Tassonomie e thesauri

ANTONIETTA FOLINO*

1. Introduzione

La gestione delle informazioni e della conoscenza relative ad uno o più domini e gli scambi comunicativi tra gli attori che operano al loro interno non possono prescindere da un utilizzo condiviso della terminologia[1], rendendo necessaria la definizione di strumenti e risorse che ne consentano un'organizzazione il più possibile coerente e non ambigua.

Si tratta di lessici, glossari, tassonomie, soggettari, sistemi di classificazione, thesauri, mappe concettuali, ontologie, ecc., che, dal punto di vista strutturale, si differenziano gli uni dagli altri sostanzialmente per la presenza o meno di relazioni di tipo semantico tra i concetti in essi rappresentati e per il diverso grado di formalismo che caratterizza la modellizzazione dell'informazione, come mostrato in Figura 1.

* Università della Calabria, Dipartimento di Lingue e Scienze dell'Educazione.

1 *«The word terminology refers to at least three different concepts: a. The principles and conceptual bases that govern the study of terms; b. The guidelines used in terminographic work; c. The set of terms of a particular special subject»*.
MARIA TERESA CABRÉ, *Terminology: theory, methods and applications*, Sager J.C. (ed.), DeCesaris J.A. (traduzione di), Philadelphia PA, John Benjamins, 1998, p. 33.

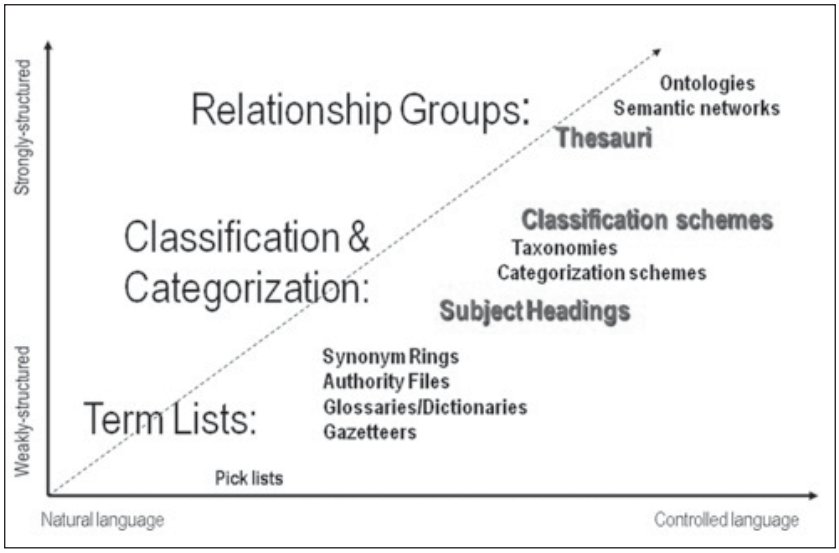

Figura 1. Controllo terminologico e strutturazione[2].

Si va infatti dai lessici specialistici, che consistono in liste non strutturate di termini afferenti ad un settore della conoscenza[3], ai glossari, che aggiungono a ciascuna voce la relativa definizione[4],

2 MARCIA LEI ZENG, ATHENA SALABA, *Toward an International Sharing and Use of Subject Authority Data*, FRBR Workshop, OCLC, 2005.

3 Un lessico specialistico può essere definito come un insieme di termini utilizzati in modo consensuale e convenzionale dagli individui che operano in uno stesso ambito al fine di scambiare e divulgare informazioni e conoscenze in maniera precisa, univoca e concisa.
Cfr. HELLMUT RIEDIGER, *Cos'è la terminologia e come si fa un glossario*, 2012.
<http://www.term-minator.it/corso/doc/mod3_termino_glossa.pdf>.

4 Se costruiti in forma di database terminologici, all'interno dei quali ciascun termine è descritto da un'apposita scheda, i glossari prevedono anche l'inserimento di relazioni tra concetti (sinonimia, iperonimia-iponimia, ecc.) sulla base di un albero concettuale che ne definisce la struttura.

ai sistemi di classificazione[5], ai soggettari[6], ai thesauri, che, pur se con modalità diverse, integrano la terminologia con relazioni semantiche tra i concetti, fino alle mappe concettuali e alle ontologie, che aggiungono un alto livello di formalismo attraverso la definizione di restrizioni nella partecipazione dei concetti alle relazioni, la rappresentazione in linguaggi interpretabili dalle macchine, l'esplicitazione della natura delle relazioni, ecc.

Il principio sottostante a tali risorse, seppur ottenibile a diversi livelli e con differenti modalità, è il controllo terminologico: a ciascun termine è attribuibile il solo significato valido per il dominio di interesse e l'interpretazione non deve essere soggetta ad ambiguità e incomprensione. Tali significati, insieme all'uso di ciascuna voce, sono condivisi e compresi dalla comunità di utenti che li impiega nelle situazioni comunicative in cui è coinvolta. Nella misura in cui rappresentano la conoscenza di dominio, invece, gran parte di queste risorse rientra in quelli che vengono comunemente definiti *Knowledge Organization Systems* (KOS)[7].

---

5 Rispetto ai thesauri, i sistemi di classificazione hanno spesso la pretesa di rappresentare l'intero scibile, mentre dal punto di vista strutturale, non presentano la relazione associativa e prevedono un sistema di notazione obbligatorio.

6 Rispetto ad un thesaurus un soggettario fornisce le regole sintattiche per la costruzione della stringa di soggetto, ovvero di «*una sequenza ordinata di termini, che rappresenta il soggetto di un documento*» (Cfr. ALBERTO CHETI, *Manuale ipertestuale di analisi concettuale*, 1996, <http://biblioteche.unibo.it/manuals/html_1/HOME.HTML>), da assegnare ai documenti, soprattutto in una logica di pre-coordinazione, nella quale la combinazione dei concetti rappresentativi del contenuto di un documento avviene già al momento dell'indicizzazione. Per tali motivi, un soggettario viene espressamente costruito per esigenze di indicizzazione, per cui anche la sua struttura viene definita in relazione all'insieme di documenti da indicizzare.

7 *«The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge*

Ci soffermeremo essenzialmente sulla descrizione delle caratteristiche e delle funzionalità di tassonomie e thesauri, in ragione della loro riscoperta e rivalutazione nell'ambito della ricerca sul Web, dopo una fase di declino che li ha considerati quasi obsoleti di fronte alle emergenti tecnologie del Web Semantico.

Se, infatti, in quanto strumenti cardine delle scienze documentali e bibliotecarie, assolvevano a funzioni di ordinamento, organizzazione e recupero dell'informazione in ambienti prettamente cartacei, oggi la loro definizione risulta indispensabile nel mondo del web e dei documenti digitali, con la conseguente acquisizione di nuove connotazioni e funzionalità e con una rivalutazione e valorizzazione delle loro potenzialità.

Ci occuperemo in un primo momento delle tassonomie fornendo una presentazione delle caratteristiche che le contraddistinguono per poi passare alla trattazione più approfondita dei thesauri, presentandone il contesto normativo, le funzionalità, le modalità di realizzazione e gli impieghi. Sebbene infatti le norme più recenti in materia di vocabolari controllati abbiano ampliato i propri interessi non limitandosi ai soli thesauri, le indicazioni più precise e complete continuano ad essere fornite solo per questi ultimi, per cui anche in questo caso, l'attenzione sarà focalizzata maggiormente sui thesauri facendo riferimento alle tassonomie laddove si ritiene interessante effettuare dei confronti.

---

management. […]. Knowledge organization systems also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies».
GAIL HODGE, *Systems of Knowledge Organization for Digital libraries. Beyond traditional authority files*, 2000.
<http://www.clir.org/pubs/reports/pub91/contents.html>.

2. Tassonomie

In senso tradizionale, l'idea di tassonomia è legata alle discipline scientifiche e alla classificazione dicotomica degli organismi nelle scienze biologiche[8]. Nell'ambito della documentazione, invece, il termine acquista un'accezione più ampia, riferendosi all'organizzazione sistematica di un soggetto o dominio. Secondo la definizione fornita dallo standard americano ANSI/NISO Z39-19:2005[9], infatti, per tassonomia si intende: *«a controlled vocabulary consisting of preferred terms, all of which are connected in a hierarchy or polyhierarchy»*. Le tassonomie sono impiegate in modo massiccio nell'organizzazione dell'informazione in ambienti digitali. Spesso, infatti, vengono citati repertori quali Yahoo! o Open directory come casi esemplificativi del loro impiego (Gnoli et alii, 2006)[10]. Tuttavia, proprio a causa di questa ampia diffusione, spesso si assiste ad un utilizzo indiscriminato del termine tassonomia, indicando con esso strutture di concetti che non rispettano le caratteristiche distintive di tale strumento e il rigore con il quale dovrebbe essere costruito per garantire il controllo terminologico e l'organizzazione della conoscenza. In taluni casi, ad esempio, per tassonomia si intende l'insieme delle categorie (o faccette) che costituiscono il menù di navigazione di un sito web, mentre il significato tradizionalmente accettato nelle scienze del libro è quello di sistema di classificazione[11].

---

8 Si pensi ai botanici Konrad Gesner e Karl von Linné.

9 ANSI/NISO Z39-19:2005, *Guidelines for the construction, format, and management of monolingual controlled vocabularies*, p. 18.

10 Claudio Gnoli, Vittorio Marino, Luca Rosati, *Organizzare la conoscenza: dalle biblioteche all'architettura dell'informazione per il web*, Milano, Tecniche Nuove, 2006, p. 44.

11 Cfr. Gail Rayburn, *Taxonomies and Thesauri*, 2011. <http://www.llrx.com/system/files?file=taxonomiesthesauri.pdf>.

La sola relazione semantica che si inserisce tra i concetti di una tassonomia è, quindi, quella gerarchica, che ne determina la tipica organizzazione ad albero che rende visibili i rapporti tra sovra- e sotto-ordinati. Si predilige la monogerarchia, quindi una collocazione unica per ciascun concetto. Sebbene i termini tassonomia e thesaurus siano spesso utilizzati indifferentemente come se fossero sinonimi, tra le due tipologie di vocabolari controllati esistono differenze significative: rispetto ad un thesaurus, infatti, una tassonomia non prevede la relazione di equivalenza, essendo costituita, come da definizione sopra riportata, da soli termini preferiti (non vi si ritrovano dunque sinonimi, quasi sinonimi, ecc. che costituirebbero il vocabolario d'accesso[12]), né la relazione associativa per l'esplicitazione di rapporti semantici diversi da quelli gerarchici. Inoltre, le relazioni non sono esplicitate per mezzo di sigle standard, come avviene invece per i thesauri. Per quanto riguarda la struttura, è possibile prevedere, così come per i thesauri, l'inserimento di faccette (vedi par. 3.4) che consentano di separare in maniera sistematica le gerarchie soprattutto ai fini della navigazione sul Web. Non prevedono, inoltre, un sistema di notazione, spesso presente nei thesauri, e i criteri che guidano la definizione delle relazioni sono meno rigorosi rispetto a quelli che le regolano in questi ultimi. Le principali differenze sono riassunte nella Figura 2, estratta dalla citata ANSI/NISO Z39-19:2005.

