Thesaurus del Nuovo soggettario fra linked data, prove di indicizzazione automatica e altri sviluppi

Anna Lucarelli - Elisabetta Viti
(Biblioteca nazionale centrale di Firenze)
anna.lucarelli@beniculturali.it elisabetta.viti@beniculturali.it

Argomenti trattati

Obiettivo dell'intervento è presentare le iniziative avviate alla Biblioteca nazionale centrale di Firenze (BNCF), grazie alle potenzialità del Thesaurus del Nuovo soggettario (NS), per la sperimentazione di procedure di indicizzazione automatica di risorse digitali, tenendo conto anche delle esperienze in corso presso biblioteche nazionali di altri Paesi.

Parole chiave: Indicizzazione automatica, Indicizzazione per sogggetto, Linked data, Nuovo soggettario.

1. La BNCF: obblighi catalografici e strumenti di indicizzazione semantica

La Biblioteca nazionale centrale di Firenze, per compito istituzionale, conserva, tutela e diffonde la memoria culturale ed il patrimonio documentario del nostro Paese, coordinandone l'informazione e giocando un ruolo importante anche a livello internazionale con il mantenimento e la diffusione di bibliografie nazionali e di strumenti di organizzazione della conoscenza (classificazioni, thesauri, authority files ecc.).

Al suo ricco patrimonio tradizionale si aggiunge da qualche anno un sempre più considerevole corpus di risorse digitali (sia *born-digital* che non) che devono essere prese in conto, sia dal punto di vista della conservazione, che del trattamento catalografico. L'incremento di acquisizioni di materiali digitali deve essere affrontato razionalizzando risorse e costi e prevedendo modalità che riducano il più possibile interventi di tipo umano/intellettuale.

Come sta avvenendo in biblioteche nazionali di altre parti del mondo, dobbiamo lavorare per soddisfare con prodotti più attuali le esigenze informative e le aspettative di un'utenza che – abituata a navigare sul web e ad interrogare motori di ricerca – si accosta ai cataloghi con atteggiamento nuovo ed esigenze mutate. Non si tratta più di mettere a disposizione soltanto opere, testi e cataloghi ma anche i nostri metadati in formati leggibili

dalla macchina favorendone così, anche in rete, visibilità, recupero, esportazione, riusabilità[1]. La Biblioteca nazionale centrale di Firenze ha inaugurato già da qualche anno le sue esperienze di *open data* con il Thesaurus del *Nuovo soggettario.*

2. Il Thesaurus del Nuovo soggettario

Il settore "Ricerche e strumenti d'indicizzazione semantica" della BNCF si dedica all'allestimento e sviluppo di uno strumento che consiste in un sistema articolato di regole concernenti sia aspetti terminologici (il Thesaurus) sia modalità di costruzione delle stringhe di soggetto[2]. Disponibile dal 2007, può essere usato per l'indicizzazione semantica di risorse bibliografiche, archivistiche e museali. Il cuore del sistema è costituito dal citato Thesaurus (http://thes.bncf.firenze.sbn.it/ricerca.php), un vocabolario controllato multidisciplinare con interfaccia anche in inglese, costruito sulla base di standard internazionali, liberamente accessibile sul web, in continuo accrescimento e con un patrimonio terminologico, ad oggi, di circa 57.000 termini. Non contiene nomi propri di autori, organizzazioni, titoli luoghi, validati negli archivi della Bibliografia nazionale italiana.
I termini del Thesaurus sono organizzati sulla base di una struttura categoriale e di relazioni semantiche (gerarchiche, sinonimiche e associative), sono corredati da note di vario tipo, da fonti repertoriali, ecc.

