UNIVERSITÀ DEGLI STUDI FIRENZE

FLORE
Repository istituzionale dell'Università degli Studi di Firenze

Ontologie e linguaggi biblioteconomici d’indicizzazione / Mauro Guerrini

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

Original Citation:
Ontologie e linguaggi biblioteconomici d’indicizzazione / Mauro Guerrini / Mauro Guerrini. - STAMPA. - (2013), pp. 13-26.

Availability:
This version is available at: 2158/870918 since: 2018-08-21T12:05:17Z

Publisher:
EDUCatt - Ente per il Diritto allo Studio Universitario dell'Università Cattolica

Terms of use:
Open Access
La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf)

Publisher copyright claim:

(Article begins on next page)

27 April 2024

TERMINOLOGIE E ONTOLOGIE

Definizioni e comunicazione fra norma e uso

a cura di
Patrizia Collesi
Anna Serpente
Maria Teresa Zanola

EDUCatt

TERMINOLOGIE E ONTOLOGIE
Definizioni e comunicazione fra norma e uso

a cura di
Patrizia Collesi
Anna Serpente
Maria Teresa Zanola

Milano 2013

© 2013 **EDUCatt** – Ente per il Diritto allo Studio Univeùdell'Università Cattolica
Largo Gemelli 1, 20123 Milano – tel. 02.7234.22.35 – fax 02.80.53.215
e-mail: editoriale.dsu@educatt.it (produzione); librario.dsu@educatt.it (distribuzione)
web: www.educatt.it/libri
ISBN: 978-88-6780-013-1

copertina: progetto grafico Studio Editoriale EDUCatt. Quarta di copertina: foto di Andrea Aschedamini.

INDICE

Maria Teresa Zanola
Vent'anni dell'Associazione Italiana per la Terminologia 5

Mauro Guerrini
Ontologie e linguaggi biblioteconomici d'indicizzazione 13

Anna Serpente
Terminologia e ontologia: organizzazione della conoscenza e comunicazione in campo medico. Esempi di schede terminologiche nella malattia di Alzheimer, dal ricercatore al paziente 27

Annibale Elia, Federica Marano, Mario Monteleone, Johanna Monti, Antonella Napoli, Alberto Postiglione, Daniela Vellutino
Applicazioni di NLP per il web semantico: il trattamento delle polirematiche terminologiche nel dominio della medicina 39

Elena Cardillo
Formalizzazione ontologica di terminologie e sistemi di classificazione in ambito biomedico: il caso ICPC2-ICD10 51

Sabrina Aulitto
Terminologie e ontologie: studi francesi 65

Donatella Pulitano
Terminologia e qualità 79

Nicola Poeta
Terminologia, corpora e contesto negli strumenti di traduzione assistita 87

Simone Torsani
Ipotesi sull'applicazione del modello relazionale a corpora linguistici 95

Mauro Guerrini[1]

ONTOLOGIE E LINGUAGGI BIBLIOTECONOMICI D'INDICIZZAZIONE[2]

Premessa

Filosofi, linguisti e terminologi si occupano dei concetti, della loro categorizzazione e, in particolare, del rapporto che intercorre con i termini e le unità linguistiche (frasi e discorsi) per cercare di spiegare le modalità di acquisizione e trasmissione della conoscenza[3]. Costoro elaborano procedure di analisi concettuale, categorizzazione e uso di terminologia semanticamente controllata propria dei modelli concettuali, che sono alla base dei linguaggi d'indicizzazione e delle classificazioni bibliografiche.

La motivazione filosofica, tuttavia, non rende, da sola, spiegazione su una tendenza sempre più evidente nell'approccio alla gestione delle informazioni, ovvero a giustificare una serie di *fenomeni* che hanno determinato un sostanziale cambiamento nell'approccio all'informazione presente sul web: quelli che dalla pratica della soggettazione e classificazione tipica dell'ambito biblioteconomico arrivano alle ontologie e al web semantico.

1. L'evoluzione del web

Come sappiamo, il web ha avuto finora tre fasi evolutive: nascita nel 1991, grazie alle intuizioni di Tim Berners-Lee; il web 2.0 o *social web*; il web 3.0 o *web semantico*. L'evoluzione del web esprime un'esigenza antica: classificare il sapere per governarne quantità e disordine.

I primi sistemi di classificazione cercavano di creare criteri di selezione

---

[1] Università di Firenze.

