The Nuovo soggettario as a service for the linked data world

Giovanni Bergamin, Anna Lucarelli

Introduction

The Nuovo Soggettario (hereinafter, NS) edited by the National Central Library of Florence (BNCF), is the main Italian subject indexing tool for various kinds of resources. It has been developed in collaboration with the Italian National Bibliography (BNI) which holds a leading role in the bulding and development of subject indexing tools in compliance with the International Federation of Library Association (IFLA) recommendations (The NS employment by Italian National Bibliography is also described in Jahns, *Guidelines for subject access in National bibliographies*) and other International standards. This tool is used by general and specialized Italian libraries (indexers, researchers, users), in particular those participating in the Servizio Bibliotecario Nazionale (SBN), and is also employable in archives, multimedia libraries and documentation centres. The NS entered into the tradition of the analytico-synthetic languages; the system consists of a semantic and syntactical apparatus and, in compliance with the uniform and specific heading principles, it is conceived as a system to be applied in both pre-coordinated (the terms are combined in subject strings) and post-coordinated indexing environments (the terms are extracted from a controlled

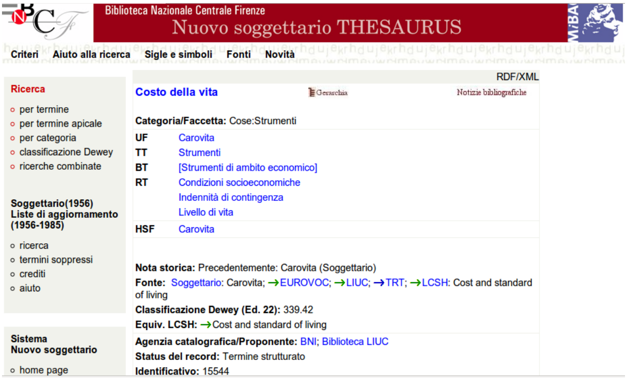
JLIS.it. Vol.4, n.1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5474

vocabulary and used as key words). The main component of the NS is a universal thesaurus built in compliance with the International standards, available online from the 2007.[1] It is a tool continuously being developed and currently accessible on the BNCF website. At the moment the Thesaurus consists of 46,000 terms derived from the 1956 Soggettario and its updates (which are being controlled and standardized), from new terms introduced for the semantic relationship network and from new terms proposed by the BNI indexers and other partners (Lucarelli et al., "The Nuovo soggettario Thesaurus: structural features and web application projects"). The terms are organized inside a structure based on four main categories and on semantic relationships determined by standards (*ISO2788:1986 – Documentation, guidelines for the establishment and development of monolingual thesauri. Documentation, principes directeurs pour l'établissement et le développement de thesaurus monolingue*; *ISO25964/1:2011 – Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*). They are equipped by a rich apparatus of notes, connections with formerly preferred terms (historical variants), an indication of the correspondent numbers of the Dewey Decimal Classification, as well as by Sources which are in constant updating and employed for the control of morphologies and meanings.[2] The Thesaurus is integrated with the BNCF opac and with the opac of the other libraries that adopt it. The users can navigate from the controlled vocabulary to the bibliographic records. Regarding Linked data, the Thesaurus is linked with other thesauri, with some encyclopedias (such as Wikipedia and the prestigious Italian Treccani encyclopedia [3]), and with other cultural instituition's digital resources. The NS thesaurus promotes the Italian language and multilingual information retrieval by its data management software, however is

[1] http://thes.bncf.firenze.sbn.it/ricerca.php.

[2] http://thes.bncf.firenze.sbn.it/fonti.php.

[3] http://www.treccani.it.



Costo della vita

Categoria/Faccetta: Cose:Strumenti

UF	Carovita
TT	Strumenti
BT	[Strumenti di ambito economico]
RT	Condizioni socioeconomiche Indennità di contingenza Livello di vita
HSF	Carovita

Nota storica: Precedentemente: Carovita (Soggettario)
Fonte: Soggettario: Carovita; →EUROVOC; →LIUC; →TRT; →LCSH: Cost and standard of living
Classificazione Dewey (Ed. 22): 339.42
Equiv. LCSH: →Cost and standard of living
Agenzia catalografica/Proponente: BNI; Biblioteca LIUC
Status del record: Termine strutturato
Identificativo: 15544

also in compliance with standards (*Guidelines for Multilingual Thesauri*). A large number of terms has a cross-language equivalence relationship with Library of Congress Subject Headings (LCSH) preferred terms, displayed and linked by "Equiv. LCSH" note. i.e. «Costo della vita»:

In the last period, the NS is developed in two ways:

1. Interoperability: since 2010, metadata are available in Resource Description Framework (RDF)/SKOS format and will be employable in the Linked data world, not only in closely librarians contexts;

2. Automatic indexing: thesaurus is testing in automatic indexing of digital resources; in particular our goal is to reduce the cataloguing expenses.

