Nuovo Soggettario Thesaurus between Linked Data, Automatic Indexing Tests and Other Developments

Anna Lucarelli – Elisabetta Viti

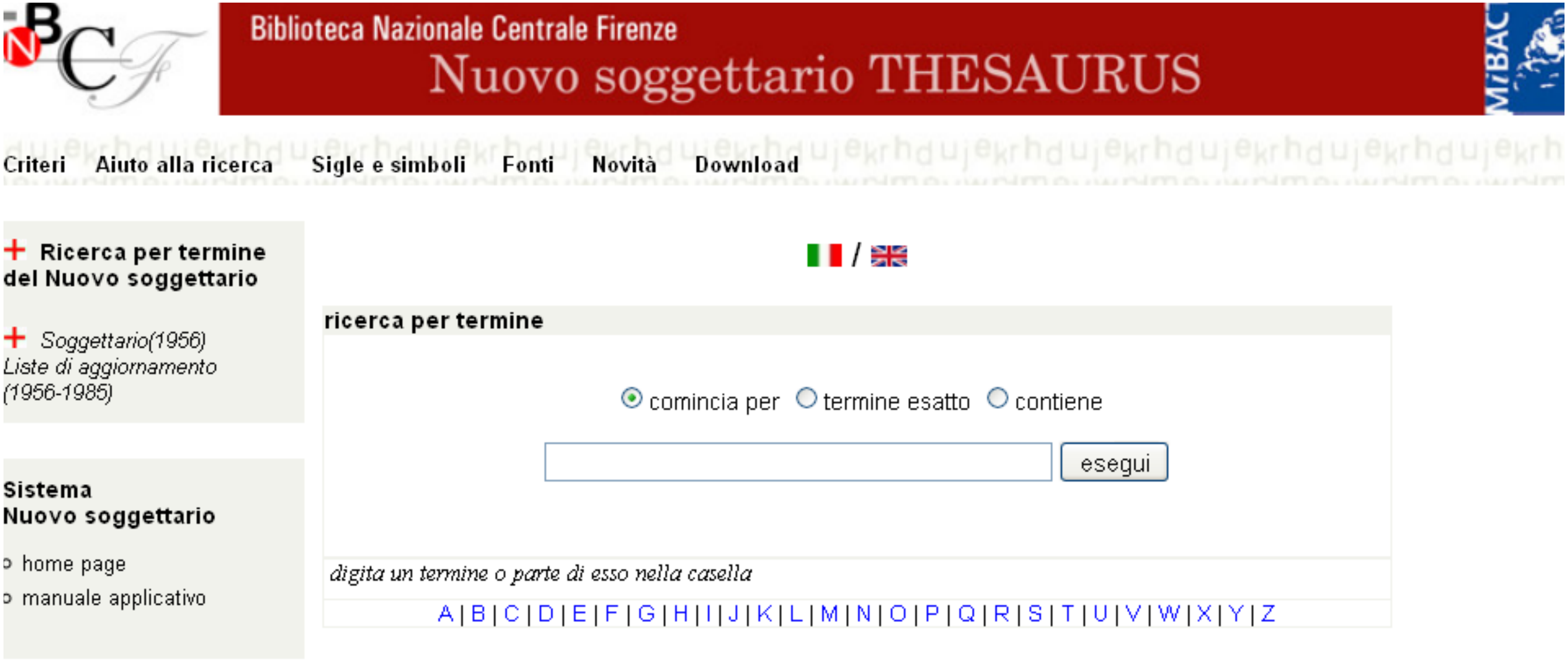
(Biblioteca Nazionale Centrale Firenze)

IRCDL 2016,
12 th Italian Research Conference on Digital Libraries ,
Florence 4-5 February

The BNCF collections

25,000 manuscripts - **4,000** incunabola - **29,123** 16th century books
1,000,000 autographs - **70,173** prints - **8,900,000** monographs
147,000 periodicals

258,000 digital resources (born digital and non)