[2] Ringrazio Valdo Pasqui, Tiziana Possemato ed Elisabetta Viti per la collaborazione alla stesura della relazione. I siti web sono stati consultati l'ultima volta il 24 maggio 2011.

[3] R. Job – R. Rumiati, *Linguaggio e pensiero*, Il Mulino, Bologna 1984, cap. 3.

per individuare *la cosa giusta nel tempo giusto* in un contesto ancora contenuto di disseminazione dell'informazione. L'obiettivo di questi strumenti, realizzati in termini di sistemi classificatori, thesauri, soggettari, è sempre stato introdurre un vocabolario controllato nel mondo della conoscenza.

Il web 2.0 esprime (o esprimeva) la medesima esigenza: creare dei puntatori a una quantità molto vasta di informazioni prodotte e sparse sulla rete, facendolo però 'a parole mie', introducendo, dunque, un concetto di classificazioni dal basso (le folksonomie), ovvero, afferma *Wikipedia*, «una categorizzazione di informazioni generata dagli utenti mediante l'utilizzo di parole chiave (o tag) scelte liberamente». Le folksonomie si pongono in opposizione alle *tassonomie* in quanto a criteri di creazione, controllo e attribuzione, ma non in opposizione, se viste come la risposta a una medesima esigenza; si tratta cioè di due mondi apparentemente opposti ma in realtà *consanguinei*.

Il web 3.0, o web semantico, è un ulteriore tentativo di governare l'enorme quantità di informazioni disponibili su web (diverse per qualità informativa, per supporti, per origini e finalità di pubblicazione); esso non può essere definito una realtà esistente e tangibile ma piuttosto come un *ambito di ricerca* proposto ancora da Tim Berners-Lee (poco affezionato al web 2.0, come ricorda Tim O'Reilly), è basato sulla considerazione del *gap semantico* che si è prodotto tra le parole come 'simboli' e il loro significato, e sul tentativo di ridurre tale gap.

Katrin Weller, in *Knowledge Representation in the Social Semantic Web*[4], afferma che scopo del web semantico è ridurre la distanza «between computers, that match word strings and humans who interpret meanings of words».

Per *semantic web* s'intende «la trasformazione del World Wide Web in un ambiente in cui i documenti pubblicati (pagine HTML, file, immagini, etc.) siano associati a informazioni e dati (metadati) che ne specifichino il contesto semantico in un formato adatto all'interrogazione, all'interpretazione e, più in generale, all'elaborazione automatica»[5].

La sfida principale è costituita dalla possibilità di individuare informazioni rilevanti all'interno di collezioni digitali integrate poiché è necessario

[4] K. Weller, *Knowledge Representation in the Social Semantic Web*, De Gruyter Saur, München 2010.

[5] Da Wikipedia, voce 'web semantico', http://it.wikipedia.org/wiki/Web_semantico.

combinare metodi tradizionali e nuovi approcci per il reperimento dell'informazione[6]. La ricerca nel web mostra che i vocabolari controllati e i linguaggi di indicizzazione continuano ad avere un ruolo fondamentale tramite la funzione collocativa (di raggruppamento) nell'organizzazione delle informazioni.

2. Ontologia e web semantico

Il concetto delle *ontologie* è strettamente relazionato al mondo del web semantico. Il web è un enorme contenitore di risorse informative, ma tali risorse mancano di collegamenti che connettano un'informazione all'altra in senso semantico; ciò produce un'incapacità del web di rispondere a domande che vadano oltre la mera ricerca formale di un termine (o dei suoi derivati evidenti) all'interno di una pagina, ed espresse, per lo più, con linguaggio naturale.

Per costruire *relazioni semantiche* tra concetti e ottenerne un utilizzo *automatico* (quindi in un formato leggibile da una macchina) è necessaria una *formalizzazione* – l'ontologia – che le descriva nel modo più completo possibile. Le *ontologie* rappresentano, dunque, quello *strato semantico* del web finalizzato a rendere possibili connessioni di significato tra concetti (e risorse). In ciò consiste quel connubio tra metodi tradizionali e nuovi approcci per il reperimento delle informazioni.

Nella prospettiva di organizzazione dei dati offerta dal web semantico, le funzioni e le modalità di applicazione, soprattutto dei thesauri, si sono ampliate, in due direzioni:

1) Reingegnerizzazione delle strutture thesaurali, come base per lo sviluppo di ontologie formali[7].