Per quanto riguarda le funzionalità alle quali assolvono, le tassonomie condividono sostanzialmente quelle proprie dei thesauri, quindi indicizzazione, recupero di informazione, organizzazione della conoscenza, descritte dettagliatamente nel seguito

[12] *«Costituito sia dai termini preferiti che dai termini non preferiti, cioè dai termini che non possono essere utilizzati per l'indicizzazione e che rimandano a termini preferiti»*.
SERAFINA SPINELLI, *Introduzione all'indicizzazione*, 2006. <http://biocfarm.unibo.it/~spinelli/indicizzazione/>.

del capitolo. Tuttavia, in ragione delle differenze pocanzi illustrate, le potenzialità di un thesaurus nella maggior parte dei contesti sono indubbiamente maggiori potendo sfruttare relazioni altre rispetto a quella gerarchica[13]. La navigazione in ambienti digitali resta la funzione principale che le tassonomie adempiono.

Property	List	Synonym Ring	Taxonomy	Thesaurus
Types of Terms				
Preferred terms	Yes	No	Yes	Yes
Entry terms	No	Yes	No	Yes
Candidate terms	No	No	No	Optional
Provisional terms	No	No	No	Optional
Deleted terms	No	No	No	Optional
Relationships	No	Yes	Yes	Yes
Equivalence		Yes	No	Yes
Hierarchy		No	Yes	Yes
Part/Whole		No	Yes	Yes
IsA		No	Yes	Yes
HasA		No	Yes	Yes
Classification		No	Optional	Optional
Related terms		No	No	Yes
Facet		No	No	Optional
Notes	No	No	Optional	Optional
Scope note			No	Optional
History note			No	Optional
Other notes			No	Optional

Figura 2. Tassonomie e Thesauri[14].

[13] Basti pensare alla funzione di mediazione tra indicizzatore e utente che il thesaurus svolge grazie alla relazione di equivalenza o all'estensione dei risultati delle ricerche per mezzo di quella associative.

[14] ANSI/NISO, *op. cit.*, p. 135.

3. Thesauri

3.1 *Definizione e struttura di un thesaurus*

La definizione di thesaurus fornita dalla recente norma ISO 25964-1:2011 che ne regola i principi di costruzione e gestione è *«controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms»*[15]. Per vocabolario controllato si intende un insieme di termini il cui significato è chiaro e non ambiguo nello specifico contesto o dominio per il quale il thesaurus è costruito. Tale controllo si esercita principalmente attraverso la strutturazione dei concetti sulla base di relazioni semantiche rese esplicite. Nel caso specifico, si tratta di tre grandi tipologie standardizzate di relazioni: di equivalenza, gerarchiche e associative.

La relazione di equivalenza consente la gestione della sinonimia e della quasi-sinonimia e delle varianti linguistiche. I termini che ai fini del thesaurus rappresentano il medesimo concetto fanno parte del cosiddetto gruppo di equivalenza[16]. Ad uno di questi termini viene attribuito lo status di preferito, mentre i restanti, in qualità di non preferiti, gli sono legati per mezzo di rinvii. Tale relazione viene esplicitata per mezzo delle sigle USE, che rimanda dal/dai termini non preferiti a quello preferito, e UF (*Used for*), che viene definita nel senso opposto. Nell'ambito della stessa relazione di equivalenza è possibile distinguere tra sinonimia assoluta e sinonimia relativa: nel primo caso essa esiste indipendentemente dall'area semantica, dalla specificità del

---

15 ISO 25964-1:2011, Information and documentation – *Thesauri and interoperability with other vocabularies*, Part 1: *Thesauri for information retrieval*, p. 12.

16 SERAFINA SPINELLI, *Introduzione ai thesauri*, 2005. <http://biocfarm.unibo.it/~spinelli/indicizzazione/thesauri.htm>.

thesaurus e dalla scelta del termine preferito e interessa varianti ortografiche[17], sigle e acronimi e relative forme sciolte[18], preferenze linguistiche[19], termini specialistici e termini utilizzati nel linguaggio comune[20], nomi comuni e nomi commerciali[21], prestiti e relative traduzioni[22], ecc.; in presenza di quasi sinonimi, invece, si parla si sinonimia relativa, in quanto la relazione è valida solo ai fini e nel contesto per il quale il thesaurus viene realizzato. La possibilità di inserire tale relazione dipende dalla copertura semantica del thesaurus e dal grado di discriminazione necessario in fase di indicizzazione e di ricerca. La si ritrova, ad esempio, in thesauri non specialistici o per tematiche che si discostano dal focus principale del vocabolario controllato[23]. Rientrano in questa tipologie di sinonimia anche gli antonimi[24], che, sempre in base alle esigenze di specificità del thesaurus, possono essere ritenuti equivalenti sebbene esprimano significato opposto, poiché il tema di un documento potrebbe essere espresso per mezzo di entrambe le forme linguistiche, e termini per i quali la relazione di sinonimia è valida solo in alcuni contesti d'u-

17 Es. Database – Data-base – Data Base.

18 Per quanto concerne la scelta del termine preferito, la normativa prevede che l'acronimo venga privilegiato qualora sia più utilizzato e conosciuto della relativa forma sciolta (es. UNICEF). Se, invece, il suo utilizzo causerebbe ambiguità, è preferibile stabilire un rinvio verso il suo scioglimento (es. CC – Corrente Continua, Conto Corrente, ecc.).

19 Es. Flusso Termico - Flusso di calore.

20 Es. Cefalea – Mal di testa.

21 Es. Penna – Bic.

22 Es. Computer – Elaboratore elettronico.

23 Una relazione di quasi sinonimia potrebbe essere definita tra Legge, Decreto Legge, Decreto Legislativo in un thesaurus non relativo a discipline giuridiche. Nel caso di un thesaurus specialistico in questo dominio, ovviamente, tra i tre concetti non potrebbe sussistere alcuna relazione di equivalenza.

24 Es. Tolleranza – Intolleranza.

so[25]. Ulteriori casi in cui è possibile ricorrere alla relazione di equivalenza consistono nel rinvio da termini più specifici ad un termine sovraordinato[26] e nella *compound equivalence*[27], nella quale, in presenza di un concetto complesso, è preferibile utilizzare i concetti semplici che risultano dalla sua scomposizione. Anche in questi casi, la scelta dipende dal contenuto dei documenti da indicizzare e, di conseguenza, dai termini potenzialmente impiegabili il loro recupero.

L'attribuzione dello status di termine preferito dipende da un insieme di fattori, tra i quali principalmente le finalità del thesaurus, soprattutto nel caso della quasi sinonimia, e gli utenti ai quali il thesaurus è rivolto e che lo utilizzeranno per finalità di recupero dell'informazione o come strumento di organizzazione della conoscenza[28]. Così come per la scelta dei termini che devono essere inseriti in un thesaurus (si veda il par. 3.5), anche in questo caso, l'autorevolezza della fonte dalla quale un termine è stato estratto, il parere di esperti di dominio e la frequenza d'uso rappresentano un supporto nella scelta del termine preferito. La relazione di equivalenza si rivela di fondamentale importanza per indirizzare nel corretto utilizzo della terminologia nei domini di realizzazione dei thesauri, limitando l'ambiguità e l'incoerenza negli scambi comunicativi tra gli attori che vi operano.

La relazione gerarchica esprime un rapporto di subordinazione/sovraordinazione tra concetti che rappresentano una classe o

---

25 Es. Faccia – Viso.

26 Es. Rocce magmatiche, Rocce sedimentarie, Rocce metamorfiche USE Rocce.

27 Es. Glicemia USE Sangue + Glucosio.

28 Un thesaurus costruito nel dominio della medicina e rivolto ai pazienti privilegerà in qualità di termini preferiti quelli quotidianamente impiegati nel linguaggio comune. La relazione di equivalenza con i corrispondenti termini specialistici, ad esempio nel contesto di un sistema di recupero di informazione, permetterà l'incontro con gli specialisti del settore.

un insieme e concetti che rappresentano elementi, parti o individui[29]. Si distingue, infatti, tra relazione gerarchica di tipo genere-specie, esplicitata tramite le sigle BTG (*Broader Term Generic*) – NTG (*Narrower Term Generic*)[30], di tipo parte-tutto, le cui sigle sono BTP (*Broader Term Partitive*) e NTP (*Narrower Term Partitive*)[31] ed esemplificativa, BTI (*Broader Term Instantial*) e NTI (*Narrower Term Instantial*)[32]. La relazione genere-specie può essere definita solo tra concetti che appartengono alla medesima categoria (oggetti, materiali, attività, proprietà, discipline, ecc.) e che rispettano l'*all-and-some test*[33], ovvero: solo alcuni membri della classe che indica il genere rientrano in quella che indica la specie, ma tutti i membri della classe che indica la specie devono rientrare in quella che indica il genere. I concetti che possono essere interessati dalla relazione parte-tutto rientrano in categorie ben identificate, quali sistemi e organi del corpo, luoghi geografici, discipline e campi di studio, strutture sociali gerarchizzate[34]. I concetti che indicano le parti, infatti, devono appartenere in maniera esclusiva al concetto che rappresenta il tutto e con il quale esiste una relazione di questo tipo. Nei casi in cui questa condizione non si verifica è preferibile ricorrere ad una relazione di tipo associativo.

La relazione associativa, la cui sigla è RT (*Related Term*), consente la gestione delle relazioni semantiche diverse da quella gerarchica che possono essere stabilite tra due concetti. La ISO 25964-1:2011 prevede la possibilità che la natura di tale re-

[29] La relazione gerarchica può essere definita su più livelli, generalmente contrassegnati da un numero progressivo, mentre nei casi in cui un concetto abbia più di un sovraordinato, si parla di poligerarchia.
[30] Es. Felini NTG Gatti – Gatti BTG Felini.
[31] Es. Apparato circolatorio NTP Cuore – Cuore BTP Apparato circolatorio.
[32] Es. Catene Montuose NTI Alpi – Alpi BTI Catene Montuose.
[33] ISO 25964-1:2011, *cit.*, p. 59.
[34] *Ivi*, p. 60.

lazione venga di volta in volta esplicitata (es. causa-effetto), agevolando la comprensione della struttura thesaurale e accorciando le distanze tra questo strumento ed un'ontologia. Tuttavia non si tratta ancora di relazioni standardizzate e le eventuali operazioni di mappatura tra thesauri esistenti potrebbero risultare compromesse dalla difficoltà di stabilire delle corrispondenze. Tale tipo di relazione interessa concetti che condividono lo stesso sovraordinato o, nel caso di un thesaurus a faccette, concetti appartenenti a raggruppamenti diversi[35] e legati tra loro da un qualsiasi legame semantico e quelli introdotti dal medesimo principio di suddivisione e collocati sullo stesso livello gerarchico.

L'esercizio del controllo terminologico avviene anche tramite l'inserimento di note d'ambito o *Scope Note* (SN), ovvero campi testuali nei quali è possibile delimitare il significato di un dato termine, fornire informazioni circa il suo impiego, eventuali usi particolari, ecc.[36], e di qualificatori, che in presenza di omografi permettono di specificare l'ambito al quale il significato di ciascuno si riferisce[37].