Biblioteca Nazionale Centrale Firenze

Nuovo soggettario THESAURUS

Criteri
Aiuto alla ricerca
Sigle e simboli
Fonti
Novità
Download

+ Ricerca per termine del Nuovo soggettario

+ *Soggettario(1956) Liste di aggiornamento (1956-1985)*

Sistema Nuovo soggettario

- home page
- manuale applicativo

ricerca per termine

● comincia per ○ termine esatto ○ contiene

digita un termine o parte di esso nella casella

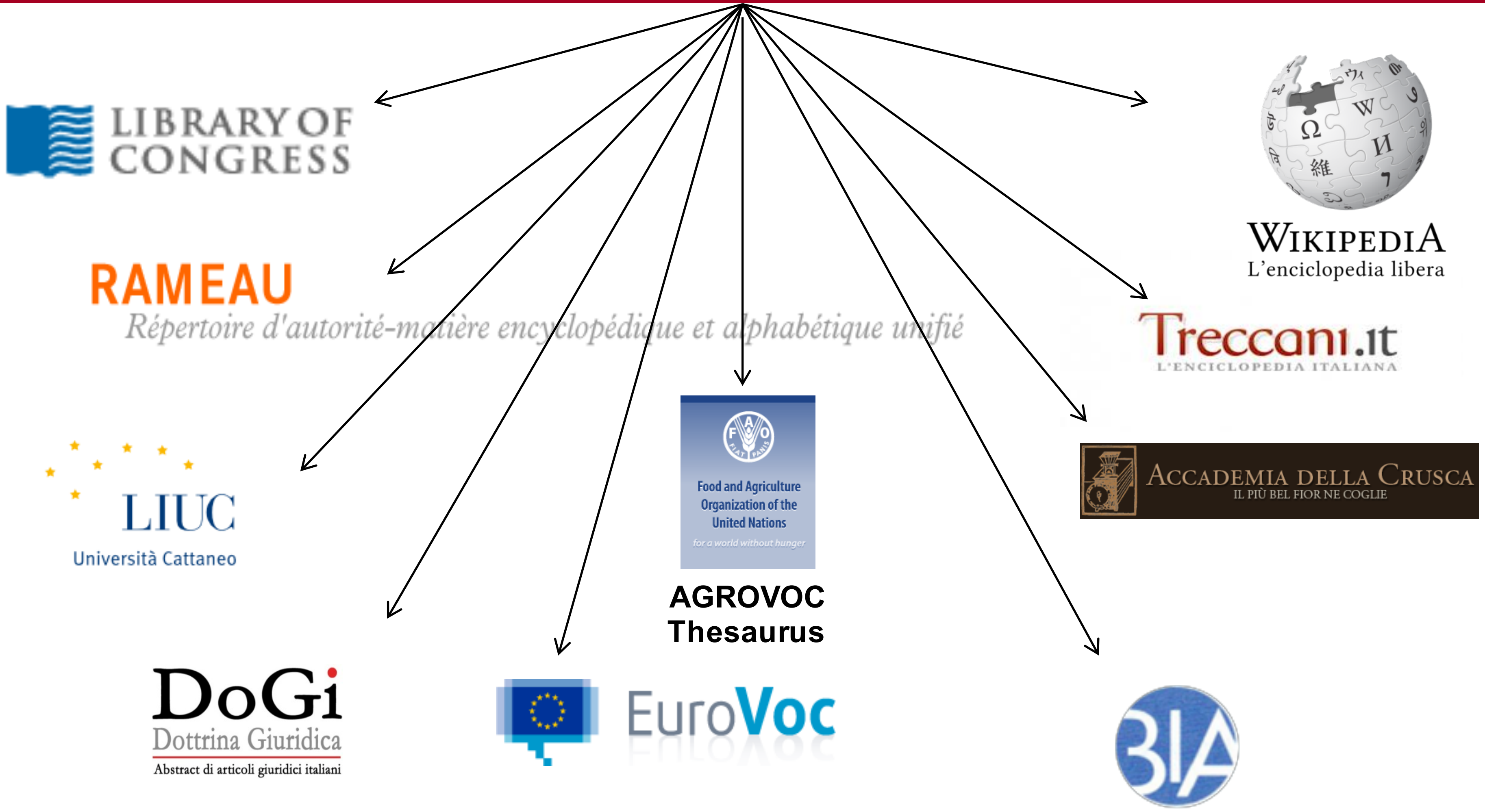
A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Universal Thesaurus with 57,000 terms

- Used for subject indexing of the bibliographic, archival and museum resources
- Available online since 2007 - English interface since 2013
- Integrated with the BNCF OPAC
- Compliant with Standard ISO 5963, ISO 2788 and ISO 25964
- SKO/SRDF and Open Data