---

[6] A. Slavic, *On the nature and typology of documentary classifications and their use in a networked environment*, in *El profesional de la información*, vol. XVI, f. 6, 2007, pp. 580-589. Vedi anche: http://webcache.googleusercontent.com/search?q=cache:evpByxN7gDoJ:www.elprofesionaldelainformacion.com/contenidos/2007/noviembre/05.pdf+slavic+on+the+nature+and+typology&cd=1&hl=it&ct=clnk&gl=it.

[7] Una prima sperimentazione è stata lo sviluppo del thesaurus AGROVOC: le relazioni genere/specie e tutto/parte sono state ampliate con relazioni di causa, somiglianza e dissimiglianza, strumenti e processi. Cfr. M.T. Biagetti, *Le ontologie come strumenti per l'organizzazione della conoscenza in rete*, «AIDAinformazioni», 28 (gen.-giu. 2010), 1-2; D. Sorgel *et al.*, *Reeingenering Thesauri for New Applications: the AGROVOC example*, «Journal of

2) Impiego ampio delle strutture thesaurali, come base di conoscenza, per l'indicizzazione automatica delle risorse digitali.

L'uso dei thesauri per l'indicizzazione automatica di record di documenti è tuttavia consolidato da anni. Oggi è cambiato il modo di far utilizzare l'informazione indicizzata: i *link* (o relazioni) sono presentate in modo molto più intuitivo, anche grafico, dal momento che è stato assimilato il modello navigazionale del web. Del resto il *www* è nato dal concetto di *hypertext*.

3. I linguaggi specialistici e la lingua comune

L'evoluzione sociale, mediante l'ampliamento dell'istruzione e la crescita dei sistemi di comunicazione e delle tecnologie, ha favorito la diffusione della conoscenza specialistica: sono aumentati i domini disciplinari e le microspecializzazioni, che hanno sviluppato settori di conoscenza nuovi, con applicazioni interdisciplinari e transdisciplinari.

I linguaggi specialistici sono una varietà funzionale di un sistema linguistico naturale e generale. Le componenti che caratterizzano la comunicazione specialistica sono:

1) la terminologia;
2) i testi, che sono strutturati e organizzati secondo un modello determinato dal contesto (per esempio, gli atti amministrativi).

La terminologia specialistica rappresenta il complesso delle conoscenze scientifiche di un settore di studio, ma anche un codice orale spontaneo che prende vita dalle esigenze di comunicazione, come la terminologia impiegata in alcuni *blog* (per esempio *shortare* che deriva da *short selling* = vendita allo scoperto)[8], oppure in relazione a eventi storici (per esempio, la terminologia legata alla crisi finanziaria).

Tre sono i meccanismi tipici di questo mondo espressivo, diviso tra i due livelli:

---

digital information», vol. IV, n. 4, 2004.

[8] D. Maldussi, *Terminologia sistematizzata e terminologia spontanea in ambito finanziario: un rapporto conflittuale*, Atti del Convegno Assiterm 2009, vol. VIII, 2010. http://publifarum.farum.it/ezine_printarticle.php?id=173.

– quello alto (formale, controllato, strutturato);
– quello basso o di massa (quello parlato, che si adatta in modo veloce ai cambiamenti della realtà e che crea o trasforma la terminologia associata ai concetti).

I tre meccanismi hanno la stessa finalità di *avvicinare mondi consanguinei* e non estranei l'uno all'altro e possono essere così esemplificati:

a. Opposizione e complementarietà tra *folksonomy* (le categorizzazioni 'dal basso', espresse per lo più in linguaggio naturale) e *taxonomy* (le categorizzazioni controllate, formalmente relazionate e strutturate).
b. Gli interventi di aggiornamento e *manutenzione* dei vocabolari controllati (thesauri, sistemi di classificazione, etc.), che devono periodicamente assorbire nuovi termini, dichiarare obsoleti termini desueti, relazionare concetti, rivedere usi ed espressioni, per rendere più attuale e aggiornato quanto in essi contenuto.
c. Creazione di ontologie e loro utilizzo in ambito informatico: la logica del web semantico è proprio quella di creare, utilizzando vocabolari controllati in formati comprensibili da macchine, un linguaggio quanto più vasto e relazionato per esprimere concetti, in forma quanto più eterogenea, considerando, dunque, sia terminologia ed approccio scientifico, sia linguaggio naturale, non controllato.

