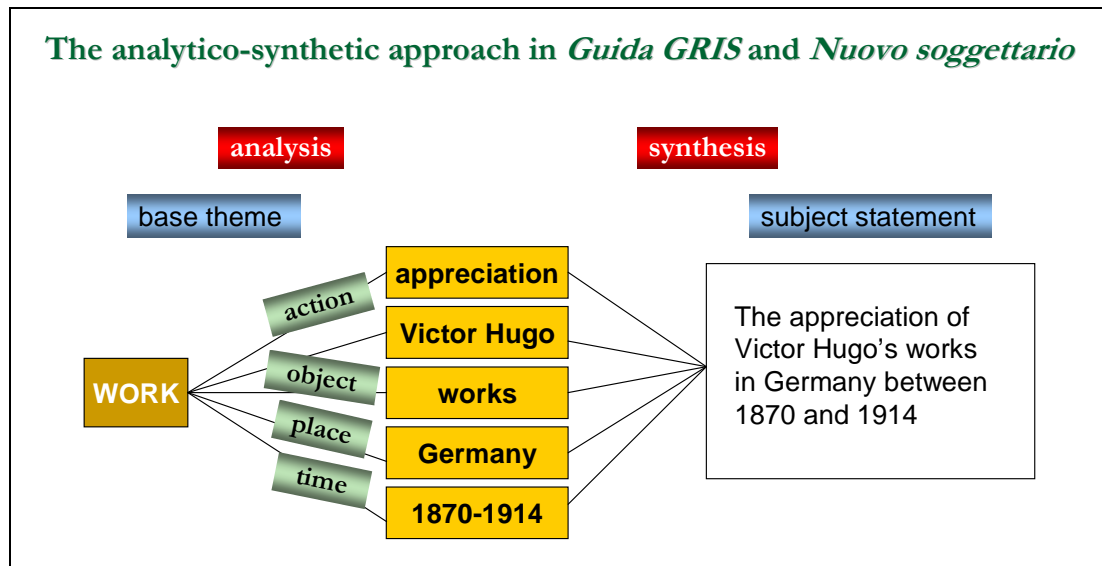
<sup>3</sup> *Soggettario per i cataloghi delle biblioteche italiane*, a cura della Biblioteca nazionale centrale di Firenze. Firenze, Stamperia Il cenacolo, 1956.

- *Syntactic analysis and synthesis*. The choice of co-extensiveness (from E. J. Coates) aims at representing the full theme in one string, no matter how complex it may be, supplying clear and complete information about the exact subject content of the work. The analysis of logic roles together with a scheme of syntactic roles (as it was earlier suggested by D. Austin and in PRECIS) forms the basis for the choice of the key concept and the citation order of the others. The output is the subject string.
- *Vocabulary control*. The choice of the preferred form for terms, the factoring of compound terms, the semantic relationships and the method for thesaurus construction are established to achieve and maintain consistency (according to ISO 2788:1986 and BS 8723-2:2005). The principle of place of unique definition (derived from J. Farradane), the analysis of the semantic category and facet analysis (from Ranganathan, CRG, BBC and ISO and BS standards again) are adopted to define semantically the concepts and to build the hierarchies of the thesaurus.