3.2 *Evoluzione concettuale e normativa*

La maggior parte dei thesauri esistenti in letteratura è stata costruita secondo la norma ISO 2788:1986[38], se monolingue, e se-

[35] Es. Attività - prodotto (Tessitura RT Tessuto); Agente – Attività (Docente – Insegnamento); Disciplina – Oggetto di studio (Anatomia – Corpo Umano); Oggetti – Proprietà (Metalli – Malleabilità); Attività – Strumento (Incisione – Bulino); ecc.

[36] Esse possono essere ulteriormente specificate ricorrendo alle sigle DEF e HN (*History Note*) qualora si voglia fornire il significato o informazioni sull'evoluzione temporale di un dato concetto.

[37] Es. Organo (strumento musicale); Organo (corpo umano).

[38] ISO 2788:1986, Documentation – *Guidelines for the establishment and development of monolingual thesauri.*

condo la norma ISO 5964:1985[39], se multilingue, che hanno rappresentato un riferimento internazionale per lungo tempo, ovvero fino alla pubblicazione della recente ISO 25964-1:2011, che le ha sostituite[40] e la cui emanazione era particolarmente attesa dalla comunità dei professionisti dell'informazione, dal momento che le norme esistenti non rispecchiavano l'evoluzione dei thesauri ed era necessario un adattamento alle nuove esigenze di gestione dell'informazione in ambienti prettamente digitali. In particolare ci si riferisce all'aumento considerevole e costante della quantità di informazioni e di documenti disponibili sul Web, alle conseguenti accresciute opportunità di recupero degli stessi, alla diversa natura delle risorse informative e all'avvento dei motori di ricerca e del metodo di ricerca full-text.

Proprio i limiti di tale metodo rendono evidente la necessità dei thesauri nei repository di documenti digitali: la presenza dei termini che compongono un'interrogazione (o query) nel testo di un documento non garantisce che gli stessi siano rappresentativi del suo contenuto concettuale. L'attribuzione di voci indice a partire da un thesaurus, invece, aumenta la probabilità che il risultato di una ricerca sia pertinente e risponda ai bisogni informativi degli utenti. Altri limiti dipendono dal fatto che la ricerca può avvenire a partire da più di un termine o essere formulata in una lingua diversa da quella in cui sono redatti i documenti.

In Figura 3 viene riportato un quadro riassuntivo delle norme che hanno interessato i thesauri, e più in generale i vocabolari controllati, negli ultimi decenni.

---

39 ISO 5964:1985, Documentation – *Guidelines for the establishment and development of multilingual thesauri.*

40 La sostituzione ha interessato anche le norme nazionali francesi definite dall'ente di normazione AFNOR (*Association française de Normalisation*), ovvero la NF Z 47-100-1981- *Règles d'établissement des thésaurus monolingues* e la NF Z 47-101-1990 - *Principes directeurs pour l'établissement des thésaurus multilingues.*

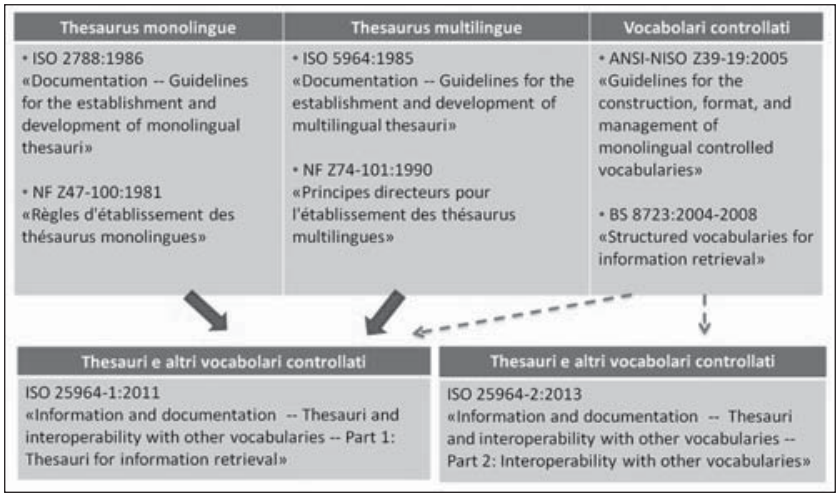

Figura 3. Thesaurus: contesto normativo

La direzione verso la quale si orienta la ISO 25964:2011, ovvero l'interoperabilità tra i vocabolari controllati, si coniuga con l'operato e gli obiettivi del *World Wide Web Consortium* (W3C)[41], concretizzatisi con la raccomandazione SKOS (*Simple Knowledge Organization Systems*)[42], finalizzata all'utilizzo dei sistemi di organizzazione della conoscenza nel Web Semantico, prevedendo sia la trasposizione di quelli esistenti sia la definizione di nuovi secondo questo formato. Il linguaggio si basa su RDF[43] e viene considerato come una tecnologia intermedia tra l'elevato formalismo logico dei linguaggi OWL[44] per la costru-

41 <http://www.w3.org/>.
42 <http://www.w3.org/2004/02/skos/>.
43 *Resource Description Framework*. <http://www.w3.org/RDF/>.
44 *Ontology Web Language*. <http://www.w3.org/TR/owl-features/>.

zione di ontologie e l'assenza o lo scarsa strutturazione delle applicazioni che attualmente caratterizzano il Web. Tra gli elementi o tag per mezzo dei quali è possibile descrivere i sistemi di organizzazione della conoscenza rientrano: *Concept, prefLabel, altLabel, broader, narrower, related, definition, scopeNote, exactMatch, collection, member, broaderTransitive, ecc.* alcune delle quali direttamente riconducibili alla struttura di un thesaurus.

Come evidenziato in figura, l'elaborazione della norma attualmente in vigore è avvenuta anche sulla base delle norme redatte in contesto inglese ed americano: la BS 8723:2004-2008[45] e la citata ANSI/NISO Z39-19:2005. Seppur non con valenza internazionale, infatti, questi documenti normativi avevano già introdotto ed anticipato alcune delle principali novità contenute nella ISO 25964[46], tra le quali l'interesse verso altre tipologie di vocabolari controllati, il formato elettronico dei thesauri, il loro

[45] BS 8723:2004-2008, *Structured vocabularies for information retrievial – Guide.*
È costituito da cinque sezioni edite separatamente.

[46] L'iter di sviluppo di una norma internazionale, infatti, viene in molti casi avviato in risposta ad esigenze e richieste provenienti da stakeholder esterni all'ente di normazione (nel caso specifico l'*International Organization for Standardization* – ISO). Tali esigenze vengono comunicate agli enti nazionali (es. UNI per l'Italia, BSI per il Regno Unito, ANSI per gli Stati Uniti, AFNOR per la Francia, ecc.), che, a loro volta, contattano l'ISO, del quale sono membri. L'iter di sviluppo degli standard è così riassunto sul sito dell'ISO (<http://www.iso.org/iso/home.html>): *«1. New standard is proposed to relevant technical committee. If proposal is accepted 2. Working group of experts start discussion to prepare a working draft. 3. 1st working draft shared with technical committee and with ISO CS. If consensus is reached within the TC 4. Draft shared with all ISO national members, who are asked to comment. If consensus is reached 5. Final draft sent to all ISO members. If standard is approved by member vote 6. ISO International Standard»*.
Cfr. anche ISABELLA FLORIO, *La normativa standardizzata per la gestione delle documentazione tra Italia e Francia*, Rubbettino Editore, 2011.

impiego per scopi di Information Retrieval (IR), l'interoperabilità tra più vocabolari attraverso la definizione di modelli[47] e formati (Calvitti, Viti, 2009)[48] (Casson, 2006)[49] (Groupe Langages documentaires de l'ADBS, 2007)[50], (Dextre Clarke, Lei Zeng, 2012)[51].

Al fine di comprendere che cosa si intende per thesaurus, quali sono le sue caratteristiche e come queste siano mutate nel tempo è opportuno riprendere le definizioni presenti nella normativa tecnica. Si riportano, quindi, sia quella prevista dalla ISO 2788:1986, nonostante la stessa non sia più in vigore, sia quella fornita dalla ISO 25964-1:2011, peraltro già riportata nel precedente paragrafo, a garanzia di completezza nella descrizione di tale strumento e a testimonianza dell'evoluzione della sua natura: *«Il thesaurus è un vocabolario di un linguaggio di indicizzazione controllato, organizzato formalmente in maniera da rende-*

---

47 Il modello Zthes (<http://zthes.z3950.org/>), ad esempio, è basato sul linguaggio XML - eXtensible Markup *Language* (<http://www.w3.org/XML/>) e si propone di facilitare l'interoperabilità tra applicazioni che utilizzano thesauri conformi a quanto previsto dalle norme ISO 2788 e ANSI/NISO Z39-19.

48 Cfr. Tiziana Calvitti, Elisabetta Viti, *Da ISO 2788 ai nuovi standard per la costruzione e l'interoperabilità dei vocabolari controllati: un'analisi comparativa*, in «Bollettino AIB», vol. 49, n. 3, settembre 2009, pp. 307-322.

49 Cfr. Emanuela Casson, *Dai thesauri ai vocabolari controllati: alcune novità introdotte nell'ultima edizione dello standard ANSI/NISO Z39.19-2005*, in «AIDAinformazioni», a. 24, n. 1-2, gennaio-giugno 2006, pp. 69-77.

50 Cfr. Groupe Langages documentaires de l'ADBS, *Les normes de conception, gestion et maintenance de thésaurus: évolution récentes et perspectives*, in «Documentaliste-Sciences de l'Information», vol. 44, n. 1, 2007, pp. 66-74.

51 Cfr. Stella G. Dextre Clarke, Marcia Lei Zeng, *From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling*, in «Information Standards Quarterly», vol. 24, n. 1, 2012, pp. 20-26.

re esplicite le relazioni 'a priori' fra i concetti»[52]; *«a controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms»*.

Dal confronto con la definizione della ISO 2788:1986 emerge che:

- Il thesaurus è un linguaggio controllato e strutturato, ma il suo utilizzo non si limita all'indicizzazione;
- Si distingue tra concetti e termini: questa distinzione restringe il divario tra thesauri e ontologie ed è indispensabile per l'uso in sistemi informatici. Il concetto è un'unità di pensiero indipendente dai termini impiegati per identificarlo, considerati come etichette. Prima della ISO 25964-1:2011 si parlava indistintamente di relazioni tra termini e di relazioni tra concetti. La norma chiarisce invece come solo per la relazione di equivalenza si possa parlare di relazione tra termini (essendo definita tra un termine preferito selezionato per rappresentare il concetto e i termini non preferiti considerati come ulteriori possibili etichette), mentre le due restanti relazioni thesaurali interessano i concetti;
- Si introduce la distinzione tra termini preferiti e sinonimi e quasi sinonimi.