These developments are outlined with the programs of other countries in the indexing domain, such as demonstrated by IFLA papers (Gömpel and Svensson, "Managing legal deposit for online publications in Germany").

SKOS standard for thesauri

Simplified Knowledge Organisation System (SKOS) is defined as a common data model,[4] developed by W3C Semantic Web Deployment Working Group (SWDWG),[5] for sharing and linking knowledge organization systems (such as thesauri, taxonomies, classification schemes and subject heading systems) within the semantic web. It is an application of the RDF. The most important thesauri, developed by National Libraries, are progressively adopting this standard for their controlled vocabularies. SKOS data are concepts which are independent of the terms used to label them, tagged as RDF triples and encoded using any concrete RDF syntax. The concepts, which are expressed by preferred terms in the thesaurus and used as descriptors in indexing system, are identified with URIs and are labeled with skos:prefLabel, expressed in one or more natural languages. The standard assigns alternative lexical labels to conceptual resources which have not a URI: skos:altLabel to represent a relationship between terms in a thesaurus that both represent the same concept; skos:hiddenLabel to represent misspelled variants of other lexical labels, abbreviations and acronyms. The standard expects the possibility to define and qualify the concept with some other information expressed by some labels which came from skos:note superclass (skos:definition; skos:scopeNote; skos:example: gives examples for the use of the terms; skos:historynote: it may be applied to a preferred or non-preferred term or to a concept. It should be used when a new preferred term is added to the thesaurus or change is made to an existing term that affects the concept's scope in different periods of application; skos:editorialnote: gives some administration information; skos:changenote: documents the different choices and modifications). The hierachical ad associative

[4] http://www.w3.org/TR/skos-reference.

[5] http://www.w3.org/2004/02/skos.

thesaural relationship, established between concepts, are labelled with skos:broader, skos:narrower, skos:related.

NS in SKOS format

Our thesaurus has been converted in SKOS format at the beginning of 2010. It was presented as a prototype at the IV Summit di Architettura dell'informazione (Motta and Rodighiero, "Il thesaurus del Nuovo soggettario interpreta SKOS") and then improved within the Digital resources automatic indexing project, developed in the BNCF since 2011 (Viti, "Interoperabilità fra thesauri generali e thesauri specialistici in ambito economico-finanziario. Il caso del Nuovo soggettario"). Our work has followed many stage and now is growing gradually in comparison with current developments. One of the most important problems starting with the prototypal stage was about the impossibility that SKOS – even if it defines an expressive array of sibling terms and collections of concepts – recognizes node labels as conceptual units which belong to hierarchical relationships; the standard calls them exclusively skos:Collection. The application doesn't establish links between the members of arrays and the general concept which expressed the same array. Instead each member of the array (skos:member) is directly linked with the concept which comes before the node label and not with the array identified by skos:Collection. Through the URI's skos:Concept we could verify if a skos:Member belongs to a skos:Collection and rebuild the whole hierarchical relationships. For example, a direct link can not be established between the skos:Concept Bambini, skos:Collection [Bambini secondo l'attività] and skos:member Bambini artisti. During our conversion we have found other problems; in particular, there where some difficulties for translation of two types of semantic relationships:

1. historical variants relationship (expressed with HSF, Historical see for) links some preferred terms with some preferred terms in the past which are no longer accepted;

2. the multi-word terms splitting relationship (expressed with USE+/UF+) create reciprocal link between multi-word terms and single word terms derived from factoring.