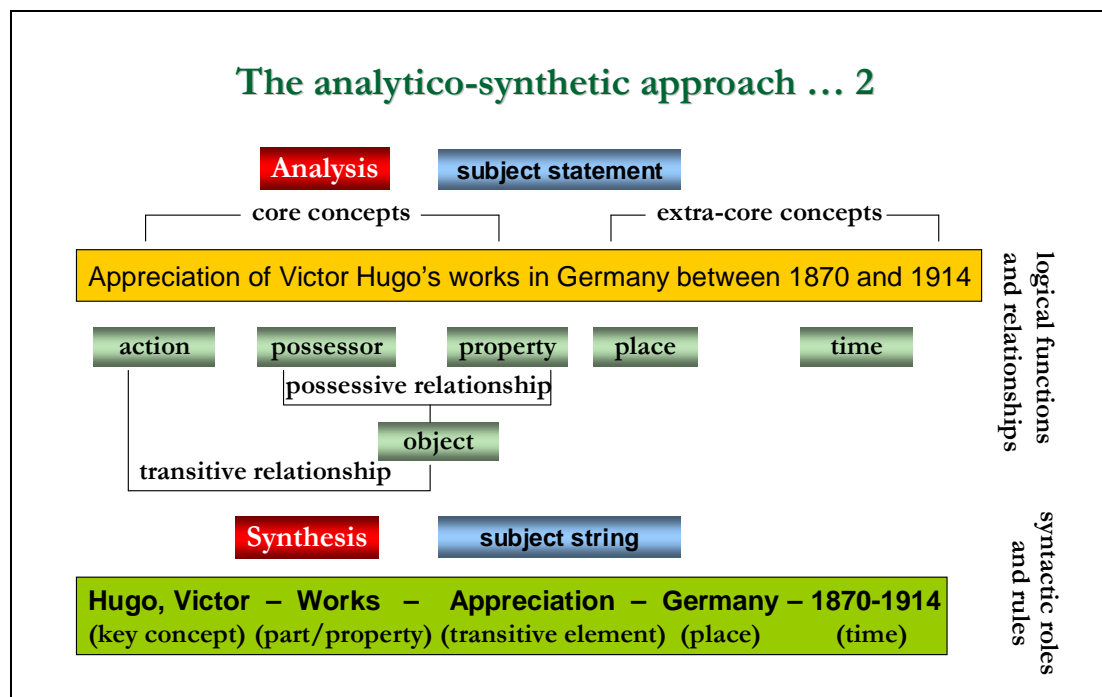
To explain the method through an example let us examine a work about “the appreciation of Victor Hugo’s works in Germany between 1870 and 1914”. It is presented as an instance of the class (entity) *concept* in FRBR<sub>OO</sub><sup>4</sup>. In our definition this is not a concept, but the base theme of the work. We find five elements in it (considering two dates as one period of time), not simply dealt with side by side, but syntactically related. The analysis – apart from their being general concepts or named entities – looks for the logical role played by each of them, starting from the existence of a concept of action (“appreciation”), with an object (“works”) and its owner (“Victor Hugo”), a place (“Germany”) and a time (“1870-1914”) where it happened and when.

---

<sup>4</sup> FRBR *object-oriented definition and mapping to FRBR<sub>ER</sub> (version 0.9 draft)*, International Working Group on FRBR and CIDOC CRM Harmonisation, supported by Delos NoE, editors Chryssoula Bekiari, Martin Doerr, Patrick Le Boeuf, [http://www.ifla.org/files/cataloguing/frbrg/frbr-oo-v9.1\\_pr.pdf](http://www.ifla.org/files/cataloguing/frbrg/frbr-oo-v9.1_pr.pdf), p. 34



To connect these elements in a meaningful and useful way, the rules for the citation order (syntactic rules) state the first position for the key concept, based on role analysis: here, the object of a transitive action. It is followed by the related transitive element (the action), and by the extra-core concepts (place and time). "Hugo" and his "works" are analysed, in a possessive relationship, as the possessor and his property, the latter being a dependent role, thus following the independent one; as the key concept, they keep the first place together, before the action.



The work *is* about the appreciation, but *per se* this concept is less interesting; Hugo's works are not studied *per se*, but only in the appreciation they had in Germany. Nevertheless, they remain the focus of

the work, with the limiting specification of the peculiar point of view considered here, the German judgement.

From a morphological and semantic point of view, each concept is represented by a preferred term, selected from natural language according to recommended procedures for vocabulary control. The accompanying construction of a thesaurus places every concept in its semantic field and helps exactly searching as well as to discover related access points of interest for similar inquiries.

The result of the synthetic process and vocabulary control is a subject string:

Hugo, Victor - Works - Appreciation - Germany - 1870-1914

that is:

- co-extensive to the subject content of the work, because it includes every concept considered,
- easily understandable, because it describes the complete theme in a logic way.
- suitable for arrangement and browsing, because the terms are syntactically ordered, and
- suitable for searching, because the terms are individually identified and uniformly selected.