In quest'ultimo punto consiste l'obiettivo primario del thesaurus in quanto strumento di recupero dell'informazione: ren-

[52] ISO 2788:1986, *op. cit*., p. 3.
Le relazioni paradigmatiche o *a priori* tra i concetti sono così definite dalla ISO 25964-1:2011: *«Relationship between concepts which is inherent in the concepts themselves»*. Si tratta cioè di relazioni che, contrariamente a quelle sintagmatiche, sono sempre valide indipendentemente dai contesti specifici di indicizzazione o di definizione del thesaurus. Le relazioni sintagmatiche sono perciò sconsigliate all'interno dei thesauri.

dere possibile l'incontro tra indicizzatore e utente e far sì che entrambi utilizzino lo stesso termine preferito per individuare un dato concetto. L'utente, cioè, attraverso i rinvii costituiti dai sinonimi e dai quasi sinonimi viene ricondotto verso il termine preferito scelto dall'indicizzatore recuperando l'informazione pur utilizzando una chiave di ricerca diversa da quella preferita[53]. In tal modo, inoltre, la ricerca può avvenire anche a partire dai termini appartenenti al vocabolario d'accesso.

Il thesaurus, in tal senso, funge da mediatore tra professionisti dell'informazione e utenti finali.

Una differenza fondamentale nell'impiego del thesaurus per la gestione dei documenti e delle informazioni in ambienti digitali riguarda gli utenti e il loro ruolo: da strumenti elaborati ed utilizzati esclusivamente dai professionisti dell'informazione a supporto delle pratiche di indicizzazione[54] e di ricerca in ambienti cartacei, diventano strumenti resi disponibili all'utenza generica impegnata in attività di ricerca, consultabili nella loro strutturazione, navigabili per permettere il recupero dei documenti e interpretabili dalle macchine. L'utente è quindi reso partecipe dell'organizzazione data alla conoscenza di un dato dominio e assume un ruolo attivo nell'utilizzo delle risorse termino-

---

53 In tal senso nella sezione *Terms and definitions* della norma si introducono anche le voci *Entry term* e *Search term.*

54 *«The act of describing or identifying a document in terms of its subject content»*.
UNI ISO 5963:1985, Documentazione - *Metodi per l'analisi dei documenti, la determinazione del loro soggetto e la selezione dei termini di indicizzazione,* 1985, p. 2.
«L'operazione mediante la quale si creano gli accessi al contenuto semantico del documento. Consta delle fasi di analisi concettuale e di traduzione dei concetti individuati e delle relazioni logiche individuate nei termini e nelle forme proprie del linguaggio di indicizzazione prescelto».
CHETI, A., *op. cit.*

logiche. Si riportano due estratti delle rispettive norme che testimoniano di questo mutamento:

> *La sua applicazione è limitata alle agenzie che utilizzano persone, quali indicizzatori, per analizzare i documenti ed esprimere i soggetti […]. Non è applicabile alle agenzie che utilizzano tecniche di indicizzazione completamente automatiche […]*[55].

> *Whereas in the past thesauri were designed for information professionals trained in indexing and searching, today there is a demand for vocabularies that untrained users will find to be intuitive, and for vocabularies that enable inferencing by machines*[56].

Pur non stravolgendo il concetto stesso di thesaurus e le relazioni che ne sono alla base, la nuova norma introduce ulteriori novità legate al ruolo prioritario di strumento di IR e si focalizza su aspetti precedentemente non contemplati per le motivazioni già illustrate. In particolare si registra la presenza di:

- Raccomandazioni sulla scelta dei software per la costruzione di thesauri, al fine di individuare le caratteristiche che gli stessi dovrebbero possedere per una corretta gestione di tali strumenti;
- *Modelli di dati* che definiscono in maniera astratta la struttura e la semantica di un thesaurus e che possono essere utilizzati per costruire strutture relazionali per database e formati di scambio, questi ultimi spesso basati su linguaggi di marcatura quale XML;
- Indicazioni relative all'integrazione dei thesauri in sistemi di indicizzazione e di ricerca dell'informazione (es. database bibliografici, centri di documentazione, banche dati docu-

[55] ISO 2788:1986, *op. cit.*, p. 3.
[56] ISO 25964-1:2011, *op. cit.*, p. VI.

mentali, basi di conoscenza, CMS[57], motori di ricerca, ecc.);
- Indicazioni più dettagliate sull'analisi a faccette per la costruzione di thesauri: come illustrato nel paragrafo 4.2, infatti, tale approccio si adatta meglio alle caratteristiche degli ambienti digitali;
- Nuove modalità di visualizzazione e di presentazione dei thesauri, date le potenzialità e la flessibilità garantite da ambienti web.

Quanto finora detto a proposito della norma ISO 25964 riguarda la prima delle due parti delle quali si compone: la seconda[58], pubblicata in data 4 marzo 2013, si sofferma sulla questione dell'interoperabilità tra vocabolari controllati (schemi di classificazione, ontologie, tassonomie, soggettari, terminologie, ecc.) e sulle operazioni di mappatura tra vocabolari diversi[59]. In alcune situazioni, quali ad esempio la ricerca in raccolte indicizzate con risorse diverse o l'utilizzo integrato di vocabolari controllati, è necessario stabilire una corrispondenza tra le differenti strutture concettuali. Pur nel pieno rispetto dei criteri di costruzione presenti nelle norme e pur interessando lo stesso dominio di conoscenza, infatti, tra due o più vocabolari possono esistere differenze sia tecniche dovute ai formati e ai sistemi informatici utilizzati, sia contenutistiche dovute all'utilizzo di terminologia settoriale o ad una diversa definizione delle relazioni, per cui in un thesaurus due termini possono essere considerati sino-

57 *Content Management Systems*.

58 ISO 25964-2: 2013, Information and Documentation – *Thesauri and interoperability with other vocabularies*, Part 2: *Interoperability with other vocabularies*.

59 O anche tra le diverse versioni linguistiche in thesauri multilingue. Finora tali operazioni sono state regolate dalle seguenti Linee Guida: Cfr. IFLA, WORKING GROUP ON GUIDELINES FOR MULTILINGUAL THESAURI, *Guidelines for multilingual thesauri*, IFLA, 2005.

nimi, mentre in un altro rappresentano due concetti distinti. Le operazioni di mappatura contribuiscono quindi ad un recupero più efficiente dell'informazione, poiché l'equivalenza tra voci appartenenti a più vocabolari permette di ritrovare tutte le risorse informative indicizzate tramite ciascuna di esse.

3.3 *Funzionalità di tassonomie e thesauri*

Come accennato, pur nella consapevolezza delle differenze tra le due tipologie di vocabolari controllati, tassonomie e thesauri assolvono pressoché alle medesime funzionalità.

I cambiamenti verificatisi negli ultimi decenni ai quali si è accennato hanno richiesto un'evoluzione del concetto di thesaurus e di tassonomia, ma al tempo stesso ne hanno valorizzato le funzionalità attraverso l'allargamento dei contesti d'uso e la dimostrazione delle loro potenzialità anche in ambienti digitali.

Riprendendo quando affermato a proposito delle figure interessate dall'utilizzo del thesaurus in relazione alle funzionalità, si può distinguere tra controllo terminologico, indicizzazione, supporto nella definizione dei metadati e classificazione da una parte, in quanto attività che continuano ad essere di competenza del professionista dell'informazione[60], e navigazione, ricerca ed espansione dei risultati delle ricerche dall'altra, poiché coinvolgono direttamente l'utente e/o la macchina.

---

[60] I più recenti sviluppi del Web stanno determinando anche per la pratica dell'indicizzazione una ridefinizione dei ruoli: si parla infatti di *social indexing* e di *folksonomy*, intese come forme di organizzazione della conoscenza nelle quali gli utenti modellizzano (attribuiscono parole chiave o classificano) sulla base del loro punto di vista e della loro visione di un dato dominio di conoscenza.
Cfr. OLIVIER ERTZSCHEID, GABRIEL GALLEZOT, *Etude exploratoire des pratiques d'indexation sociale comme une renégociation des espaces documentaires. Vers un nouveau big bang documentaire?*, in Document numérique et société, Charton G., Broudoux E. (a cura di), ADBS Éditions,

Controllo terminologico

Attraverso il controllo terminologico è possibile gestire l'ambiguità del linguaggio naturale e limitare il significato di un dato concetto al contesto di applicazione del thesaurus. In particolare tale funzione viene esercitata per disambiguare i concetti interessati dai fenomeni della polisemia e della sinonimia, attraverso la scelta del termine preferito, la strutturazione dei concetti per mezzo delle relazioni, che fornendo il contesto semantico di ciascuno di essi contribuiscono ad esplicitarne il significato, l'inserimento di note d'ambito, che forniscono definizioni o indicazioni riguardo all'impiego dei termini, e di qualificatori, che specificano l'ambito o la disciplina alla quale i concetti appartengono (soprattutto nei thesauri multidisciplinari).

Indicizzazione e supporto nella metadatazione

Come più volte menzionato, l'indicizzazione, intesa come l'azione di descrivere o identificare un documento nei termini del suo contenuto concettuale, è stata riconosciuta per lungo tempo come la funzionalità principale di un thesaurus, tant'è che lo stesso era definito *vocabolario di un linguaggio di indicizzazione controllato.* Seppur in contesti diversi, l'indicizzazione rimane una funzione fondamentale e i thesauri rappresentano fonti autorevoli dalle quali estrarre i concetti da attribuire ad una risorsa informativa di qualsivoglia natura al fine di permetterne la descrizione e il recupero.

2006; ZACKLAD, M., *Classification, thésaurus, ontologies, folksonomies : comparaison du point de vue de la recherche ouverte d'information (ROI),* in CAIS/ACSI 2007, 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information. Partage de l'information dans un monde fragmenté: Franchir les frontières, Montréal, 10-12 maggio 2007, Arsenault C., Dalkir, K. (a cura di), 2007; ÉLIE FRANCIS, ODILE QUESNEL, *Indéxation collaborative et folksonomies*, in «Documentaliste – Sciences de l'Information», vol. 44, n. 1, 2007, pp. 58-63.

Tassonomie e thesauri forniscono un supporto all'indicizzatore anche in virtù della loro struttura: l'organizzazione dei concetti gli consente non solo di identificare i termini, ma anche di determinare il livello di specificità con il quale si vuole rappresentare il contenuto concettuale dei documenti e che più si adatta alle caratteristiche della collezione e alle esigenze dei potenziali utenti, guidandolo verso concetti più generici, più specifici, o, solo per i thesauri, semanticamente correlati.

In ambiente digitale l'indicizzazione rientra nell'operazione di metadatazione delle risorse: gli standard di descrizione documentale, quale *Dublin Core* (DC)[61], prevedono, all'interno del set di metadati, un apposito elemento per l'inserimento di parole chiave rappresentative del contenuto delle risorse (*subject* nel DC). Anche in questo caso, come espressamente raccomandato dagli stessi standard, i vocabolari controllati possono essere utilizzati come fonte per la compilazione di metadati semantici[62] o indicizzazione semantica.

L'utilizzo di un linguaggio controllato a fini di indicizzazione, sebbene più dispendioso rispetto all'uso del linguaggio naturale o dei termini presenti nei titoli o nel testo dei documenti, contribuisce a limitare la soggettività e l'incoerenza che inevitabilmente caratterizza il lavoro di indicizzazione, soprattutto se eseguito da diversi indicizzatori. Presenta, inoltre, indubbi vantaggi in fase di ricerca data la scarsa coincidenza tra i termini utilizzati da questi ultimi e quelli utilizzati dagli utenti. Ne deriva anche un'applicazione più rigorosa del controllo terminologico favorendo in ogni situazione l'impiego dello stesso termine per rappresentare uno stesso concetto. Tale scelta è alla base dell'in-

61 <http://dublincore.org/>.