Nella lingua comune, designare concetti differenti tramite una stessa parola è un elemento fondamentale di economia linguistica, a differenza dei linguaggi specialistici in cui la terminologia è caratterizzata dai seguenti requisiti funzionali:

- Rapporto biunivoco tra termine e concetto designato.
- Oggettività.
- Economia, intesa come l'equilibrio tra la necessità di differenziare i singoli elementi linguistici e la capacità da parte dei destinatari di elaborarli con il minimo sforzo; per esempio, l'uso di simboli in luogo della forma sciolta: PIL per 'prodotto interno lordo')[9].

Nel mondo odierno, i fattori 'moda' e 'media' contribuiscono a svincolare la terminologia specialistica dagli originari confini semantici, e a favorire la

[9] F. Scarpa, *Terminologia e lingue speciali*, in M. Magris *et al.* (a cura di), *Manuale di terminologia: aspetti teorici, metodologici e applicativi*, Hoepli, Milano 2006, pp. 27-47.

sua irruzione nel linguaggio comune e la sua assunzione trasversale in altri contesti scientifici, nonché ad acquisire poliedricità concettuale[10]. Accade inoltre che termini acquisiti da un sistema generale siano usati in un dominio disciplinare specifico e, a distanza di tempo, riacquisiti in contesti generali perdendo però alcune caratteristiche peculiari antecedenti al processo di terminologizzazione (per esempio, *mouse*).

4. Ontologia formale fondazionale e tecnologie dell'informazione

La necessità di scambiare dati tra diverse comunità scientifiche, culturali e linguistiche ha fatto nascere l'esigenza di un colloquio semantico tendenzialmente privo di ambiguità terminologica, e di impiegare strutture concettuali in grado di fornire una rete di relazioni variegate e dettagliate.

Nell'ambito del web semantico hanno assunto maggiore rilievo le ontologie. Katrin Weller ne fornisce questa definizione:

> Ontologies can represent a vocabulary for a domain of interest in form of general *concepts* in this domain, particular *instances* that embody these concepts, *relations* which describe the properties of concepts, and *axioms* that capture general facts about them.

In altri termini, un'ontologia permette di definire in modo significativo i concetti e le relazioni che caratterizzano un certo ***dominio di conoscenza***.

Un altro concetto determinante. Si parla, per esempio, di 'carriera' sia riferendosi al personale sia agli studenti, ma vi sono differenze profonde. Inoltre, si può reperire sul mercato un programmatore esperto che ha piena padronanza degli strumenti, ma fino a quando non si cala nello specifico dominio di conoscenza e non ne apprende la terminologia e i concetti, ovvero se non impara il linguaggio necessario per capirsi con gli addetti ai lavori, non è in grado di produrre risultati. Un problema tipico di ogni *softwarehouse*, quando consegna un'applicazione, è lo scostamento tra i requisiti (le attese) e quanto realizzato. Ciò dipende quasi sempre da un'analisi carente e dall'incapacità di calarsi pienamente nel dominio di conoscenza.

---

[10] M. T. Cabré, *La terminologia tra lessicologia e documentazione: aspetti storici e importanza sociale*, http://www.assiterm91.it/wp-content/uploads/2010/10/cabr%C3%A9-2000-testo-italiano. pdf.

Oggi l'*ontologia formale fondazionale* e le tecnologie dell'informazione lavorano insieme per creare applicazioni che siano in grado di individuare, reperire, organizzare e descrivere le informazioni, in base al contenuto o a un dominio disciplinare più o meno specifico. Il vocabolario di base dell'ontologia fondazionale è in grado di descrivere concetti generali da vari punti di vista, creando uno schema generale strutturato secondo categorie e relazioni di base tra entità che siano sufficientemente intuitive, cioè che derivino dal linguaggio naturale e dal senso comune[11]. Esso risulta un linguaggio basato su uno schema generale adattabile a diversi contesti, il cui scopo è rispecchiare le strutture del linguaggio e della cognizione dell'essere umano e favorire la cooperazione e l'interazione tra diverse tipologie di agenti (umani e non).

Le ontologie fondazionali si basano su strutture categoriali impiegabili e riutilizzabili in tutti i contesti; negli ultimi anni sono state applicate per sviluppare ontologie di ambiti specialistico-settoriali specifici, come, per esempio, la struttura categoriale della *Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE).*

DOLCE è un'ontologia descrittiva, cioè parte dall'assunto che l'uso quotidiano e il senso comune abbiano una rilevanza ontologica: la struttura del mondo viene data in base agli agenti o fenomeni che la concettualizzano. Le categorie dell'ontologia DOLCE dipendono pertanto dalla percezione umana, da tracce culturali e da convenzioni sociali così come le relazioni di base sono intuitive ed applicabili a più domini[12].