## 2. A binary scheme

After this simple description, the focus moves to the conceptual model underlying the Italian system, a model valuable also for other indexing systems based on the same principles. We can recognise a binary structure, where each aspect is made up by a couple of fundamental elements:

- two kinds of *entity*: *concept* and *theme*
- two kinds of *relationship*: *semantic* and *syntactic*
- two kinds of *language*: *vocabulary* and subject *strings*
- two kinds of *operations*: *vocabulary* construction and subject *strings* construction
- two steps in *searching* (“two-steps search”): by *terms* and by *strings*
- two main kinds of *users’ interest*: wide survey and exact theme
- two kinds of *quality ratio*: *recall* and *precision*.

All these elements are logically connected on both the horizontal (within each couple: *concept–theme*, and so on) and the vertical plane, along both semantic and syntactic sequences, as we see in the diagram:

**Subject**

<i>entity</i>	concept	theme
<i>relationships</i>	semantic relationships	syntactic relationships
<i>language</i>	vocabulary	subject strings
<i>operations</i>	vocabulary construction	subject strings construction
<i>searching</i>	by terms	by strings
<i>Users' task</i>	wide survey	exact theme
<i>quality ratio</i>	recall	precision

The horizontal lines link together the couples of elements, defining each aspect of the conceptual universe (e.g. entities, relationships, language, etc.)

The vertical columns link all the aspects of the conceptual universe to each element and vice versa (e.g. a wide survey is linked to the term(s) used in searching, to the vocabulary and its semantic relationships representing concepts, to the recall as principle of evaluation).

Without examining each element in detail, let us explain some of them for a clearer comprehension and some amplification.

First, the couple of entities. In an analytico-synthetic approach we are immediately aware that, even if a subject should be represented in a unitary way in relation with the work considered, when we want to build a subject catalogue, on the contrary, each complex subject should be examined distinguishing in it every constituent element. The analytic phase leads to discover the concepts used in the discourse developed in the text. They are interlaced components, but each is single in itself, and from this singularity and the variety of possible combinations derives the very possibility of any communication and dialogue, of comparing discourses, of recognizing known and shared notions, as well as new and available knowledge.

Therefore, *concept* is assumed in the precise meaning fixed by the norm ISO 5963 “a unit of thought”, expressed by a single “indexing term”.

To express the overall meaning of *subject* of a work, what it is about, the term *theme* has been chosen. It is not simply a term to avoid the semantic heritage of other long used words, like “subject” or “topic”. The notion of theme is particularly important in this model, as it is the core concept in the definition of subject. According to ISO 5963, subject is “any concept or combination of concepts representing a

theme in a document” (but already in Cutter’s *Rules for a dictionary catalog* a subject was “the theme or themes of the book, whether stated in the title or not”). *Guida GRIS*, on the basis of a deep reflexion, has defined the base theme as “the unitary object of knowledge to which any particular theme discussed in the document may be referred, and to which all information intentionally given by the author is correlated in the text, since the fundamental aim of the intellectual production of the whole document is just the will of communicating direct and specific notions about that object of knowledge”. Furthermore, the base theme “is a necessary property of any text perceptible as a consistent unit. The text, as a whole, may be considered as the answer to a single question, whose content is coincident with the base theme of the document”<sup>5</sup>.

Following this line of reasoning, we discover that the notion of *theme* is essential not only with regard to the definition of *subject*, but also for the conceptual organization of a document and the communicating process.

Looking at the following three couples in the diagram above, let us run the three steps on each column separately, semantic and syntactic. The distinction between syntactic and semantic relationships is well-known and is expressed in standards as follows:

- semantic (or paradigmatic, a priori) relationships are those “between terms assigned to documents and other terms which, because they form part of common and shared frames of reference, are present by implication” (ISO 2788:1986), or “that are valid in almost all contexts, especially when they are inherent in the definitions of the concepts which the terms represent” (BS 8723-2:2005)<sup>6</sup>.
- syntactic (or syntagmatic, a posteriori) relationships are those “between the terms which together summarize the subject of a document” (ISO 2788:1986), or “which exist only because the terms are used together in the context of a particular document” (BS 8723-2:2005).