62 Relativi cioè al contenuto del documento. I metadati forniscono anche informazioni gestionali e relative alla proprietà intellettuale delle risorse informative.

dicizzazione detta *assegnata* o *per concetti*, che si contrappone a quella *derivata* o *per termini*[63].

Rappresentare il contenuto di un documento e, di conseguenza, recuperarlo in fase di ricerca, può comportare l'uso di più termini, la cui combinazione può avvenire in maniera pre- o post-coordinata[64]. Nel primo caso la modalità di combinazione è prevista e definita a priori e in fase di indicizzazione sulla base di regole di citazione che stabiliscono la sequenza secondo la quale i termini devono comparire in un'intestazione o stringa di soggetto[65] e tale rigidità potrebbe in molti casi compromettere il buon esito delle operazioni di ricerca. I contesti d'uso più frequenti sono, quindi, l'indicizzazione per soggetto e la collocazione di materiale librario. Nel secondo caso, invece, i termini vengono combinati solo al momento della ricerca e, per tale ragione, la post-coordinazione è la scelta più comune in ambiente digitale, vista la semplicità di effettuare delle ricerche utilizzando uno o più termini come chiave di accesso all'informazione. La metodologia di classificazione a faccette predilige tale approccio, che permette, attraverso la fase di sintesi, di non inserire termini eccessivamente lunghi e di non enumerarne quanti più possibile come nei sistemi tradizionali[66]. In una logica di recupero dell'informazione, sebbene un documento possa essere ritrovato anche a partire da uno solo dei descrittori attribuitigli, la

63 Cfr. SPINELLI, S., *op. cit.*

64 Cfr. CLAUDIO GNOLI, *Coordinazione, ordine di citazione e livelli integrativi in ambiente digitale*, in «Bibliotime», a. 6, n. 1, marzo 2003. <http://www.spbo.unibo.it/bibliotime/num-vi-1/gnoli.htm>.

65 Come nel caso del Nuovo Soggettario della Biblioteca Nazionale Centrale di Firenze, che prevede, oltre all'ordine di citazione, anche delle regole sintattiche che derivano da un'analisi dei ruoli svolti dai concetti contenuti nelle faccette e nelle categorie.

66 DOUGLAS TUDHOPE, CERI BINDING, *Faceted Thesauri*, in «Axiomathes», vol. 18, n. 2, giugno 2008, pp. 217-218.

post-coordinazione riduce il grado di precisione della ricerca, poiché il documento potrebbe rispondere solo parzialmente alle esigenze dell'utente[67].

Thesauri e tassonomie forniscono un supporto anche nelle tecniche di indicizzazione e di classificazione automatica[68]. Esemplificativo a tal proposito è il sistema AgroTagger[69], un estrattore di termini che attribuisce le voci indice ai nuovi documenti sulla base dei descrittori contenuti all'interno del thesaurus AGROVOC[70].

Rispetto ai tradizionali software di estrazione terminologica, basati prevalentemente su misure statistiche che calcolano la frequenza delle occorrenze dei termini all'interno dei testi, senza distinguere quelli che effettivamente sono rappresentativi del dominio di interesse da quelli relativi al linguaggio comune, gli applicativi basati su thesauri identificano all'interno dei testi stessi solo le voci che con molta probabilità ne rappresentano il contenuto concettuale. Mentre quindi nel primo caso l'indicizzazione richiede una fase di validazione e di selezione manuale, nel secondo avviene un'operazione di indicizzazione vera e propria.

Recupero dell'informazione

Sempre più frequentemente tassonomie e thesauri sono integrati in CMS, centri di documentazione, banche dati, ecc. come

---

67 Richiamo e precisione sono misure che in IR consentono di valutare l'esito delle ricerche di informazione. Il richiamo esprime il rapporto tra documenti rilevanti trovati e il totale dei documenti rilevanti esistenti, mentre la precisione indica il rapporto tra documenti rilevanti trovati e il totale dei documenti trovati. L'utilizzo di strumenti come i thesauri contribuisce ad accrescere il valore di queste misure.

68 Cfr. OLENA MEDELYAN, IAN H. WITTEN, *Thesaurus Based Authomatic Keyphase Indexing*, in Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA, ACM, 2006, pp. 296-297.

69 <http://aims.fao.org/agrotagger>.

70 <http://aims.fao.org/standards/agrovoc/about>.

strumenti di recupero dell'informazione e dei documenti.

In alcune applicazioni la struttura è resa disponibile all'utente, il quale può navigare al suo interno attraverso le relazioni precedentemente definite, al fine di recuperare le risorse informative associate a ciascuna voce durante la fase di indicizzazione (elemento *subject*). Ciascun documento può essere assegnato ad una o più entrate del thesaurus o della tassonomia, favorendo una classificazione multipla, e perciò flessibile, e garantendo più punti di accesso per il suo recupero. La navigazione del thesaurus, inoltre, permette all'utente di farsi un'idea sul contenuto dei documenti della collezione, sull'utilizzo della terminologia (in particolare in presenza di sinonimi e varianti) e sull'organizzazione concettuale del relativo dominio di conoscenza, soprattutto se non esperto. L'accesso all'informazione per mezzo del *browsing* può avvenire sia a partire dalla presentazione alfabetica delle entrate lessicali, sia da quella sistematica, ovvero dalla struttura classificatoria basata sull'organizzazione in categorie rappresentative del dominio per il quale il thesaurus o la tassonomia sono realizzati; in genere la prima modalità viene preferita di fronte ad un bisogno informativo meglio definito.

L'integrazione di una struttura thesaurale in un sistema di gestione del contenuto o in un software di ricerca e la possibilità di recuperare i documenti o le informazioni direttamente collegate ai concetti presuppone una scelta a monte in ordine alla definizione dell'architettura dell'informazione: inserire tutte le voci di un thesaurus esistente, anche se non esistono contenuti indicizzabili per mezzo di alcuni dei suoi descrittori o inserire (e integrare di volta in volta) solo le voci alle quali possono essere associate delle risorse informative. Nel primo caso il rischio è quello di effettuare delle ricerche che non producono alcun risultato, anche se disporre del thesaurus nella sua interezza permetterebbe di meglio gestire l'ampliamento della collezione documentale e di comprendere l'organizzazione concettuale del dominio; nel secondo il rischio risiede appunto nella parzialità della rap-

presentazione dei concetti, ma le ricerche produrrebbero in tutti i casi un set di risultati.

In aggiunta alla navigazione, la struttura del thesaurus può fornire un supporto all'utente nella scelta dei termini da impiegare come chiavi di ricerca e quindi nella formulazione della query.

Nei sistemi di Information Retrieval che integrano tassonomie e thesauri per migliorare i risultati ottenuti a seguito di un'interrogazione, sfruttando le relazioni semantiche definite tra termini e concetti utilizzati per indicizzare i documenti, le parole che compongono la query vengono confrontate con i termini del vocabolario controllato. Al termine di una ricerca il sistema può proporre all'utente delle possibilità di raffinamento dei risultati ottenuti attraverso la visualizzazione delle relazioni, oppure può automaticamente inserire nei risultati i documenti indicizzati con termini correlati a quelli impiegati per l'interrogazione.

Nello specifico in un thesaurus:

- La relazione di equivalenza fa sì che un documento venga recuperato anche se l'utente, per effettuare l'interrogazione, non ha utilizzato il termine preferito attribuito alla risorsa dall'indicizzatore. In tal senso i termini del gruppo di equivalenza rivestono un'importanza notevole, in quanto punti di accesso all'informazione che anticipano le possibili modalità di ricerca da parte degli utenti finali e per tali motivi è importante inserire tutti quelli potenzialmente utili a favorire l'incontro tra il professionista dell'informazione e l'utente;
- La relazione NT consente di specificare meglio la ricerca e ridurre gli item che potrebbero essere recuperati. Si parla in questo caso di *query extension*;
- Le relazioni BT e RT consentono di ampliare la ricerca nel caso vengano restituiti pochi documenti. Si parla in questo caso di *query expansion*.

La Figura 4 mostra l'utilizzo del LISA (*Library and Information Science Abstracts*) Thesaurus per il raffinamento della ricerca. L'utente può migliorare i risultati ottenuti attraverso i concetti sovra e sotto ordinati e i concetti correlati.

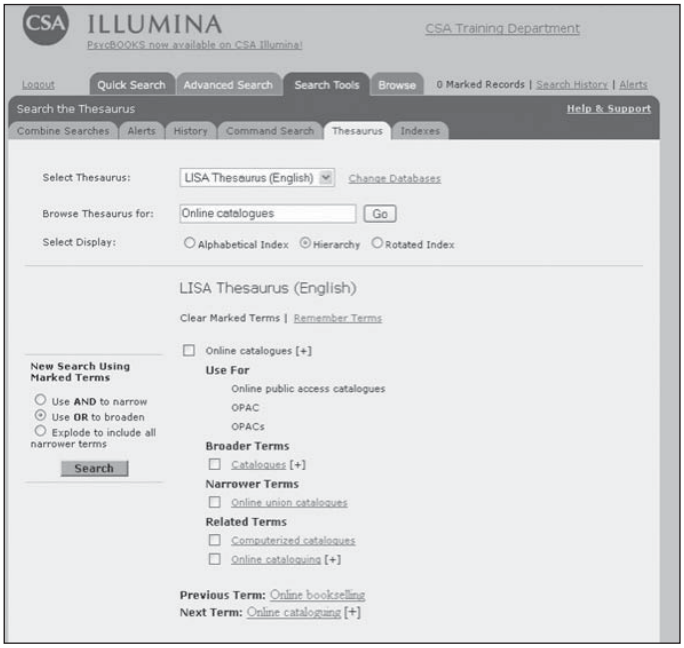

Figura 4. Thesaurus e IR[71].

Organizzazione della conoscenza di dominio

Thesauri e tassonomie possono essere progettati e realizzati anche per la rappresentazione di un dominio di conoscenza o di una disciplina: in questo senso i concetti in esse contenuti sono rappresentativi di un settore della realtà nella sua interezza piuttosto che di una collezione di documenti da indicizzare e da re-

71 <http://www.csa.com/factsheets/supplements/LISAguide.pdf>.

cuperare. Le relazioni semantiche e, laddove presente, l'organizzazione dei concetti per mezzo di categorie rappresentative, forniscono anche una classificazione del dominio di interesse, che può diventare un punto di riferimento per la sistematizzazione della conoscenza, la predisposizione di documenti, lo scambio non ambiguo di informazione, la normalizzazione e la guida nell'utilizzo della terminologia.

In questo caso, quindi, i termini da inserire all'interno del thesaurus non devono essere scelti sulla base della collezione da indicizzare (inserendo quindi solo i termini potenzialmente utilizzabili per l'indicizzazione e per il successivo recupero delle risorse informative) ma devono costituire un set il più possibile rappresentativo del dominio da rappresentare.

3.4 *Tipologie di thesauri*

È possibile distinguere tra diverse tipologie di thesauri sulla base della copertura semantica, delle lingue di compilazione e della struttura classificatoria sottostante. Rispetto al primo criterio i thesauri possono essere generali[72] (o multidisciplinari o ad ampio spettro) o speciali[73].