Viene pertanto naturale aggiungere che pensando a strumenti generalisti come OPAC e IR, *Institutional repository*, vi è la difficoltà di creare ontologie che travalicano i singoli domini di conoscenza. L'ambizione del web

[11] S. Gaio *et al.*, *Un'introduzione all'ontologia DOLCE*, «AIDAinformazioni», 28 (gen.-giu. 2010), n. 1-2, pp. 107-125.

[12] Le categorie principali sono: continuanti (entità che esistono nel tempo, per es., gli oggetti) e occorrenti (entità che accadono nel tempo, per es. gli eventi); qualità (che possono variare nel tempo ma che sono una caratteristica legata all'esistenza dell'entità); entità astratte (non hanno qualità temporali né spaziali). Le relazioni di base di DOLCE sono: relazione di parte (caratteristica dei continuanti) e di parte temporale (occorrenti e entità astratte sono atemporali); dipendenza e dipendenza spaziale; costituzione; partecipazione (un oggetto partecipa ad un evento senza essere parte di questo); inerenza qualitativa e valore qualitativo; tipi di continuanti; tipi di occorrenti; tipi di qualità (cfr. GAIO, cit., 2010).

semantico mi sembra sia quella di codificare le relazioni direttamente insieme ai dati (*linked data*) in modo da consentire la navigazione, scoperta e reperimento su base semantica indipendentemente da dove sono contenute le risorse.

5. Il modello concettuale del *Nuovo Soggettario*

In quest'ottica di neutralità da un preciso contesto e di applicabilità a più contesti di categorie ed entità a queste legate, potrebbe essere impiegato il modello concettuale del *Nuovo Soggettario* della Biblioteca Nazionale Centrale di Firenze (BNCF): la sua struttura categoriale e relazionale potrebbe essere sviluppata e maggiormente articolata da un punto di vista ontologico. Ciò dovrebbe essere reso più facile dalla conversione del thesaurus in SKOS, Simple Knowledge Organization System; si basa, cioè, sul linguaggio RDF, che è anche alla base di OWL, OWL Web Ontology Language.

L'RDF, *Resource Description Framework*, è un modello definito dal W3C per descrivere le *risorse* sul web. Il modello di descrizione si basa su tre tipi di oggetto:

- *Risorsa*: una qualsiasi entità identificabile con un *Universal Resource Identifier* (URI), un riferimento d'identificazione come l'indirizzo web.
- *Proprietà*: una specifica caratteristica o attributo di una risorsa.
- *Asserzione*: definita dall'insieme di una *risorsa*, una *proprietà* e uno *specifico valore* per quella proprietà; costituisce l'elemento descrittivo, le caratteristiche di una risorsa e le relazioni con altre risorse.

L'RDF è un linguaggio basato su XML, utilizzato per descrivere oggetti digitali *identificabili* nel web. Una struttura RDF diventa così disponibile sul web, accessibile e comprensibile (soprattutto alle macchine), che può essere utilizzata per finalità e contesti differenti. La traduzione del *Nuovo Soggettario* in RDF/SKOS è stata voluta e realizzata proprio per questa sua finalità di apertura alla comunità del web. Il *Nuovo Soggettario* non è un'ontologia complessa, ma sostanzialmente un thesaurus, che grazie alla sua ricchezza e attualità di terminologia e relazioni, contribuisce a diminuire quel *gap semantico* tra la parola come *segno* e parola come *significato*.

Il progetto del *Nuovo soggettario* è pertanto nato con lo scopo di creare un linguaggio d'indicizzazione codificato, basato sul modello analitico-sintetico[13], conforme agli standard internazionali, e, al contempo, flessibile e modulare, ovvero adattabile a rappresentare soggetti di diverso livello di complessità, in modalità precoordinata e postcoordinata (dove i singoli termini, utilizzati come parole chiave, possono essere combinati con gli operatori booleani)[14] per descrivere qualsiasi tipologia di risorsa documentaria, in domini generali e specialistici, e fruibile da contesti bibliotecari e della documentazione.