Analysing what a work is about we find the core concepts it deals with. Each concept we meet in indexing and that we isolate, is nevertheless linked in our mind and in common knowledge, and surely

---

<sup>5</sup> *Guida all’indicizzazione per soggetto*, p. 13

in the work too, to other concepts, broader or narrower or belonging to another category but strictly linked to it (like the typical action of an agent, e.g. teaching and teacher). Generally we can define a concept by using some of these other concepts, for instance qualifying a thing as a particular kind of a more generic class of things, or explaining an action with the tool used to do it. Somehow all the concepts implied in the concept we are dealing with are worth of mention together with it, so it is convenient to enrich every concept we use in indexing with a constellation of related concepts, by means of hierarchic and associative relationships. In the meanwhile we know or discover that the same concept may be expressed by more than one term. So we need to choose one term as the preferred one among all, and to establish relationships between equivalent terms (preferred and not preferred), completing the semantic relationships of each term considered. In doing so, independently from the particular discourse a work is developing, we fix semantic relationships, uniform terms and access vocabulary, and build a structured vocabulary that, in the case of *Nuovo soggettario*, takes the shape of a thesaurus.

Analysing the concepts forming the theme of a work, we need to analyse also the relationships connecting each other in that particular setting. Through the role analysis, these syntactic relationships are shown, the core or key concept is chosen on the basis of a logical construction of the subject statement, and the other concepts follow it in a logical order, lead by syntactic rules. The resulting subject string is readable and expressive of the full theme.

Vocabulary control is independent from subject strings construction. Vocabulary (i.e. semantic) relationships do not appear in strings. Strings construction depends upon vocabulary control only for the form of terms, not for their order or relationships.

Also searching is seen as a couple of elements: search (a) by subject strings (typically browsing through a list) to find directly the exact theme of works, or (b) by terms, finding all the occurrences of a term, that is to say all the works where a concept is relevantly treated and, thanks to vocabulary control, without the inclusion of other concepts in case of polysemy or omography, without loss of occurrences

---

<sup>6</sup> BS 8723-2:2005 - *British Standards Institution. Structured vocabularies for information retrieval. Part 2: Thesauri*. London, British Standards Institution.

as it happens in case of expression of one concept by equivalent words. But the best achievement of the binary scheme is the possibility of doing the “two-steps search”: searching for one or more terms and finding first all the strings containing the wished term(s), with the opportunity to choose those relevant to the search, avoiding examining the themes that are less or not at all interesting.

A clear distinction of the two kinds of search corresponds to two main kinds of interest in users:

- a wide survey of all the works that deal with one concept (as the first moment of a comprehensive research, for instance), including the exploration of its domain, by means of semantic relationships, and
- a need of information about an exact theme of a work or a particular association of concepts in it (as a means to identify or select a useful document, for instance), avoiding not pertinent or meaningless co-occurrences of the terms searched.

At last, also the upshot of searching is clearly distinct along the two lines of semantics and syntax: searching by terms satisfies the requirements of the best recall, while searching by strings satisfies precision.

The two columns are autonomously conceivable (e.g. the semantic column, from identifying a concept, to the choice of a preferred term, to the net of relationships, can stand without any involvement in a particular theme developed in a work). At the same time there are many points of contact and, generally speaking, they are reciprocally necessary (e.g. well constructed strings are useless without control of terms).

### **3. Subject indexing in conceptual models**

Looking at the Italian system in a more abstract way, and including it in the wider context of the bibliographic universe, we can propose an enriched version of the FRBR model.