---

[72] Di questa categoria fanno parte ad esempio: l'UNESCO Thesaurus (<http://databases.unesco.org/thesaurus/>) che interessa le seguenti discipline: istruzione, cultura, scienze naturali, scienze umane e sociali, comunicazione e informazione, l'AAT - *Art and Architecture Thesaurus* (<http://www.getty.edu/research/tools/vocabularies/aat/index.html>) che organizza concetti relative ad oggetti, artisti, luoghi legati all'arte, all'architettura e più in generale alla cultura; l'AGROVOC thesaurus (<http://aims.fao.org/standards/agrovoc/about>) che si occupa di alimentazione, agricoltura, pesca, ambiente e altri domini correlati; ecc.

[73] Relativi a domini specifici. Ne sono esempi il MeSh – *Medical Subject Headings* (<http://www.nlm.nih.gov/mesh/>), il NASA Thesaurus (<http://data.nasa.gov/nasa-thesaurus/>), l'*Alzheimer's Disease Thesaurus* (<http://adlib.alzheimers.org/adear/alzdb/thesaurus.aspx>), lo *European Education Thesaurus* (<http://www.freethesaurus.info/redined/en/index.php>), ecc.

Anche se multidisciplinari o generali, i thesauri interessano sempre domini della conoscenza ben identificati e in numero limitato. Essi infatti, anche in ragione dei principi del controllo terminologico, non nascono con la pretesa di rappresentare l'intero scibile, contrariamente agli obiettivi propri dei sistemi di classificazione tradizionali. Ciò rappresenta, infatti, una delle principali differenze tra queste due tipologie di sistemi di organizzazione della conoscenza.

Relativamente alle lingue di compilazione si distingue tra thesauri monolingue e thesauri multilingue, interessati da due norme distinte fino all'emanazione della ISO 25964-1:2011 e ora in essa confluite essendo alcuni principi validi per entrambe le tipologie.

La predisposizione di thesauri multilingue si rende necessaria al fine di consentire l'accesso e il recupero dell'informazione a partire da risorse informative disponibili in più lingue e indipendentemente dalla lingua di indicizzazione.

Vi si stabiliscono quindi relazioni di equivalenza interlinguistica così che sia l'indicizzatore sia l'utente non si vedano imposto l'utilizzo di una lingua dominante. Le problematiche che emergono per la definizione di tale strumento dipendono principalmente dalle differenze non solo linguistiche, ma anche concettuali, che esistono tra le lingue e, conseguentemente tra le relative culture, nelle quali il thesaurus viene compilato.

Le versioni linguistiche devono essere quanto più possibile sovrapponibili; la ISO 25964-1:2011 stabilisce che ciascuna lingua debba avere lo stesso status e che ciascun concetto debba essere rappresentato in ognuna di esse. La sola distinzione possibile è quella tra lingua d'origine e lingua di destinazione: la prima rappresenta il punto di partenza dal quale promana la traduzione o la ricerca degli equivalenti nelle lingue restanti, mentre la seconda è la lingua verso la quale si indirizza l'attività di traduzione o di ricerca.

Rappresentare i concetti in ciascuna lingua significa identifi-

care il corrispettivo semanticamente più vicino di ciascuno di essi e stabilire relazioni di equivalenza interlinguistica bidirezionali. Tuttavia, così come accade all'interno di un thesaurus monolingue, i possibili livelli di equivalenza si dispongono lungo un continuum che va da quella esatta all'assenza di equivalenza. Il primo caso, che corrisponde alla sinonimia assoluta, si verifica nel momento in cui è possibile identificare termini preferiti in ciascuna delle lingue di compilazione senza differenze di tipo semantico o culturale[74]. È possibile che tra due o più concetti esista invece una relazione di equivalenza non assoluta o quasi equivalenza, del tutto simile alla quasi sinonimia nel thesaurus monolingue: tra i termini nelle diverse lingue esistono differenze di significato, spesso dovute a differenti connotazioni culturali[75]. La relazione viene comunque stabilita se si ritiene che ai fini del thesaurus esse rappresentino il relativo concetto, affidando in alcuni casi ad una nota d'ambito l'esplicitazione della diversa copertura semantica. Qualora, invece, una delle lingue possieda un termine che rappresenta un concetto sovraordinato rispetto a quello rappresentato dai termini delle altre, si parla di equivalenza parziale o *broader/narrower equivalence.* Rientra in questo caso anche la *compound equivalence*, che si verifica quando una lingua possiede uno o più equivalenti parziali, che, in combinazione, rappresentano il concetto identificato da un solo termine nella lingua d'origine[76]. Infine, la non equivalenza interessa quei casi in cui non esistono, in una o più lingue, termini rappresentativi di un concetto espresso in un'altra lingua[77]. Spesso alla ba-

[74] Es. IT Sole – EN Sun – FR Soleil – DE Sonne.
[75] Es. IT Educazione – EN Education.
[76] Es. IT Sicurezza – EN Security + Safety.
[77] Es. *Grandes-écoles* è un concetto tipico del contesto francese, in quanto indica istituti di istruzione superiore di livello universitario che mancano nel sistema scolastico italiano. Nel caso di un thesaurus multilingue il termine verrebbe inserito in quanto tale o verrebbe tradotto letteralmente, ri-

se vi sono differenze nei contesti e nelle culture dei paesi in cui vengono parlate le lingue di compilazione del thesaurus. In quest'ultimo caso gli espedienti ai quali si ricorre possono essere quello di accettare un prestito linguistico o quello di coniare un neologismo o un calco, soprattutto laddove il prestito non sarebbe di immediata comprensione per gli utenti.

Sia la ISO 25964-1:2011, sia le Linee Guida IFLA (*International Federation of Library Associations and Institutions*) suggeriscono modalità di gestione di tutti i casi in cui non è possibile stabilire una relazione di equivalenza esatta. In generale si tratta di: note d'ambito, qualificatori soprattutto in presenza di omografi, l'assegnazione dello status di termine preferito ad un termine più generico che più si avvicina al concetto espresso in un'altra lingua.

La relazione di equivalenza interessa solo i termini non preferiti: tra quelli non preferiti non è necessario stabilire alcuna relazione, anche perché il loro numero e il loro significato potrebbero cambiare significativamente da una lingua all'altra.

La trasposizione interlinguistica deve interessare tutte le relazioni del thesaurus: in tal senso si distingue tra struttura simmetrica e asimmetrica. Nel primo caso esiste una corrispondenza tra tutti i concetti contenuti nel thesaurus in tutte le lingue di compilazione e la struttura concettuale determinata dalle relazioni gerarchiche e associative è condivisa. Nel caso di asimmetria, invece, non è possibile stabilire una corrispondenza esatta perché le relazioni definite in una lingua potrebbero non essere valide per un'altra, soprattutto nel caso di lingue molto distanti tra di loro. È necessario, quindi, ricorrere a operazioni di mappatura tra le diverse versioni linguistiche parallele del thesaurus multilingue al fine di stabilire delle corrispondenze che consentano la na-

chiedendo sempre la presenza di una nota d'ambito che ne espliciti il significato e il contesto d'uso.

vigazione e l'interoperabilità al fine di non privilegiare nessuna delle lingue del thesaurus forzando la corrispondenza tra i concetti e creando strutture che non sarebbero accettate e condivise dagli utenti. La definizione delle modalità di mappatura sono specificatamente demandate alla seconda parte della ISO 25964.

La Figura 5 mostra il dettaglio del termine *Delinquenza* nel thesaurus multilingue dell'Unione Europea EuroVoc. La ricerca può avvenire a partire da ciascuna delle lingue previste dal thesaurus e per ogni termine scelto vengono mostrati gli equivalenti interlinguistici preceduti dai codice identificativi delle relative lingue, definiti dalle ISO 639-1 e ISO 639-2. In generale, nella

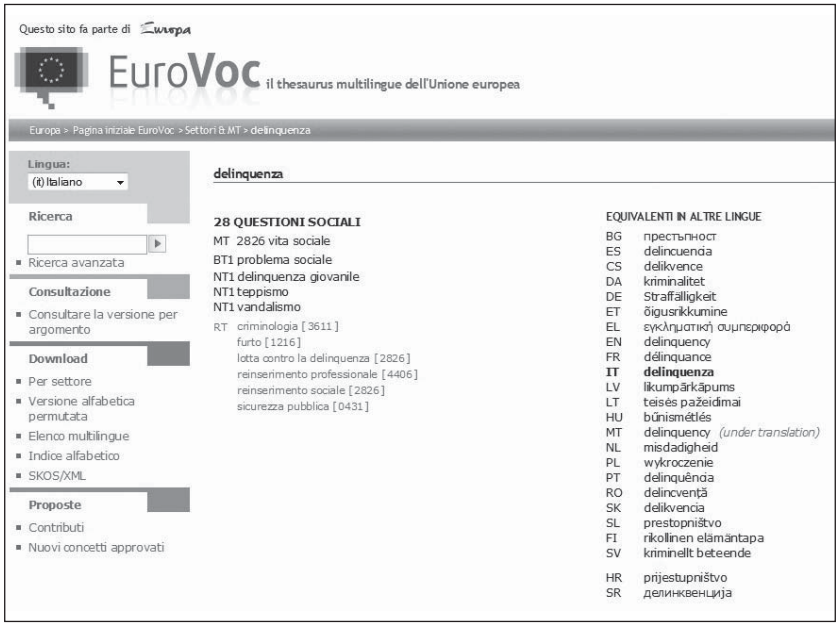

Figura 5. Thesaurus multilingue EuroVoc[78].

78 <http://eurovoc.europa.eu/drupal/?q=it>.

presentazione alfabetica viene mostrato il contesto strutturale di ciascun concetto, mentre in caso di presentazione sistematica, sarebbe opportuno poter visualizzare la struttura di due o più versioni linguistiche contemporaneamente al fine di comprenderne le corrispondenze.

Per quanto riguarda, invece, la costruzione di thesauri multilingue, la norma illustra tre approcci, che nell'ordine presentano un grado di complessità di costruzione crescente, ma consentono di ottenere risultati migliori in termini di rispetto del contesto culturale delle lingue di compilazione:

- Traduzione di un thesaurus monolingue esistente: in questo caso è alto il rischio che la lingua d'origine diventi dominante rispetto a quella di destinazione e che la struttura rispecchi poco le aspettative degli utenti;
- *Merging* di diverse versioni linguistiche: questo approccio presterebbe maggiore attenzione alle differenze linguistico-concettuali dei thesauri da integrare non assegnando a nessuna lingua un ruolo predominante. La complessità risiede nelle differenti scelte, soprattutto nel grado di specificità, compiute nei diversi thesauri;
- Costruzione ex novo del thesaurus multilingue: la costruzione simultanea fa sì che ogni lingua sia a turno d'origine e di destinazione.

3.4.1 *Thesauri e analisi a faccette*

La stretta interdipendenza tra classificazione a faccette e thesaurus è messa in evidenza in (Broughton, 2008a)[79], che afferma che *«quando si costruisce una classificazione a faccette, si prepara anche un thesaurus e per costruire un thesauro, si deve passare per una classificazione»*. In un altro studio (Broughton,

[79] BROUGHTON , V., (a), *op. cit.*, p. 13.