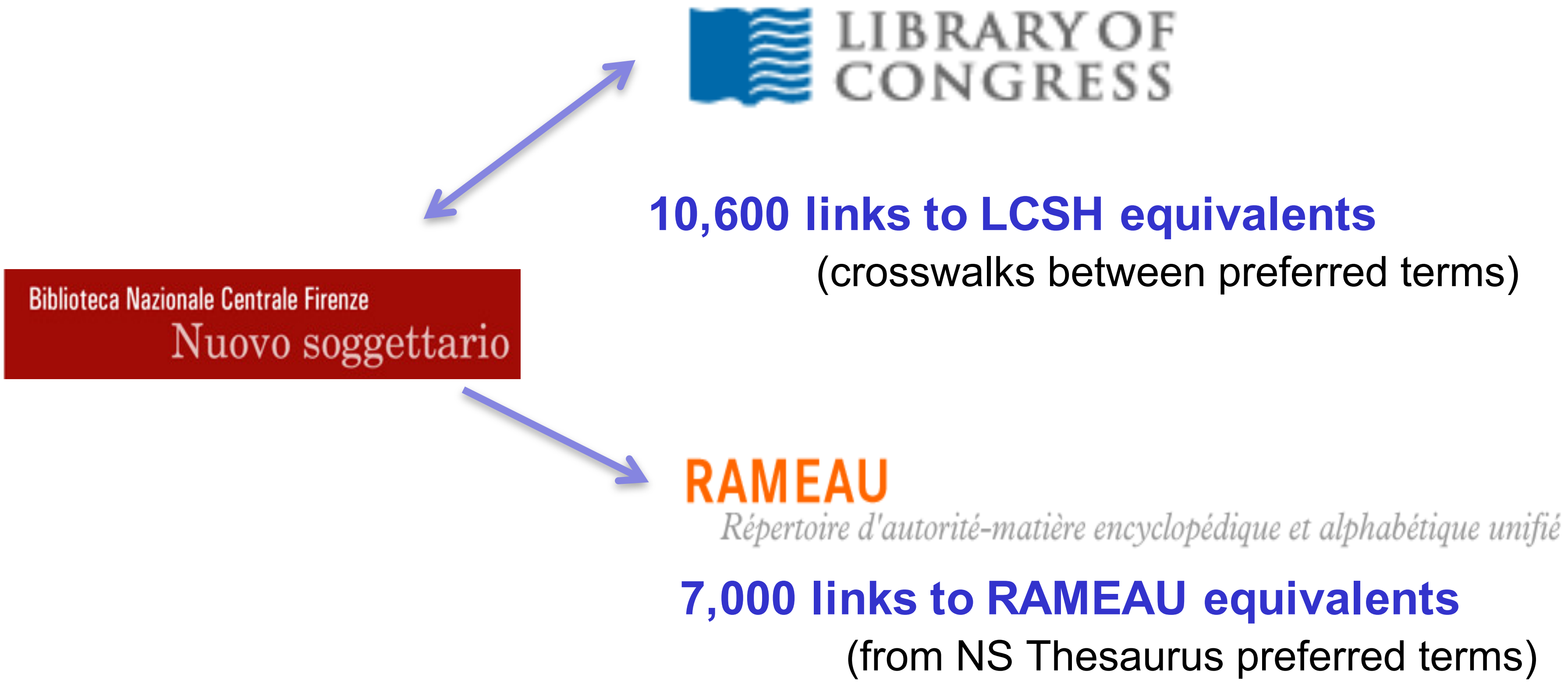
http://thes.bncf.firenze.sbn.it/ricerca.php

Mapping with Knowledge Organization Systems


LIBRARY OF CONGRESS
RAMEAU
Répertoire d'autorité-matière encyclopédique et alphabétique unifié
LIUC
Università Cattaneo
DoGi
Dottrina Giuridica
Abstract di articoli giuridici italiani
EuroVoc
FAO
FIAT PANIS
Food and Agriculture Organization of the United Nations
for a world without hunger
WIKIPEDIA
L'enciclopedia libera
Treccani.it
L'ENCICLOPEDIA ITALIANA
ACCADEMIA DELLA CRUSCA
IL PIÙ BEL FIOR NE COGLIE
BIA

AGROVOC Thesaurus

Linking with English and French Equivalents



LIBRARY OF CONGRESS

10,600 links to LCSH equivalents

(crosswalks between preferred terms)