3. SKOS, interoperabilità e multilinguismo

Stiamo lavorando per incrementare il multilinguismo e l'interoperabilità del Thesaurus con data set di altre "istituzioni della memoria" e di amministrazioni pubbliche impegnate in analoghi progetti di linked open data. In quest'ottica, è stato sviluppato il colloquio semantico e tecnico del Thesaurus con altri sistemi di indicizzazione (non solo italiani) e con strumenti lessicografici e repertoriali, tramite procedure in certi casi intellettuali, in altre semi-automatiche. La disponibiltà dei suoi metadati in formato SKOS/RDF lo rende attivo nell'universo del web semantico e dei linked data[3]. Come è noto, SKOS associa (tramite il predicato *rdf:type*) concetti espressi dai termini a classi specifiche (*skos:Concept*) grazie all'assegnazione di identificativi univoci e facilitando così mappature e interoperabilità tra

1 Sull'argomento anche Anna Lucarelli, *Web dei dati alla Biblioteca nazionale centrale di Firenze,* (intervento presentato al Salone del libro di Torino il 15 maggio 2015 nell'ambito della Tavola rotonda Biblioteche digitali verso il futuro: accesso ai linked open data - Book to the future), *"DigItalia",* (in corso di pubblicazione)

2 http://thes.bncf.firenze.sbn.it/index.html

3 Sull'argomento già Giovanni Bergamin – Anna Lucarelli, *The Nuovo soggettario as a service for the linked data world*, «JLIS», vol. 4, 2013, n. 1 (http://leo.cineca.it/index.php/jlis/article/view/5474/7903)

Knowledge Organization Systems (KOS).

Il software impiegato consente di gestire equivalenti in altre lingue; abbiamo dato precedenza a quelli inglesi, previsti da *Library of Congress Subject Headings* (LCSH) e a quelli francesi del *Répertoire d'autorité-matière encyclopédique et alphabétique unifié* (RAMEAU). Gli equivalenti attivano link ai rispettivi strumenti di indicizzazione online, offrendo agli utenti la possibilità di navigare tra patrimoni documentari di prestigiose istituzioni culturali italiane e straniere. Ad oggi gli equivalenti con LCSH sono circa 10,600 (molti dei quali con link reciproci), quelli con RAMEAU circa 7,000.

4. Indicizzazione intellettuale vs. Indicizzazione automatica

Un fronte di sviluppo del Thesaurus riguarda l'indicizzazione automatica di risorse digitali.

A livello internazionale, anche nel mondo delle biblioteche, si sta diffondendo l'idea che, soprattutto per certe tipologie di pubblicazioni, l'indicizzazione semantica tradizionale di tipo intellettuale, molto costosa per le istituzioni che la praticano, possa essere sostituibile da procedure di tipo automatico (o semi-automatico), meno dispendiose in una visione a lungo termine[4]. La nuova edizione delle *Guidelines for Subject Access in National Bibliographies* dedicano alla tematica del contenimento dei costi uno specifico spazio e, per raggiungere l'obiettivo di documentare comunque i "soggetti" della crescente produzione editoriale in formato digitale, propongono l'applicazione di procedure di indicizzazione automatica o semi-automatica, da estendere anche a tipologie documentarie come immagini, musica, ecc.[5]

L'indicizzazione automatica di risorse digitali è una procedura in fase di studio e sperimentazione in contesti di lavoro e di ricerca anche diversi fra loro; prima ancora che nelle biblioteche, è stata affrontata da comunità che si occupano di intelligenza artificiale, di linguistica computazionale, ecc. Non possiamo non citare gli studi condotti nell'ambito del CNR di Pisa e del Dipartimento di Ingegneria dell'Informazione dell'Università di Firenze.

Come è noto, esistono varie modalità per automatizzare l'indicizzazione, distinte per il livello di automatismo, per l'impiego o meno di vocabolari controllati, ecc.

4 Il processo di indicizzazione per soggetto, basato sull'individuazione del contenuto concettuale e sulla sua traduzione in un particolare linguaggio di indicizzazione, richiede conoscenze personali, esperienza e una formazione che necessita di risorse non indifferenti. Sugli aspetti cognitivi dell'indicizzazione, si veda anche il recente saggio di Alberto Cheti, *I processi cognitivi nell'analisi concettuale dei documenti. Una postilla tra biblioteconomia e linguistica*, "AIB Studi", 2016, (in corso di pubblicazione)

5 IFLA Working group on guidelines for subject access by National bibliographic agencies, *Guidelines for subject access in National bibliographies*, edited by Yvonne Jahns. Berlin-Boston: De Gruyter Saur, 2012

In ambito catalografico, l'indicizzazione automatica ha una sua utilità se algoritmi e tecniche specifiche possono sostituire o integrare l'intervento umano nell'analisi e indicizzazione di documenti.