In the first case, we have refined the historical variants tagged skos:altLabel class as sogi:obsoleteTerm. Practically, the the historical variants begin a non preferred term. About the splitting of the complex concepts, at the moment, we have decided not to implement the SKOSXL extension (which identifies also the terms by an URI, not only the concepts), because about this we have not found some examples of applications. At the moment, the splitting relationship is expressed by a note in a specific field. The apparatus of note (definition, scope note, history note, sources, DDC...) is suitably expressed by SKOS. The syntactical note, that in the thesaurus guides the subject strings constructions, is labelled with skos:example. The assignment of an URI to the concepts promote the interoperabilty between different KOS, that is the possibility of mapping the semantic entities of different conceptual schemes. To realize this aim, the standard establishes three different equivalence levels: skos:closeMatch; skos:exactMatch; skos:broaderMatch e skos:narrowerMatch; skos:relatedMatch.[6] About this, we are testing the creation of equivalences to support the linked data between NS terminology and its equivalents in another vocabularies. We have chosen an empiric approach, based on an international reconnaissance of others SKOS applications. During the creation or maintenance of the NS equivalences can be activated by:

[6] http://www.w3.org/TR/skos-reference.

1. entering in a specific field (Source) the name of the vocabulary you want to cite: if the cited vocabulary is available SKOS, SKOS relationship of NS will be enriched with skos:closeMatch. If the the cited vocabulary is not available in SKOS this citation will be used for the creation of a deep link to the vocabulary (i.e. a direct link to the corresponding term);

2. entering the equivalence in a specific field (Equiv. LCSH) which refers to the Library of Congress Subject Headings equivalences: also in this case we use closeMatch relationship which is conceptual wide-ranging than exactMatch which was used in the initial stage.[7]

AGROVOC	1070
DBPEDIA	800
LCSH	750
ThESS	450
RAMEAU	240
EUROVOC	80

We are testing the settlement of equivalence semantic levels, between NS and ThESS (the thesaurus of Mario Rostoni Library of the LIUC University), by skos:broaderMatch, skos:narrowerMatch, skos:relatedMatch tags.

[7] About this, we have analysed matching procedures between RAMEAU and LCSH, in which the link is an exactMatch or a closeMatch without equivalence level's identification. At the moment, the links between RAMEAU and LCSH are established with a closeMatch (one sense relationship: RAMEAU -> LCSH) while those between LCSH and RAMEAU are established with an exactMatch LCSH<>RAMEAU.

The NS for automatic indexing of digital resources

As already mentioned, in BNCF has been running since 2011 a prototype test for the use of NS for semiautomatic subject indexing of digital resources acquired through legal deposit.[8] The BNCF initiative is in line with other European national libraries initiative (for instance, the Deutsche Nationalbibliotek project in this field is a relevant one (Junger, "Can indexing be automated? - the example of the Deutsche Nationalbibliothek") and takes into account two objectives:

1. the need for change in cataloguing practices due to rising amount of publications in digital format;
2. the sustainability of subject indexing.

Here "automatic indexing" refers to procedures using algorithms and techniques – coming also as result of the latest technological research – that can be used for automatic (or semi-automatic) extraction from a text of "relevant" keywords / key phrases. These procedures may be based on keywords / key phrases extraction and assignment with or without support of a controlled vocabulary. According to recent tests in progress at the international level, automatic indexing seems to produce better results – in term of precision and recall – if assisted by controlled lists (such as thesauri). In our prototype, the process of extraction of keywords / key phrases is managed by the software application Keyword indexer (Biblioteca Nazionale Centrale di Firenze, "Procedure automatizzate di estrazione di parole e frasi chiave: specifiche tecnico-funzionali"). This application requires, as preliminary step, the

[8] The prototype was developed in collaboration with two Italian companies: Casalini libri http://www.casalini.it and @Cult http://www.atcult.it.

creation of a knowledge base (also called learning model) based on sample documents (with associated metadata) and a vocabulary in SKOS format. In particular, as a first test, we created a thematic learning model on the economic and financial sectors, using the following structural components:

1. set of digital full-text documents: a sample of Italian doctoral thesis belonging to the economic and financial sector according to the classification system determined by the MIUR (Ministry of Education, University and Research): the classification symbols are SECS-P/01-13 and SECS-S/01-06;

2. set of metadata associated with the selected set of documents;

3. Nuovo Soggettario (NS) in SKOS format;

This model has been then applied to indexing the 2010-2011 issues of the digital journal LIUC Papers.[9]Keyword Indexer software, using TF/IDF (Term Frequency/Inverse Document Frequency) algorithm, was used to determine the ranking of terms . Obviously final results were affected by every variation of the above parameters.[10] Obviously final results were affected by every variation of the above parameters. For the time being, considering the last configuration of our test (choice of metadata closest to the semantic content

[9] Italian monthly journal focused on social science and in particular on Economics and Management http://www.biblio.liuc.it/pagineita.asp?codice=82. It is edited by Mario Rostoni Library of Carlo Cattaneo University in Castellanza (LIUC) which cooperate with the NS project.