Biblioteca Nazionale Centrale Firenze
Nuovo soggettario

RAMEAU
Répertoire d'autorité-matière encyclopédique et alphabétique unifié

7,000 links to RAMEAU equivalents

(from NS Thesaurus preferred terms)

Automatic Indexing at the BNCF

Goals

- Adaptation of the traditional cataloguing tools to the online resources
- Saving financial and human resources
- Matching the Thesaurus to terms extracted from texts on the basis of semantic relevance with the aim of creating keywords

Partners

- BNCF
- Casalini libri
- @cult

Tests at BNCF (1)

First step

- Defining workflow
- Choosing a set of PDF documents
- Formatting of results
- Identifying the methods of using the Thesaurus for extraction procedures

Tests at BNCF (2)

Second step

- Using KEA (Keyphrase Extraction Algorithm) for the extraction of keywords in the texts. Some features are calculated for each candidate term extraction:
  - TF/IDF
  - Distance of a phrase from the beginning of a document (the number of words that precede its first appearance, divided by the number of words in the document)
  - Length of phrases (usually from one to three words)
  - Matching the presence of the terms with controlled vocabularies

- Using open source software (Apache Tika, Maui, JBoss, MySql)
- Constructing and installing the Keyword Indexer (KI) application
- Reviewing of SKOS/RDF format of the Thesaurus by the BNCF

Learning Schemes

Knowledge bases to compute the frequency and the meaning of the terms which are extracted from a set of training documents and then used as a basis of comparison for controlled keywords extracted from new documents which are later indexed

Learning Schemes at the BNCF

Method used

Selection of a set of doctoral theses (Italian language, PDF, full text, abstracts and MIUR classes)

Schemes created

- Variable numbers of the theses used (from a minimum of 100 to a maximum of almost 500)
- Multidisciplinary or specific domain (e.g. Engineering and Architecture)
- Italian abstract
- Use of the NS Thesaurus in SKOS/RDF (matching with preferred terms or all terms)
- Stopwords

Indexed resources

1. Issues of Italian journals edited by the Florence University Press
2. Papers of the Carlo Cattaneo University (*LIUC Papers*)
3. Monographs (inside BNCF Teca and other institutions) in digital format

Results and Open Issues

The results obtained are not yet satisfactory

Aspects to be explored:

- Method used for the creation of the Learning Models
  - *Efficacy of KEA*
  - *Type of intellectual intervention*
  - *Opportunity to use multidisciplinary or specialistic models*
  - *Methods of using the Nuovo Soggettario Thesaurus: should it be integrated with the Authority file of proper and geographic names and/or with the subject strings archive of the BNCF OPAC?*
- Choices regarding the language of the texts, metadata, thesauri
- Problems related to formal presentation of the texts

The automatic indexing project cannot ignore intellectual contribution

The Deutsche Nationalbibliotek Petrus Project

Goals

- Reduce the cost of cataloguing using already assigned metadata
- Automatic assignment of the Schalgwortnormadatei (SWD) terms to digital resources

Steps

- Acquire and license the Averbis Extraction Platform
- Start tests for German language texts and plan for the same with English language texts

Averbis Extraction Platform

What does Averbis do?

1. Analysis of the online publications based on the extraction of the terms taken from textual parts and titles
2. Ranking the extracted terms according to their meaning and importance
3. Matching the extracted terms to the SWD controlled vocabulary

Two main components in the system

- Averbis Concept Mapper, tool based on a dictionary. It combines method of learning machine with morphological and syntactical analysis. The dictionary allows for the integration of synonyms and various attributes for terms, e.g. classificatory information
- Dictionary Configurator, user interface to create and modify user-specific concepts regarding the dictionary

DNB Tests to Measure the Quality of the Automatically Generated Subject Headings

Evaluation of the intellectual results by subject indexing specialists

Creation of an evaluation database

Results achieved are not yet satisfactory

- ambiguous terms
- no discrimination between names of people and topical terms
- Etc.

The DNB is working to find acceptable solutions

Lucas Cranach der Ältere

Das goldene Zeitalter
1530 ca

Thanks !

Anna Lucarelli *and* Elisabetta Viti

anna.lucarelli@beniculturali.it
elisabetta.viti@beniculturali.it

IRCDL 2016, Florence 4-5 February