In FRBR final report the entities traditionally coded in lists of proper names are considered for the subject relationships. They are distinct by tangibility (*concept* and *object*) and by the dimension where they exist, in space or in time (*place* and *event*). The entities of group one and two (*work*, *expression*,

*manifestation, item, person, corporate body*) are added as they are already present in the model. There is no consideration for the classes collecting these individual entities (e.g. battles, to collect The Battle of Hastings with other great battles when treated together in one work) and consequentially for relationships between entities as subjects (semantic relationships, e.g. between battles and wars). There is no consideration for the relationships between the entities that are subject of a *work*, and are connected to one another (syntactic relationships). The entity *concept*, in the broader meaning assumed in FRBR, could function as the collector of all the subjects more complex than an individual entity, including or not including named entities that should stand by themselves, like in the example used above. In it “Victor Hugo” and “Germany” appear as parts of one *concept* (the full theme), but they are a *person* and a *place* respectively. Thus, according to the model, three relations may be provided:

*Work* ==> has as subject ==> *concept*: the appreciation of Victor Hugo’s works in  
Germany between 1870 and 1914  
==> has as subject ==> *person*: Victor Hugo  
==> has as subject ==> *place*: Germany

To overcome this inconsistency or incompleteness within an abstract model, it is worth considering the deep structure of subject indexing, instead of starting from the evidence of current indexing systems (i.e. the surface of indexing). From this point of view, we consider the aboutness of a *work*, particularly as the theme around which a discourse is organised, the base theme, independently from the different theories “about aboutness”. This is not an alternative, but an integration to the model of fragmented relationships towards individual concepts, objects, etc., provided that we put together the two levels

- the level of the structured discourses combined in the *theme* and taken as a whole, and
- the level of the modular plurality of the *concepts* serving to build discourses.

So, the model underlying the Italian system, and suitable for any other indexing system, considers only two entities for the subject relationship in group three: *theme* and *concept*, and new relationships between them and with the other entities.

Instead of four juxtaposed entities plus pre-considered entities from group one and two, i.e. a sort of categorisation not exhaustive (*concept* serves also as a residual class) and partly overlapping (an *item* is an

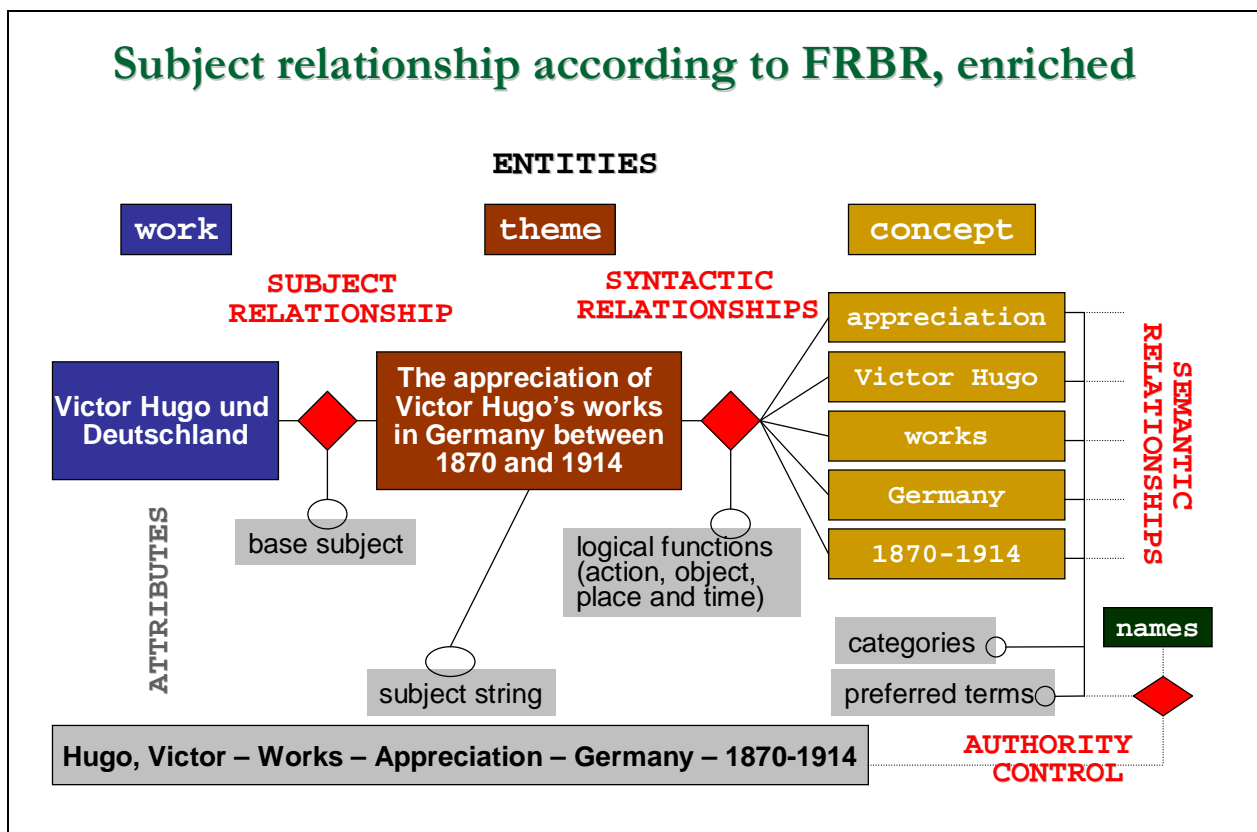
*object*), the nature of aboutness requires an approach similar to that of group one entities. The “products of intellectual or artistic endeavour” are analysed in the inner relationships between the different aspects of the same thing (a book or a disc are seen as a *work*, as an *expression*, as a *manifestation*, as an *item*). In a similar way the aboutness of a *work* should be analysed in the inner relationships of the overall *theme* and of the *concepts* contributing to its making.