[10] «The TF/IDF weight (term frequency–inverse document frequency) is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The TF/IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others».http://en.wikipedia.org/wiki/Tf*idf

of the document such as title + abstract, title + MIUR - Ministero dell'Istruzione, dell'Università e della Ricerca - classification symbol), findings are to be considered provisional: in this case automatic indexing is not closest enough to intellectual indexing. For these reasons we plan to continue our tests taking into account:

1. a multidisciplinary learning model for the general needs of a National library;
2. refinement of procedures for preparation of metadata to be used for building the learning model: we are considering both intellectual indexing and/or new automatic procedures for extracting topic keywords which could be used as metadata.

In any case is worth considering that all the tests are based on reuse of open source software components freely available on the net.

NS and the Semantic web

For interoperability with other applications, NS is available through the Zthes protocol.[11]Zthes is essentially an evolution of Z39.50-based information retrieval protocol, where the targets are not library catalogs but controlled vocabularies in compliance with ISO 2788 and ISO 5964. Through Zthes, applications can exchange data using the well-known and established mechanism of application interfaces known as Application Programming Interface (API)s. In particular Zthes uses SRU syntax (Search-Retrieval via URL) where requests for access to a controlled vocabulary are included as a parameters within a URL and response messages are tagged using XML syntax: in other words, Zthes uses http protocol - designed for interaction between the user (browser) and machine (web

[11] http://zthes.z3950.org.

Server) - for communication between machine and machine. API based on Zthes are easy to implement but they must however deal with the limitations of all the API based on HTTP protocol and XML syntax for the exchanged messages. In particular, an important limitation is the fact that in general API are not reusable - either at the protocol level or at message encoding level - in different contexts (a custom API is need for different kind of application). Interoperability through the infrastructure of the semantic web (RDF language and the SPARQL protocol in particular) certainly overcomes APIs limitations and this fact is the main reason for making available the NS using SKOS/RDF. As many have noted the success of the semantic web depends on widespread use including the ability of penetration into everyday applications that we use to access information. Among them search engines play an important role and, on the other hand, one can certainly argue that if the search engines are not interested in semantic web, there are little chance to establish semantic web as a widespread infrastructure for the information access. Search engines have long been interested in the semantic of documents (interested in indexing coded data within documents). The recent agreement – known as schema.org[12]– between the most important search engines (Google, Yahoo, Bing and Yandex) with the purpose to commonly define a standard to describe elements within HTML (HTML5) pages based on RDF, can (or should) be an interesting opportunity for libraries (Ronallo, "HTML5 Microdata and Schema.org"). Use of schema.org metadata set - in fact a simple extension of the HTML tags - will allow search engines to "understand" the structure and the nature of a given document. To remain in the library world, as we know search engines already can index the bibliographic records but treating them like any HTML page losing the ability to identify the semantic structure

[12] http://schema.org.

(the elements that characterize the bibliographic record). With a HTML/RDF coding based on schema.org our catalogs, thanks to metadata they contain, will be interpreted as "semantic objects" by the major search engines. This will increase the value of the information produced by the libraries increasing also the likelihood of bringing together '"supply and demand". Of course schema.org is not proposing a new model for bibliographic record, but within the library world schema.org can be used as as strategy to promote on the web here and now the information we produce (value and limitations included). Schema.org has recently decided also to maintain a list of suggested extensions.[13] This list will include both basic and widely used vocabularies (e.g. Wikipedia), and vocabularies produced thanks to a "significant professional contribution" (LCSH is cited as an example). Since for schema.org there is no limitations for extensions, this list will be used by search engines as a priority indication for inclusion of content in the new "semantic indexing service". NS, available as SKOS/RDF, is ready to become also an extension for schema.org for people accessing search engines using Italian language.