La tabella che segue rappresenta un tentativo di mettere a fuoco le principali differenze tra le due procedure d'indicizzazione.

INDICIZZAZIONE UMANA	**INDICIZZAZIONE AUTOMATICA** (o semi-automatica)
Sistema più costoso (tempo, risorse finanziarie e umane)	**Sistema più economico** (tempo, risorse finanziarie e umane)
Catalogatori con formazione specifica	**Sistemi complessi** (*software)* non completamente standardizzati
Uso di regole e di standard	**---------**
Processo cognitivo legato anche alla cultura dell'indicizzatore • Analisi del testo • Comprensione del testo • Interpretazione e individuazione del contenuto concettuale • Traduzione del tema di base in un linguaggio documentario	**Processo meccanico su base statistica** • Analisi del testo • Nessuna comprensione umana del testo
I termini d'indicizzazione sono scelti in vocabolari controllati (soggettari, thesauri, etc.) **La scelta dei termini d'indicizzazione si basa su conoscenza e comprensione**	**I termini d'indicizzazione sono estratti in automatico** (eventuale match con vocabolari controllati) **La scelta dei termini di indicizzazione si basa su parametri statistici e automatici**
Livello alto di Precisione **Livello basso di Richiamo**	**Livello basso di Precisione** *individua parole con grado maggiore di occorrenza a discapito di parole/espressioni meno comuni ma significative* **Livello alto di Richiamo**

5. Prove di indicizzazione automatica in BNCF: obiettivi e partners

Forme di indicizzazione automatica possono dare risultati migliori se agganciate a strumenti di controllo terminologico (mappe topiche o thesauri). In questa direzione vanno ricerche, come ad esempio, quella spagnola (Università di Alcalà) sull'applicabilità di

AGROVOC (thesaurus che copre le aree d'interesse della FAO) nell'estrazione di parole chiave da un campione della collezione open acces AGRIS del repository VOA3R[6].

Il nostro obiettivo di partenza era di provare ad adeguare strumenti catalografici tradizionali al crescente sviluppo del mondo dell'informazione in rete, individuando modalità per razionalizzare risorse umane e finanziarie, con un abbassamento dei costi della catalogazione per soggetto.

Il lavoro si è basato sul tentativo di mettere a punto tecniche per associare il Thesaurus del Nuovo soggettario a termini estratti da testi sulla base della loro rilevanza semantica, con lo scopo di generare parole chiave riferite al contenuto di documenti digitali.

Il lavoro è iniziato dalla collaborazione fra tre partner (uno pubblico e due privati):

- la BNCF che ha messo a disposizione le proprie risorse online, ha sperimentato le procedure di estrazione automatica di parole chiave e, tramite catalogatori esperti di indicizzazione, ha valutato i risultati;
- Casalini libri, azienda che da tempo si occupa di editoria digitale ed è interessata all'indicizzazione automatica con il Nuovo soggettario
- @cult, azienda che ha avviato attività nel campo del web semantico (RDF e Linked data) e dei metadati bibliografici. Il suo ruolo nel progetto è stato quello di specificare i requisiti funzionali e, sulla base di librerie e applicazioni open source, di mettere a punto un'applicazione che è stata chiamata KI (Keyword Indexer).

6. Prime sperimentazioni d'indicizzazione automatica in BNCF

La prima fase (dicembre 2010 - ottobre 2011) è consistita nella definizione delle procedure (scelta di un set di documenti in formato PDF, formato dei risultati, flussi di lavoro, ecc.) e nella scelta del Thesaurus del Nuovo soggettario come componente base per le procedure di estrazione.

La seconda (novembre 2011 - dicembre 2013), si è articolata in varie tappe.