The syntactic relationships grant this kind of connection between the two levels. The attribute of this relationship is the logical function of the *concept* in the *theme*:

*theme* ==> has as logical function ==> *concept*  
*concept* ==> is a logical function in ==> *theme*

It specifies the logical roles played by each component *concept* in the specific *theme* through their values (e.g. “The appreciation of Victor Hugo’s works in Germany between 1870 and 1914” has as action “Appreciation”, and “Appreciation” is the action in “The appreciation of Victor Hugo’s works in Germany between 1870 and 1914”).

A diagram shows the proposed model applied to our example:



The syntactic relationships allow any *concept* to be part of many *themes*, i.e. in subject relationship with many works, where it may play different logical functions or roles, always keeping its own unity and uniqueness (the same happens to a *work* in group one that may be realised through many *expressions* and embodied in many *manifestations*, but still remains the same *work*).

The component *concepts*, freed from accidental belonging to a specific *theme* and playing a specific role, are related to other *concepts* by implication or by other kinds of a priori relationship and should be connected to them by semantic relationships, as it happens in every indexing language, thus contributing to the complete mapping of the domain covered. But semantic relationships, like the definition of categories and the choice of terms, are a task of indexing languages. Therefore, they should be considered in the model only as the necessary expansion into the field of the implementations and of the distinct solutions supplied by different systems. In the diagram above it is shown on the right side, together with the work of authority control, which is done on the names for the concepts and on the subject strings for the themes.

What is said as regards subject headings languages is valid, to a large extent, also for classifications, where a notation, no matter how expressive it may be, stands for a complex subject. The component *concepts* are considered in the analysis and, somehow, in the scheme, even if they are neither represented on their own nor by the level of low specificity of the classification.

The conceptual model for *Functional Requirements for Subject Authority Data (FRSAD)* has just been published in its 2<sup>nd</sup> draft 2009-06-10 for world wide review<sup>7</sup>. Despite the focus on authority data, the first goal of the study was to build a conceptual model of group three entities within the FRBR framework. Two new entities have been pointed out by the FRSAR Working Group: *thema* and *nomen*, and two relationships: *work* has as subject *thema*, and *thema* has appellation *nomen*.