2008b)[80], l'autrice mette in evidenza come il valore dell'analisi a faccette in quanto supporto notevole nella realizzazione di un thesaurus non sia stato riconosciuto dagli standard o dalle linee guida esistenti, se non molto di recente. A ciò si aggiunga l'impiego spesso inappropriato del termine *faccetta*, che ha spesso ingenerato confusione circa il suo reale significato e, di conseguenza, il suo corretto utilizzo nei sistemi di classificazione e nei thesauri. Come dimostrato in (Spiteri, 1999)[81], infatti, ne vengono fornite definizioni diverse anche all'interno di quegli stessi thesauri che sono basati su tale sistema:

> *IBE, and UNICEF, for example, define facets as groups that cover related concepts. In BINDING and GENRE, facets are "gathering terms" used to arrange the hierarchical relationships amongst broader and narrower terms. ROOT and YOUTH both state that facets are fundamental categories, but do not explain what this means. In AAT, facets are homogeneous, mutually exclusive units of information that share characteristics that demonstrate their differences from each other.*

Come accennato, invece, la norma ISO 25964-1:2011 dedica un'intera sezione alla presentazione dell'analisi a faccette e all'applicazione della stessa nel processo di costruzione di un thesaurus. Vengono a tal proposito fornite le seguenti definizioni: «*Facet: Grouping of concepts of the same inherent category*»; «*Facet analysis*: *Analysis of subject areas into constituent con-*

---

80 VANDA BROUGHTON (b), *A faceted classification as the basis of a faceted terminology: conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification*, ed. 2, in «Axiomathes», vol. 18, Springer, 2008, p. 196.

81 LOUISE F. SPITERI, *The Essential Element of Faceted Thesauri*, in «Cataloging & Classification Quarterly», vol. 28, n. 4, The Haworth Press, Inc, 1999, p. 7.

cepts grouped into facets, and the subdivision of concepts into narrower concepts by specified characteristics of division»; «*Faceted classification scheme: Classification scheme in which subjects are analyzed into their constituent facets*».

Riprendendo la citazione precedente di Vanda Broghton, si può quindi affermare che applicare i principi dell'analisi a faccette alla costruzione di un thesaurus significa classificarne i concetti organizzandoli sulla base di un set di categorie precedentemente identificate e rappresentative del dominio per il quale lo stesso viene costruito e che tale strutturazione fornisce un supporto significativo nella corretta definizione delle relazioni thesaurali e quindi nella costruzione del thesaurus nella sua presentazione gerarchica[82]. Anche (Aitchison et alii, 2000)[83] sostengono che una classificazione a faccette[84] possa rappresentare un punto di partenza o una fonte per la costruzione di un thesaurus.

In ogni caso, la struttura a faccette e la visualizzazione gerarchica del thesaurus risultano complementari: in una presentazione sistematica a faccette, infatti, se si escludono le relazioni gerarchiche di tipo genere-specie, graficamente rappresentate per mezzo di rientri, non si tiene traccia delle relazioni parte-tutto e di quelle associative e di equivalenza, definite, invece, nella presentazione gerarchico-alfabetica del thesaurus.

---

[82] La relazione gerarchica (BT-NT) può essere derivata dall'organizzazione dei termini in sottofaccette, nel senso che i concetti introdotti da un principio di suddivisione sono sotto-ordinati del concetto al quale il principio stesso viene applicato (ciò è valido su più livelli di strutturazione), mentre la relazione associativa riguarda i concetti introdotti dal medesimo *node label* e quindi collocati sullo stesso livello gerarchico o concetti appartenenti a diverse faccette e tra i quali si identifica una relazione semantica.

[83] JEAN AITCHISON, DAVID BAWDEN, ALAN GILCHRIST, *Thesaurus Construction and use: a practical manual*, ed. 4, Londra, ASLIB, 2000, p. 69.

[84] Per una presentazione delle differenze in termini di obiettivi e struttura sistema di classificazione e thesaurus si veda TUDHOPE, D., BINDING, C., *op. cit.*, pp. 211-222.

Le origini dell'analisi a faccette risalgono all'ideazione e alla pubblicazione nel 1934 della *Faceted Classification* (FC) o *Colon Classification*[85] (CC) da parte di Ranganathan, bibliotecario e matematico indiano, che propose un approccio decisamente innovativo rispetto alle tradizionali classificazioni biblioteconomiche comunemente adottate per la sistemazione del materiale librario[86]. A partire dalla propria esperienza nell'utilizzo della CDD, egli ne individuò i principali limiti, quali l'impossibilità a rappresentare tutti i temi trattati in un'opera, ad enumerare tutti i soggetti o ad accoglierne di nuovi e formulò i principi di un nuovo approccio che avrebbe consentito di superarli attraverso un sistema basato su operazioni di scomposizione e ricomposizione dei soggetti da classificare[87]. Alla base di tale principio vi è l'i-

---

85 Così chiamata perché utilizza il segno di punteggiatura dei due punti (in inglese Colon) come separatore tra i soggetti.

86 La Classificazione Decimale Dewey (CDD) Progettato da Melvil Dewey per l'Amherst College nel 1873, rappresentò una vera e propria rivoluzione del campo della biblioteconomia in quanto introdusse il metodo della notazione decimale che consente un «*ordinamento monodimensionale di ospitalità infinita*» (ALFREDO SERRAI, *Le classificazioni: idee e materiali per una teoria e per una storia*, Firenze, Leo S. Olschki Editore, 1970, p. 283), nel senso che la successione delle classi avviene secondo un solo principio di suddivisione per volta, ma è possibile estendere la struttura in maniera potenzialmente infinita. La Classificazione Decimale Universale (CDU), derivata dalla precedente, fu elaborata da Paul Otlet e Henri La Fontaine nel biennio 1893-1894. Rispetto alla Dewey, il suo obiettivo principale fu quello di classificare oggetti documentali piuttosto che quello di collocare il materiale bibliotecario. Tali sistemi sono ancora oggi ampliamente applicati.

87 SHIYALI RAMAMRITA RANGANATHAN, *Colon Classification, I: Schedules for Classification*, ed. 7, Gopinath M.A. (a cura di), Sarada Ranganathan Endowment for Library Science, 1989, (ed. 1, 1933), p. 3.
La classificazione del materiale bibliotecario richiede, dunque, una fase di analisi e di scomposizione del soggetto sulla base delle categorie identificate, seguita da un'attività di traduzione del linguaggio naturale in linguaggio controllato attraverso la verifica dei concetti nelle tavole di clas-

dentificazione di cinque categorie fondamentali o faccette che permettono di analizzare qualsiasi soggetto, poiché ognuna di esse ne mette in evidenza un particolare aspetto. Esse sono: *Personality*, *Matter*, *Energy*, *Space*, *Time* (PMEST)[88].

Nonostante la classificazione a faccette abbia rappresentato un'intuizione innovativa per il superamento dei limiti propri dei sistemi gerarchici, la sua applicazione in contesto bibliotecario è fin da subito risultata complessa, anche solo per il semplice fatto di dover assegnare ad un volume più collocazioni. Tuttavia, le sue potenzialità sono state riscoperte nelle pratiche di indicizzazione e di recupero dell'informazione in ambiente digitale data la virtualità degli oggetti da classificare e la necessità di sistemi multidimensionali (Marino, 2004)[89]. Lo sviluppo quindi di thesauri a faccette[90] ha ricevuto e sta ricevendo un forte impulso e l'attenzione dedicatagli dalla recente norma ISO ne è una conferma.

L'organizzazione dei contenuti nel web ha visto l'applicazione massiccia tanto di sistemi gerarchici che di sistemi a faccette, seppure non nel pieno rispetto dei principi sui quali sono fondati (Rosenfeld, Morville, 2002)[91]. I sistemi di classificazione tra-

---

sificazione e nel linguaggio ordinale attraverso l'attribuzione del codice di notazione a ciascuno di essi. La fase di sintesi prevede la definizione di un codice unico dato dalla ricomposizione dei singoli codici attribuiti agli aspetti nei quali il soggetto è stato analizzato.

88 *Personality*: oggetti di studio delle varie discipline; *Matter* proprietà o materiali; *Energy*: le azioni o i processi che si verificano in una disciplina; *Space*: concetti relativi allo spazio; *Time*: concetti relativi al tempo.

89 Cfr. VITTORIO MARINO, *Classificazioni per il Web. I vantaggi dell'adozione di schemi a faccette*, Associazione Italiana Biblioteche (AIB) - WEB, 2004.
<http://www.aib.it/aib/contr/marino1.htm>

90 Tra i principali thesauri a faccette rientra il già citato AAT, che rappresenta un'applicazione rigorosa ed esemplificativa dei principi alla base di tale approccio.

91 LOUIS ROSENFELD, PETER MORVILLE, *Information Architecture for the World Wide Web*, ed. 2, O'Reilly, 2002, p. 208.

dizionali a cui si è accennato, e più in generale, i sistemi gerarchici, sono basati sull'enumerazione di tutte le classi e sono caratterizzati dalla difficoltà di accogliere integrazioni, se non a condizione di modifiche consistenti dello schema di base, e da una struttura che costringe l'utente a navigare solo secondo il percorso definito. I sistemi a faccette sono, invece, come decisamente più flessibili. Questi infatti, non precludendo integrazioni successive in termini di categorie e principi di analisi, si adattano all'evoluzione e alle esigenze di aggiornamento dei contenuti sul Web, garantendo pluridimensionalità, persistenza, scalabilità e flessibilità (Rosati, 2003)[92].

Significativo, in tale ottica è anche il lavoro svolto dal *Classification Research Group (CRG)*[93], che ha accolto i principi della metodologia a faccette approfondendoli e perfezionandoli soprattutto in riferimento alla revisione del suddetto schema di faccette fondamentali di Ranganathan e alla conseguente definizione di un set di faccette più ampio e di più immediata applicabilità. Lo schema definito dal CRG è così costituito: *thing, types, parts, properties, materials, processes, activities, products, by products, patients, agents, space and time*[94].

Il punto di forza di tale schema risiede nella sua potenziale

---

92 LUCA ROSATI, *La classificazione a faccette fra Knowledge Management et Information Architecture (parte I)*, It Consult, 2003. <http://www.itconsult.it/knowledge/articoli/pdf/itc_rosati_faccette_e_KM.pdf>.
Classificazione sulla base di molteplici attributi; Cambiamenti limitati dovuti al fatto che le proprietà rappresentano attributi intrinseci dei concetti; Possibilità di aggiungere nuove faccette e nuovi principi di suddivisione; Ricerca a partire da un solo attributo da più attributi in combinazione.

93 Gruppo di ricercatori inglesi attivi nel campo della biblioteconomia e della classificazione costituitosi a Londra all'inizio degli anni 50 del secolo scorso.