Per l'estrazione di frasi chiave dai testi, è stato scelto l'algoritmo KEA (Keyphrase extraction algorithm),[7] realizzato in un'università neozelandese ed usabile per indicizzazioni sia libere che supportate da vocabolari controllati[8]. KEA utilizza TF/IDF, cioè

6 *Evaluating the practical applicability of thesaurus-based keyphrase extraction in the agricultural domain: insights* from the VOA3R Project, "Knwoledge Organization. Internationa Journal", vol. 42 (2015), n. 2, p. 76-89

7 http://www.nzdl.org/Kea/

8 KEA è un algoritmo diffuso e, in associazione a thesauri (come AGROVOC, MESH, HEP), ampiamente sperimentato. Notizie sul tema si trovano anche in: *Domain-Specific Keyphrase Extraction* (http://ijcai.org/Past%20Proceedings/IJCAI-99%20VOL-2/PDF/002.pdf) e nel più recente paper di

il rapporto tra la frequenza di un termine all'interno di un documento, e quella del termine all'interno dell'intero set di documenti; impiega la distanza di una frase (composta da uno più termini) dall'inizio del testo del documento fino alla sua prima occorrenza (i termini presenti all'inizio o alla fine del documento sono solitamente più rilevanti);[9] prende in considerazione anche la lunghezza della frase e verifica infine la presenza del termine all'interno di vocabolari controllati.

La realizzazione della componente informatica da parte della società @cult ha comportato la scelta del software open source (Apache Tika, Maui, JBoss, MySql) e la realizzazione ed installazione dell'applicazione Keyword Indexer (KI)[10].

Per la BNCF è stata anche l'occasione per rivedere la versione SKOS/RDF del Thesaurus.

Tramite il software realizzato, sono stati creati “modelli di apprendimento” e avviate, infine, prove di indicizzazione automatica.

I “modelli di apprendimento” consistono in una sorta di base di conoscenza finalizzata a misurare frequenza e significatività dei termini estratti da un set di documenti di dominio (attribuendo loro un peso in funzione di alcuni parametri) e successivamente utilizzabili come base di confronto per l'estrazione di parole chiavi pertinenti e controllate dalle risorse da indicizzare.

Nella creazione dei modelli possono essere impiegati anche set di metadati associati ai documenti (parole chiave o classificazioni assegnate manualmente). Possono, inoltre, essere specificati alcuni parametri, tra cui i più importanti sono: il vocabolario controllato da usare; l'eventuale stemming; la lingua.

7. Metodo seguito

Per la creazione del “modello di apprendimento” abbiamo scelto un campione di tesi di dottorato (in lingua italiana), in formato PDF, raccolte negli open archive delle Università e acquisite dalla BNCF tramite una procedura di harvesting (protocollo OAI-PMH). Le tesi in full text sono corredate da abstract (in italiano e/o inglese) e dai record bibliografici che descrivono il documento e la classe disciplinare di afferenza.

Dal corpus individuato per la sperimentazione sono state escluse tesi con elementi grafici, formule matematiche, statistiche ecc.

Inizialmente, per la creazione della base di conoscenza, si pensava di utilizzare descrittori

ricercatori spagnoli già citato (vedi nota 4)

9 La distanza è calcolata come il numero delle parole che precedono la prima apparizione di una frase diviso il numero delle parole nel documento.

10 La metodologia seguita è descritta nel dettaglio nel report prodotto da @cult, *Procedura automatizzata di estrazione parole e frasi chiave. Specifiche tecnico-funzionali*, (documento interno).

assegnati intellettualmente a risorse cartacee digitalizzate. In realtà, abbiamo optato per procedure esclusivamente automatiche, acquisendo unicamente i metadati semantici assegnati dalla classificazione disciplinare del MIUR. E' stata creata una tabella di corrispondenza tra la decodifica verbale della classificazione ed i termini del Thesaurus.
È stato scelto il Thesaurus come componente base per le procedure di estrazione.

8. Modelli creati

Abbiamo creato diversi modelli:

Modello A)

- 200 tesi afferenti a varie discipline
- Abstract in lingua italiana
- Metadati semantici MIUR
- Thesaurus del Nuovo soggettario in formato SKOS/RDF (solo termini preferiti)

Modello B)

- 100 tesi afferenti a uno specifico ambito disciplinare (Area 08 - Ingegneria civile e Architettura)
- Abstract in lingua italiana
- Metadati semantici MIUR
- Thesaurus del Nuovo soggettario in formato SKOS/RDF (solo termini preferiti)

I Modelli A e B sono stati duplicati ed integrati impiegando il Thesaurus non solo con i termini preferiti, ma anche con i relativi sinonimi.