Il *Nuovo Soggettario*, in quanto linguaggio bibliografico[15], si configura come un sistema integrato di quattro componenti progettate per operare insieme, per poter supportare il processo d'indicizzazione:

1. Vocabolario
2. Semantica
3. Sintassi
4. Pragmatica

L'architettura relazionale del thesaurus è basata su quattro categorie generali

[13] Il linguaggio sintetico d'indicizzazione si basa sulla combinazione di termini, derivati da un vocabolario controllato e strutturato, in base a precise norme sintattiche. Il modello analitico sintetico investe quindi due piani: semantico, basato su relazioni a priori che non hanno corrispondenza con i documenti analizzati; sintattico, basato su relazioni create a posteriori, sulla base dell'analisi di una documentazione precisa. Cfr. P. Buizza, *Subject analysis and indexing: an 'Italian version' of the analytico-synthetic model*, relazione tenuta all'IFLA Satellite Pre-Conference of the Classification and Indexing Section "Looking at the Past and Preparing for the Future", Firenze, 20-21 agosto 2009, http://www.ifla2009satelliteflorence.it/meeting2/program/assets/Buizza.pdf.

[14] L. Franci *et al.*, *The Nuovo Soggettario thesaurus: structural features and Web applications project*, relazione tenuta a IFLA Satellite Pre-Conference of the Classification and Indexing Section 'Looking at the Past and Preparing for the Future', Firenze, 20-21 agosto 2009, http:// www.ifla2009satelliteflorence.it/meeting2/program/assets/FranciLucarelliMottaRolle.pdf, pp. 3.

[15] Tutti i linguaggi bibliografici sono linguaggi artificiali che assumono il ruolo di 'ponte' tra il linguaggio naturale, usato dal lettore in fase di ricerca, ed il linguaggio documentario, che contiene informazioni. I linguaggi bibliografici descrivono unità informative e consistono di vocabolario (espressioni semplici o complesse che identificano entità, attributi e relazioni), semantica (identifica le relazioni che ci possono essere tra i nomi o tra nomi e soggetti), sintassi (ordinamento degli elementi che costituiscono il linguaggio) e pragmatica (apparato normativo che guida nell'applicazione del linguaggio). Cfr. E. Svenonius, *Il fondamento intellettuale dell'organizzazione dell'informazione*, Le Lettere, Firenze 2008, pp. 75-83.

(azioni, agenti, cose e tempo) mutualmente esclusive, che raccolgono ed esauriscono al loro interno tutte le caratteristiche delle tredici sottocategorie sotto cui è elencata la terminologia.

Il principio di concettualizzazione e di categorizzazione della terminologia si basa sulla definizione attribuita da enciclopedie e dizionari: il valore semantico di ciascun lemma viene decontestualizzato, anche se il riferimento al contesto d'impiego è sempre 'a monte e a valle' del processo d'indicizzazione.

Ad aprile 2011 il thesaurus contiene circa trentottomila termini e viene quotidianamente alimentato in base alle proposte della Bibliografia nazionale italiana (BNI) o alle proposte inoltrate dagli enti scientifici e di ricerca che collaborano al progetto.

Il thesaurus ha raggiunto una consistenza pari alla metà dei termini che caratterizzano una lingua (circa settantamila); la sua terminologia è sufficientemente rappresentativa dei termini *types* che caratterizzano i diversi ambiti disciplinari. Il Settore Ricerche e strumenti d'indicizzazione della Biblioteca Nazionale Centrale di Firenze sta impegnandosi ad alimentare questo strumento e a sperimentarne l'impiego su diversi fronti. In particolare:

1. Mappatura della terminologia, con altri strumenti di indicizzazione e con strumenti semantico-lessicografici: nel campo 'fonte' è stato creato un *deep link* con le LCSH, AGROVOC, Treccani.it.
2. Colloquio semantico e tecnico tra il *Nuovo Soggettario* e thesauri specialistico-settoriali.
3. Impiego del thesaurus per l'indicizzazione automatica delle risorse digitali.
4. Ha stabilito contatti con Tullio De Mauro per poter collaborare con il gruppo di Senso comune[16].

[16] «Senso Comune è un'associazione che ha sede presso il CNR ISTC. La sua attività consiste nello sviluppo, manutenzione e distribuzione di una base di conoscenza della lingua italiana attraverso la collezione e la formalizzazione delle competenze linguistiche della comunità dei parlanti condotta mediante strumenti informatici cooperativi. Si propone di attuare iniziative di ricerca relative a: semantica del linguaggio naturale, ontologia formale e legame tra ontologia e lessico, formalismi di rappresentazione della conoscenza linguistica, costruzione di risorse per la codifica delle conoscenze linguistiche dei parlanti, elicitazione della conoscenza attraverso i mezzi informatici, metodi per il ragionamento automatico su basi di conoscenza, riconoscimento automatico del significato» (http://www.sensocomune.it/portale/index.php?option=com_content&task=view&id=24&Itemid=115).