References

Lucarelli, Anna, et al. "The Nuovo soggettario Thesaurus: structural features and web application projects". *Subject access. Preparing for the future*. Ed. Patrice Landry et al. Berlin/Munich: De Gruyter Saur, 2011. 155–168. (Cit. on p. 214).

ISO2788:1986 – Documentation, guidelines for the establishment and development of monolingual thesauri. Documentation, principes directeurs pour l'établissement et le développement de thesaurus monolingue. Geneva: International Organization for Standardization, 1986. (Cit. on p. 214).

Classification, IFLA and Indexing Section. Working Group on Guidelines for Multilingual Thesauri. *Guidelines for Multilingual Thesauri*. 2009. (Cit. on p. 215).

[13] http://www.w3.org/wiki/WebSchemas/ExternalEnumerations.

Gömpel, Renate and Lars G. Svensson. "Managing legal deposit for online publications in Germany". IFLA, 2011. (Cit. on p. 215).

Motta, Marta and Dario Rodighiero. "Il thesaurus del Nuovo soggettario interpreta SKOS". IFLA, 2010. (Cit. on p. 217).

Viti, Elisabetta. "Interoperabilità fra thesauri generali e thesauri specialistici in ambito economico-finanziario. Il caso del Nuovo soggettario". Diss. Università degli studi di Udine, 2012.

Junger, U. "Can indexing be automated? - the example of the Deutsche Nationalbibliothek". 2012. (Cit. on p. 220).

Biblioteca Nazionale Centrale di Firenze, @Cult. "Procedure automatizzate di estrazione di parole e frasi chiave: specifiche tecnico-funzionali". 2011. 9–11. [Documentazione interna alla BNCF, non pubblicata]. (Cit. on p. 220).

Ronallo, Jason. "HTML5 Microdata and Schema.org". *Code4lib jurnal* 16. (2012). <http://journal.code4lib.org/articles/6400>. (Cit. on p. 223).

Jahns, Y., ed. *Guidelines for subject access in National bibliographies*. Berlin/Münich: De Gruyter Saur, 2012. 77–79. (Cit. on p. 213).

Biblioteca Nazionale Centrale di Firenze. *Nuovo soggettario. Guida al sistema italiano di indicizzazione per soggetto. Prototipo del Thesaurus*. Milano: Editrice Bibliografica, 2006.

ISO25964/1:2011 – Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Geneva: International Organization for Standardization, 2011. (Cit. on p. 214).

GIOVANNI BERGAMIN, National Central Library of Florence.
giovanni.bergamin@bncf.firenze.sbn.it

ANNA LUCARELLI, National Central Library of Florence.
anna.lucarelli@bncf.firenze.sbn.it

Bergamin, G., A. Lucarelli. "The Nuovo soggettario as a service for the linked data world". *JLIS.it*. Vol.4, n.1 (Gennaio/January 2013): Art: #5474. DOI: 10.4403/jlis.it-5474. Web.

ABSTRACT: Nuovo soggettario (NS), edited by the National Central Library of Florence, is the Italian subject indexing tool for various types of resources. It has been developed in compliance with the IFLA recommendations, and other international standards in the field of subject indexing. This tool has been created for general and specialized Italian libraries, and for museums, multimedia libraries, archives and documentation centres. The main component of the NS is a general thesaurus available on the web since 2007 (http://thes.bncf.firenze.sbn.it/ricerca.php). The thesaurus comprises nowadays approximately 46.000 terms and is updated. It supports the new subject indexing practices and manages terminology deriving from collaboration between the BNCF and other libraries. The project is evolving in many directions and supporting interoperability. The main goal of the availability – since November 2010 – of the NS dataset in SKOS/RDF format, is to promote the use of this tool also beyond the traditional library environment. In this context three working areas have been taken into account: 1) improve accessibility and usability of the NS in the linked data environment: SPARQL endpoint, mapping to other datasets (including LCSH, RAMEAU, AGROVOC, EUROVOC, DBpedia); address the costs of bibliographic control starting from a project of automatic indexing (quality controlled) using NS in SKOS /RDF format and open source software tools; 3) cooperate with other institutions that are publishing linked open data.

KEYWORDS: Library linked data; Nuovo soggettario; SKOS

ACKNOWLEDGMENT: Thanks to Maria Grazia Pepe and to Elisabetta Viti

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15