Modello C)

- 436 tesi scelte con procedure intellettuali per coprire in modo uniforme tutti gli ambiti afferenti ai 14 domini del MIUR
- Abstract in lingua italiana
- Metadati semantici MIUR
- Thesaurus del Nuovo soggettario in versione SKOS/RDF (solo termini preferiti);

Tutti i modelli hanno utilizzato stopwords (articoli, preposizioni, pronomi/aggettivi dimostrativi, indefiniti numerali, avverbi, etc...).

9. Risultati e problemi aperti

Utilizzando i modelli descritti sono stati indicizzati: alcuni fascicoli di periodici scientifici italiani pubblicati da Firenze University Press; alcuni papers dell'Università Carlo Cattaneo (*LIUC Papers*); alcune monografie in formato digitale (della Teca BNCF e di altre istituzioni). I modelli impiegati (multidisciplinari o specialistici) hanno previsto dunque sia un impiego completo del Thesaurus, sia un impiego parziale, limitato ai soli descrittori in forma preferita.

I risultati ottenuti non sono ancora soddisfacenti e per questo il nostro percorso di sperimentazioni non è affatto concluso. Vorremmo continuare i lavori, partendo da un'analisi degli aspetti da approfondire; solo per citarne alcuni:

- il metodo seguito nella creazione dei “modelli di apprendimento”, definendo meglio:
    - l'efficacia dell'algoritmo KEA
    - il tipo di intervento intellettuale necessario (ad esempio, nell'attribuzione dei metadati semantici)
    - l'opportunità di usare modelli multidisciplinari o specialistici
    - la modalità di impiego del Thesaurus del Nuovo soggettario, considerando la possibilità di affiancarlo a liste di autorità di altri termini (nomi propri, geografici, ecc.), oppure con l'archivio delle intestazioni per soggetto del nostro Opac
- le scelte inerenti la lingua dei testi, dei metadati, dei vocabolari controllati
- i problemi legati alla presentazione formale dei testi.

Certo è che l'esperienza condotta sinora in BNCF ha confermato che qualsiasi progetto di indicizzazione automatica non può prescindere dal contributo intellettuale e umano di esperti di indicizzazione, come del resto dimostrano esperienza straniere che stiamo valutando anche in vista di ulteriori sviluppi del progetto.

A livello nazionale e internazionale, negli ultimi anni, varie biblioteche, istituti di ricerca e network hanno provato ad avviare progetti di indicizzazione automatica; solo per citarne alcuni: il progetto inglese MERLIN,[11] quello della Bibliothéque nationale de France,[12] della

11 http://www.ucl.ac.uk/ls/merlin/about.shtml

12 Gildas Illie, *Le dépôt légal de l'internet en pratique: les moissonneurs du web*, “BBF- Bulletin des

Deutsche Nationalbibliothek, della rete RERO a cui afferiscono le biblioteche della Svizzera occidentale,[13] ecc.

10. Il progetto di indicizzazione automatica della Deutsche Nationalbibliothek (DNB)

Dal 2010 la DNB ha avviato il progetto PETRUS per l'indicizzazione automatica di risorse monografiche online e tesi di dottorato acquisite per deposito legale, con l'obiettivo di risparmiare i costi della catalogazione e utilizzare il più possibile metadati già assegnati[14].
Obiettivo di PETRUS è l'attribuzione automatica a risorse digitali di termini (parole chiave) previsti dal vocabolario controllato *Schalgwortnormadatei* (SWD), usato in Germania e in altri paesi germanofoni per l'indicizzazione semantica delle risorse cartacee.
SWD è in forma di thesaurus e contiene vari gruppi di intestazioni: termini relativi a concetti, nomi geografici ed etnografici, nomi di enti e titoli di opere.
Dopo un periodo di analisi di sistemi vari, la Deutsche Bibliothek ha deciso di acquisire Averbis Extraction Platform, un software disponibile sul mercato, sviluppato dall'azienda Averbis (che ha sede a Friburgo),[15] e di iniziare la sperimentazione a partire da testi in tedesco, ma prevedendola anche per testi in inglese.
Il processo eseguito dalla piattaforma Averbis Extration Platforme consiste di queste fasi:

1) analisi testuale delle pubblicazioni online, con estrazione di termini dai contenuti e titoli delle pubblicazioni online;
2) raggruppamento dei termini estratti secondo significato ed importanza;
3) verifica dei termini estratti all'interno del vocabolario controllato SWD

Il sistema si basa su due componenti:

1) Averbis Concept Mapper, uno strumento basato su un dizionario che combina metodi di apprendimento automatico (machine learning) (sulla base di algoritmi specifici) con analisi

bibliothèques de France", 2008, n. 6 (http://bbf.enssib.fr/consulter/bbf-2008-06-0020-004).