6. La terminologia di una singola disciplina

È in corso una tesi di dottorato di Elisabetta Viti, che ha come scopo principale:

a. L'elaborazione di criteri e modalità per ricostruire la terminologia di una singola disciplina all'interno di un contesto generale multidisciplinare, in particolare l'Economia finanziaria.
b. L'interoperabilità semantica e tecnica tra due realtà thesaurali differenti, *Nuovo Soggettario*-thesaurus della LIUC di Castellanza (Varese).

La sperimentazione si basa sulla definizione di un corpus terminologico rappresentativo il settore dell'Economia finanziaria, sulla cui base mappare la terminologia che appartiene ai due differenti contesti thesaurali. L'analisi della terminologia ha rilevato l''essenza' dei termini e quindi anche la differenza di categorizzazione e di relazioni semantiche che si possono stabilire in relazione a domini disciplinari differenti. Da un primo confronto sono emerse le seguenti differenze semantiche e strutturali:

- *Differenza nella copertura*: differente stato di sviluppo della struttura thesaurale, differente scopo o differenze nelle necessità degli utenti. In particolare, alcuni temi sono stati sviluppati più di altri o con maggiore o minore specificità.
- *Uso di parole differenti*: differenza nella scelta del livello linguistico; in un thesaurus di carattere generale il tasso di polisemia di un termine sarà maggiore rispetto a un linguaggio specialistico dove c'è la tendenza a un rapporto biunivoco tra termine e concetto; all'opposto, in un linguaggio specialistico l'uso della sinonimia sarà ridotto.
- *Differenze semantiche*: differenti interpretazioni semantiche dipendenti dal contesto.
- *Differenze strutturali*: differenti aspetti di categorizzazione o differenze nell'applicazione della monogerarchia o poligerarchia; oppure, se il thesaurus è poligerarchico, dai criteri di applicazione della poligerachia.

7. L'indicizzazione semi-automatica di tesi di dottorato in BNCF

Il progetto di indicizzazione semi-automatica di tesi di dottorato in formato digitale, tramite l'attribuzione di *keyword* e *keyphrase* prodotte dalla

comparazione dei termini estratti dal testo con i termini definiti nel *Nuovo Soggettario* in formato SKOS come concetti, non ha l'ambizione di sostituirsi alla fase di catalogazione semantica elaborata da esperti, non prevista per questa tipologia di materiale. Intende, invece, colmare un vuoto informativo, penalizzante e grave per i ricercatori, dovuto dalla mancanza di un'approfondita analisi semantica.

Il trattamento delle risorse selezionate dal personale della Biblioteca nazionale centrale di Firenze è rappresentabile in pochi passi:

- *L'indicizzazione dei documenti*
  I formati dei documenti soggetti a tale processo possono essere diversi (pdf, word, excel, access, jpg, etc.) per garantire che nessun documento sia escluso per ragioni tecniche. I documenti selezionati sono elaborati con specifiche funzioni di indicizzazione, in una fase che può essere considerata *propedeutica* delle successive. Si tratta di funzioni di normalizzazione che trasformano in un formato testuale intermedio documenti che in origine hanno formati diversi.
- *L'estrazione dei metadati dai documenti*
  La *prima* fase del processo consiste nell'estrarre i metadati (eventualmente) presenti all'interno dei documenti. In relazione alla qualità del documento digitale originale potranno essere estratti metadati quali Titolo, Autore, Anno di creazione.
- *L'estrazione del testo dai documenti*
  Una *seconda* fase del processo prevede l'estrazione del testo dei documenti. Insieme ai metadati estratti nella fase precedente, il contenuto testuale della risorsa rappresenta un'informazione fondamentale per la comprensione del *dominio*, del *contesto di analisi* e per la successiva estrazione delle *parole-chiave* e *frasi-chiave*. Questa fase di estrazione del testo dalla documentazione selezionata prevede un'attività automatica di esclusione di termini non significativi (sulla base di *stopwords* per specifiche lingue) e caratteri non utili ai fini delle ricerche.
- *L'estrazione di parole e frasi chiave*
  La *terza* fase del processo prevede l'estrazione dalla risorsa di *termini* o *frasi* ritenute rilevanti sulla base di logiche e algoritmiche di frequenza, di vicinanza ecc. (algoritmi statistici di rilevazione del dato). Le *keyword* e le *keyphrase* ritenute significative all'interno della risorsa saranno asse-

gnate, in fase provvisoria, al documento come ulteriore set di metadati descrittivi.