*Thema* refers to “anything that can be subject of a *work*” and includes any FRBR entity, like a super-class, or a generic term without a meaning on its own. Despite the almost coincident word, *thema* looks like our *concept*, representing separate ideas (see the example for *A history of time*), not like our *theme*.

Complex *themata* (or *themas*) are admitted without facing their inner relationships, and pushing them to the representation side and to the differences among indexing systems. But this is a core of the *work-thema* relationship, that is marked many to many, and a clear distinction should be drawn between the aboutness of a *work*, as resulting from its conceptual analysis, and the indexing policy of the agencies.

Complexity in aboutness is in the variety of *concepts* treated and in their relations. *Concepts* may be (a) juxtaposed as not connected *themes*, or (b) related to one another to form one complex *thema*. The latter case requires, from a logic point of view, the two levels of *theme* and *concept* of the binary model in order to represent both the whole aboutness of the *work* and the presence of singular *concepts*. A different point is the indexing policy chosen by the agency and strictly linked to the adopted system. A system may provide that all *concepts* are summarised in one subject, or that any distinct *concept* treated in a *work* is indexed, like the index of a book listing all the noteworthy things mentioned in the text. But the latter choice does not delete the existence, in any structured text, of one unitary and intentional complex *theme* where the *concepts* converge. Therefore, the binary scheme is useful to exchange information among systems even when two or more *themes* coexist or when the overall level is not used in the implementation. Moreover, this is the way:

- to move categories out of the tangle of general and additional attributes, towards the clear attributions of categories to *concepts*, as a device to build semantic relationships, and
- to include syntactic relationships, as required for the completeness of the model and in the goals of the study.

The latter new entity *nomen* in FRSAD has a parallel in the entity *name* in FRAD. Here again the binary scheme of the analytico-synthetic model offers a better base than a mono-linear one for a model devoted to authority data. It allows a distinct and more accurate control both on terms for single *concepts* (including their categorisation) and on strings, classification numbers or equivalent representations of *themes* composed by related *concepts*, not only formally (e.g., the consistence of citation order) but also applying the syntactic analysis of logical roles. *Nomina* (or *nomens*) representing simple *concepts* need

---

<sup>7</sup> <http://nkos.slis.kent.edu/FRSAR/report090623.pdf>.

vocabulary control and semantic relationships, and do not admit role analysis and syntactic relationships; the opposite for *nomina* representing *themes*. Therefore, it is difficult to imagine an authority system that does not distinguish the two levels.

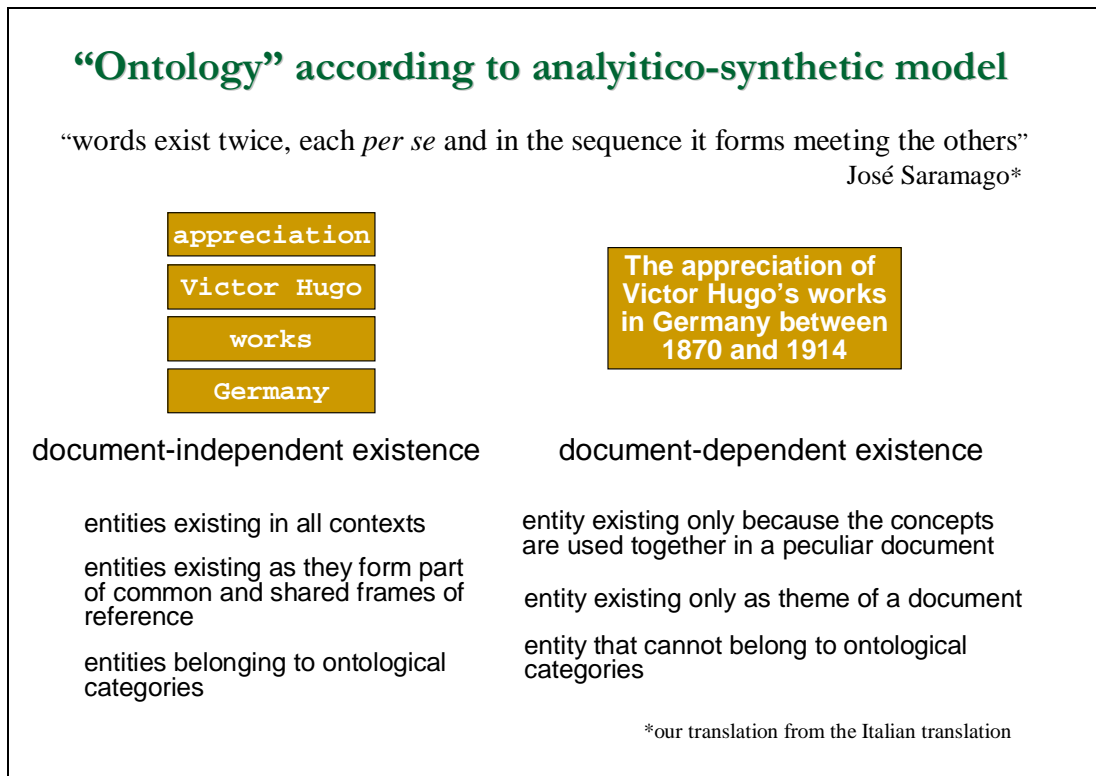
#### 4. Which kind of ontology?