13 https://www.rero.ch/pdfview.php?section=infos&filename=RERO_IM_complement_20110707.pdf

14 http://www.dnb.de/DE/Wir/Projekte/Abgeschlossen/petrus.html. Un resoconto sul progetto è stato presentato nell'ambito di IFLA 2012 e successivamente pubblicato in: Ulrike Junger, *Can indexing be automated? The example of the German National Library*, "Cataloging & Classification Quarterly", vo. 52, no. 1, 2014, p. 102-109 (http://www.catalogingandclassificationquarterly.com/ccq52nr1.html); si veda anche Reinhard Altenhöner, *Access to knowledge: Text mining and information extraction in the German National Library* (https://www.youtube.com/watch?v=g5gqCtf3sYg). Per informazioni recenti anche: Elke Jost-Zell, *News from the German National Library (DNB)*, "IFLA Metadata Newsletter", vol. 1, n. 2, December 2015, p. 34-37 (http://www.ifla.org/files/assets/cataloguing/scatn/metadata-newsletter-201512.pdf).

15 https://averbis.com/. Averbis dichiara di basarsi su Apache UIMA (Unstructured Information Management Architecture) ovvero su un diffuso framework per il trattamento e l'analisi di informazioni semi-strutturate, e l'estrazione di informazioni da esse.

morfologica e sintattica. Il dizionario è flessibile e permette l'integrazione di sinonimi e vari attributi per i termini, per es. informazioni classificatorie.

2) Il Dictionary Configurator: un interfaccia utente per creare e modificare concetti specifici usati dall'utente e che riguardano il dizionario.

Varie sezioni di SWD/PND sono state integrate nel dizionario come base per l'assegnazione di intestazioni di soggetto. Ad oggi ci sono: 170.000 argomenti relativi a soggetti, 153.000 registrazioni di nomi geografici ed etnografici, 311.000 registrazioni di nomi di persone.

Come già detto, il vocabolario così costruito può essere configurato con Averbis Dictonary Configurator che permette la realizzazione di concetti d'indicizzazione confezionati per certi gruppi di oggetti/pubblicazioni. Il dizionario è in continuo aggiornamento.

Dopo l'acquisizione di Adverbis sono stati condotti test per individuare le configurazioni che portassero risultati migliori. Inoltre nei primi test sono stati integrati nel dizionario solo gli argomenti e i nomi geografici e solo successivamente i nomi di persona.

Il progetto, infine, ha previsto vari test per misurare la qualità delle intestazioni di soggetto generate automaticamente: si è trattato di una valutazione intellettuale dei risultati, fatta da specialisti di indicizzazione. È stato creato un database di valutazione contenente l'autore, il titolo, link al full text del documento e la lista di intestazioni per soggetto assegnate automaticamente. Ogni intestazione ha avuto un giudizio espresso con una scala da 1 a 4; sono state segnalate intestazioni o aspetti mancanti, e sono stati valutati parametri come il livello di precisione e di richiamo, altri parametri sulla completezza delle intestazioni.

I test hanno prodotto risultati non del tutto soddisfacenti, a causa di errori e problemi legati ad ambiguità semantiche, al fatto che le SWD (essendo un vocabolario generale) contengono molti termini generali che frequentemente occorrono nei documenti ma senza un significato specifico, ecc., ed infine anche alla difficoltà di discriminare omografi relativi a nomi di persona, nomi geografici e nomi comuni.

I nostri colleghi tedeschi hanno concluso che i metodi da loro usati per calcolare ranking e affidabilità debbano essere rivisti e stanno studiando soluzioni. Se, da un lato, hanno chiaro che non sia facile sviluppare ed implementare processi di indicizzazione automatica usando vocabolari controllati generali, dall'altro sono convinti che queste procedure possano comunque essere accettabili, pur non producendo gli stessi risultati dell'indicizzazione intellettuale.

Ultima consultazione siti web 28/01/2016