Il processo di trattamento di dati sopra descritto è frutto di automatismi complessi, basati su criteri statistici di estrazione dei termini o delle frasi ritenute significative. Un intervento umano di selezione dei termini, ove raramente possibile, renderebbe certamente il processo di indicizzazione più controllato e puntuale.

I costi di realizzazione di questi interventi di indicizzazione e soggettazione della documentazione prodotta sono altissimi e non accessibili a pubbliche amministrazioni e, ove possibili, limitati alla documentazione standard, oggetto di operazioni catalografiche (monografie, articoli ecc.). L'utilizzo di specifici strumenti per l'estrazione di parole-chiave controllate, di seguito descritto, è centrato sull'obiettivo della creazione di un set di metadati semantici associati al documento, anche in assenza di catalogazione descrittiva, in modalità automatica o semi-automatica.

Per produrre una lista di *keyword* e *keyphrase* davvero significative per il documento, e orientare meglio la ricerca degli utilizzatori del sistema informativo, il processo di indicizzazione delle risorse è completato dall'uso del *Nuovo Soggettario* in RDF come base di conoscenza, a garanzia di una più ampia 'pertinenza' delle parole chiave assegnate automaticamente al documento. In questa fase del lavoro quanto estratto precedentemente dal documento (testo, metadati) viene automaticamente 'confrontato' con i concetti, i suoi attributi, le sue relazioni espresse, in termini leggibili dalla macchina, nel *Nuovo Soggettario*: questo processo garantisce una più puntuale e scientificamente pertinente assegnazione di parole e frasi chiave alla risorsa. I termini estratti dal documento, *confrontati con i termini presenti sul thesaurus utilizzato*, sono:

- estesi: tramite l'uso di *Related term*, di *Broader term* o *Narrower term* correlati al termine in oggetto;
- sostituiti: sostituzione di voci accettate a voci non accettate;
- filtrati: esclusione di termini non pertinenti rispetto al 'contesto' individuato.

Il risultato di questa delicata e potente operazione di elaborazione del testo è un *elenco di keyword o keyphrase* che, nel caso del progetto specifico, sono *automaticamente proposte* a un gruppo di selezionatori, che conferma-

no o eliminano le associazioni proposte per la specifica risorsa.

Si tratta di un progetto che nasce dalla collaborazione tra BNCF, Casalini Libri e la società @Cult. Il prototipo dovrebbe basarsi sull'indicizzazione automatica delle tesi di dottorato, depositate per diritto di stampa presso la BNCF. Ad aprile 2011 sono in fase di definizione le modalità di estrazione dei dati e di confronto con la base di conoscenza designata.

Conclusioni

In questa breve panoramica delle tendenze e dei progetti che stanno realizzandosi nell'impiego di una terminologia controllata e strutturata che sia alla base di un processo d'indicizzazione e usata come strumento efficace di *information retrieval*, nonostante le tecnologie siano cambiate, sembrano ancora validi i principi della soggettazione che Cutter enunciava nelle *Rules* del 1876, e che Camassima ha ribadito nel 1966 in *Soggettazione*[17]: precisione, univocità, uniformità, modernità (nel rispetto della correttezza linguistica) sono alla base di tutti i linguaggi d'indicizzazione:

1. la voce a soggetto deve essere diretta e specifica;
2. la voce a soggetto deve essere formulata secondo l'uso linguistico attuale, corrente.

[17] E. Casamassima, *La soggettazione*, in *La documentazione in azienda*, vol. II (1966), pp. 235-256.

TERMINOLOGIE E ONTOLOGIE
Definizioni e comunicazione fra norma e uso

EDUCatt
Ente per il Diritto allo Studio
Universitario dell'Università Cattolica
Largo Gemelli 1, 20123 Milano -
tel. 02.7234.22.35 - fax 02.80.53.215
e-mail: editoriale.dsu@educatt.it (produzione);
librario.dsu@educatt.it (distribuzione)
web: www.educatt.it/libri
ISBN: 978-88-6780-013-1

euro 6,50