The analytico-synthetic model proposes an issue as a logic conclusion of its adoption: which kind of “ontology” is needed for the world of information? That is to say, in our example: do simple *concepts* like “appreciation”, “works”, “Victor Hugo”, “Germany” and their categorization and relationships have the same nature as the complex *concept*, that we have called *theme*, “the appreciation of Victor Hugo’s works in Germany” or not?

The solution suggested is the “double existence” of *concepts*:

- a document-independent existence, typical of simple *concepts*, of their categories, of their semantic relationships;
- a document-dependent existence, typical of complex *concepts* and of their syntactic relationships.

In this way, we have a double “inventory of the world”. In the former, the *concepts* exist *per se*, and we call them properly “*concepts*”. In the latter, they exist as they are treated together with others in particular documents, and we call these associations “*themes*”.



A *concept* keeps its identity in its double existence, while the nature of *concepts* and *themes* is different in that:

- a *concept* exists in any context, it forms part of common and shared frames of reference and belongs to an ontological category;
- a *theme* exists only because it has been conceived in the context of a peculiar work, where its component *concepts* are associated by peculiar relations, and it cannot belong to any ontological category, as it is formed by a combination of categories.

In the structure of indexing languages, *concepts* are treated in semantics and their representation is subject to authority control on vocabulary, while *themes* are formed in a consistent manner according to syntactic rules and their representation is subject to authority control on subject strings, class numbers, etc. In library catalogues the *themes* propose the aboutness of *works*, while the *concepts* are the most immediate nuclei of mental approach for searchers, independently from the associations in which they might occur, and unaware of which *themes* actually were treated in catalogued documents. Saving the independent value of singular *concepts* in the context of the *themes* where they appear is the condition to find known and unknown themes when searching concepts: exactly and without noise.

The binary interpretation supplied by the analytico-synthetic model allows users to satisfy any different kind of task, as it is founded on both the particles and the aggregations of the meaning of works (of what is meant in works). The clear and consistent distinction between semantics and syntax, and their intersection in the “double existence” of *concepts* allows and helps both discrete and joined searching at will. Moreover, it makes information exchange easier in multilingual and multicultural context, as it does not stand on the signifiers of languages and allows to reduce some differences in structure between very far languages by means of analysis, that is to say factoring complex syntagms into simple concepts, translating and recombining them in the other language.

The same is valuable also between different indexing systems, between alphabetical indexing and classifications, for instance, discovering the deep common origin of strings and notations in the same sets of concepts combined in different ways.

The analytico-synthetic model of subject indexing, restored and developed in the Italian renewal by *Guida GRIS* and *Nuovo soggettario*, shows a wide range of efficient functions and suggests consistent improvements on IFLA’s conceptual models. It supplies the right premises even for effective searching in the web and may serve as the bases for designing and implementing high quality automatic searching.