94 Le definizioni di ciascuna faccetta sono tratte da BROUGHTON, V. (a), *op. cit.*, pp. 259-281.

applicazione a qualsiasi dominio oggetto di interesse, data la genericità delle categorie e, al tempo stesso, l'elevata probabilità di essere rappresentative degli aspetti di un dato ambito semantico. Data la notevole specificità di ciascun settore e le differenti finalità che possono essere alla base della costruzione di un simile strumento di classificazione, è possibile adottare anche parzialmente lo schema proposto, scartando alcune faccette, che risultano non applicabili, accorpandone delle altre o anche prevedendone alcune aggiuntive, laddove quelle iniziali non dovessero rivelarsi sufficienti per la descrizione completa del dominio.

Applicare i principi della classificazione a faccette alla costruzione di un thesaurus implica la scomposizione del dominio di interesse in categorie (faccette) rappresentative dello stesso: così come un oggetto viene analizzato nelle sue caratteristiche intrinseche, un dominio viene analizzato negli aspetti che ne ricoprono l'intero ambito semantico. All'interno di ciascuna faccetta i concetti possono essere ulteriormente organizzati attraverso principi di suddivisione o *node labels* o etichette di snodo

«Cose: comprende i concetti che sono i principali oggetti di studio per un argomento o disciplina; Parti: comprende i concetti che sono parti dei concetti della categoria delle entità; Proprietà: concetti che sono proprietà o attributi di concetti appartenenti alla categoria principale; Materiali: Raccoglie i concetti collegati a sostanze e materiali di tutti i tipi […]; Processi: raccoglie i concetti di azioni che accadono spontaneamente, non compiute da agenti umani; Attività: raccoglie i concetti di azioni condotte su di un oggetto da un agente umano; Pazienti: raccoglie i concetti che sono oggetti di azioni, […] Dovrebbe comprendere gli oggetti impiegati in fasi intermedie di processi produttivi quando i prodotti finali sono le entità primarie; Prodotti: comprende i prodotti di attività quando questi non appartengono alla categoria primaria delle entità; Prodotti intermedi: raccoglie i prodotti intermedi di attività […];Agenti/Strumenti: comprende i concetti per mezzo dei quali si compiono delle azioni […]; Spazio: raccoglie i concetti relativi a luoghi […]; Tempo: comprende i concetti legati al tempo […]».

che rappresentano caratteristiche intrinseche dei concetti stessi[95]. La Figura 6 mostra un estratto del Nuovo Soggettario della Biblioteca Nazionale Centrale di Firenze, nel quale è possibile distinguere i principi di suddivisione (tra parentesi quadre) e la strutturazione dei termini al loro interno.

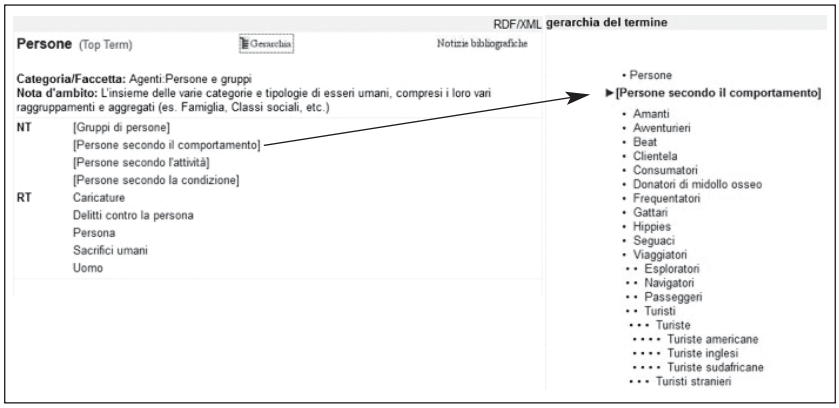

Figura 6. Nuovo Soggettario.

Tra i principi fondamentali alla base di un sistema a faccette rientrano:

– Approccio analitico – sintetico: rappresenta l'aspetto che maggiormente contraddistingue tale metodo rispetto alle classificazioni enumerative. Si susseguono due fasi: la prima orientata all'analisi del dominio di interesse o degli oggetti da classificare al fine di individuare gli aspetti secondo i quali possono essere scomposti; la seconda a sintetizzare e combinare i concetti appartenenti a più suddivisioni in stringhe più complesse[96];

[95] Data la loro funzione, non sono considerati termini o concetti all'interno del thesaurus.

[96] Il processo di sintesi interessa anche i codici eventualmente associati a ciascun concetto ed è in tal senso che si parla di notazione sintetica.

- Mutua esclusività: garantisce che ciascun insieme o sottoinsieme di termini venga introdotto da un sola caratteristica di suddivisione per volta e che si evitino fenomeni di sovrapposizione semantica[97];
- Ordine di elencazione delle faccette[98]: stabilisce l'ordine in base al quale le faccette devono essere presentate in un layout sistematico del thesaurus o in un menù di navigazione; va dal generale al particolare, secondo un principio di concretezza crescente (quindi, nel caso del set del CRG, da *Tempo* a *Cose*), in quanto ritenuto più intuitivo da parte degli utenti;
- Ordine di citazione standard: ordine che determina la sintassi in base alla quale i concetti appartenenti a diverse faccette dovrebbero essere inseriti all'interno della stringa di soggetto in un contesto di pre-coordinazione (Iyer, 2012)[99]; va dal particolare al generale, quindi prevede una sequenzialità inversa rispetto al principio precedente.

Per quanto riguarda, invece, l'ordinamento dei concetti all'interno dei raggruppamenti, è consigliabile inserirli in base ad un principio che non sia quello alfabetico, molto poco significativo,

---

97 *«[…] i termini che appartengono ad una stessa sottofaccetta come mutuamente esclusivi, ossia escludentisi a vicenda. Ciò significa che, diversamente dai termini che appartengono a differenti faccette o a differenti sottofaccette di una stessa faccetta, i termini che appartengono ad una stessa sottofaccetta non possono combinarsi tra loro, non possono dare luogo ad una sovrapposizione o intersezione di classe (p.e., Coniugi celibi). Questa proprietà (mutua esclusione) indica la fine del processo di divisione di una classe»*.
BIBLIOTECA NAZIONALE CENTRALE DI FIRENZE, *Nuovo Soggettario*, Milano, Editrice Bibliografica, 2006, p. 82.

98 BROUGHTON, V. (a), *op. cit*., p. 201.

99 HEMALATA IYER, *Classificatory Structures: Concepts, Relations and Representation*, Würzburg, Ergon Verlag, 2012, p. 128.

ma piuttosto quello cronologico, o in funzione alla complessità, della sequenzialità, ecc.

L'approccio a faccette risulta estremamente adatto ad un dominio specialistico data l'omogeneità degli oggetti da classificare, così che gli stessi possano essere analizzati a partire da caratteristiche comuni. L'adozione in contesti multidisciplinari o multi tematici risulterebbe, per tale motivo, decisamente più complessa. Lo svantaggio che anche le precedenti norme riconoscevano alla struttura a faccette, infatti, risiedeva nella separazione, attraverso il loro inserimento in faccette diverse, di concetti appartenenti ad una stessa disciplina o ambito.

3.5 *Costruzione di un thesaurus*

Così come per le funzionalità, anche le modalità di costruzione di una tassonomia non sono esplicitamente descritte nelle citate norme. Tuttavia, se si escludono le attività e le indicazioni che conducono verso la definizione di elementi propri dei thesauri, in particolare la gestione della sinonimia e dei concetti correlati, le fasi di costruzione si possono considerare comuni ai due sistemi. La raccolta dei termini, l'organizzazione gerarchica e l'eventuale identificazione delle faccette, infatti, interessano entrambi. Si farà comunque riferimento ai thesauri, essendo il relativo processo di costruzione maggiormente standardizzato.

Preliminare al processo di costruzione vero e proprio è la fase di progettazione, durante la quale, oltre allo studio di fattibilità che interessa aspetti propriamente gestionali (es. risorse umane e materiali), è importante stabilire:

- a chi si rivolge, intendendo soprattutto gli utenti finali che lo utilizzeranno come supporto nelle loro ricerche o come fonte di termini, distinguendo principalmente tra esperti di dominio e utenti comuni senza competenze specifiche nel settore;
- per quali scopi viene costruito, ovvero indicizzazione di una specifica collezione documentale, IR, organizzazione

della conoscenza, ecc., dal momento che alcune scelte dipendono strettamente dalla funzionalità alla quale lo strumento dovrà assolvere (Folino et alii, 2012)[100];
- l'eventuale integrazione in un sistema di *content management*;
- la costruzione ex novo o la traduzione/adattamento di un thesaurus esistente;
- la struttura di classificazione alla base dell'organizzazione dei concetti.

La prima fase di costruzione del thesaurus consiste nella raccolta dei termini da inserire al suo interno. Tale attività può avvenire a partire da risorse terminologiche esistenti in letteratura e relative al medesimo dominio di interesse del thesaurus, quali lessici o glossari, e dall'estrazione dei termini a partire da un corpus documentale[101] appositamente costituito.

D'accordo con la linguistica dei corpora, di cui il corpus rappresenta l'oggetto di studio e la cui finalità è quella di analizzare l'utilizzo della lingua in contesti d'uso reali (Lenci et alii, 2005)[102], il set di documenti costituito deve essere un campione rappresentativo rispetto alla popolazione di riferimento, nel sen-

[100] ANTONIETTA FOLINO, FRANCESCA IOZZI, MARIA TAVERNITI, *Gestione documentale in ambiente digitale*, in «Archivistica e Documentazione», Guarasci R. (a cura di), vol. 7, Marzi, Cosenza, Comet Editor Press, 2012, pp. 152-158.

[101] *«A collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research».* JOHN SINCLAIR, *Trust the text: language, corpus and discourse*, Londra, Routledge, 2004, p. 14.

[102] ALESSANDRO LENCI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, *Testo e computer: Elementi di Linguistica Computazionale*, Roma, Carocci Editore, 2005, p. 25.

so che le informazioni ottenute dalla sua analisi devono poter essere generalizzate alla popolazione intera, ovvero alla lingua o ad una sua varietà. Per tali ragioni, deve essere sufficientemente grande[103] ed equilibrato in termini di tipologie testuali in esso contenute (es. leggi, norme, articoli scientifici, riviste di settore, opuscoli informativi, ecc.), sempre nel rispetto degli obiettivi del thesaurus. I tipi di risorse infatti si differenziano per il pubblico a cui si rivolgono e per il grado di specializzazione della terminologia impiegata, caratteristiche importanti per la selezione dei termini da inserire in un thesaurus.

I documenti non sono, quindi, creati ad hoc, ma vengono selezionati tra quelli esistenti sulla base di determinati criteri, che, nel caso specifico, variano in funzione degli obiettivi e delle caratteristiche che il thesaurus dovrà possedere, affinché possano rendere conto del reale utilizzo della terminologia di interesse. Tra i criteri più rilevanti ai fini della costruzione di un thesaurus rientrano la lingua dei documenti selezionati (corpus monolingue o multilingue) e l'ambito tematico (corpus specialistico o generale).

La scelta di costruire un corpus documentale nel dominio di interesse del thesaurus permette di non limitare i termini a quelli estratti dalla collezione di documenti da indicizzare[104], sia in ragione di un eventuale ampliamento di tale insieme, sia per prevedere il maggior numero di termini potenzialmente utilizzabili dagli utenti.

L'estrazione terminologica eseguita sul corpus può avvenire manualmente o, nella maggior parte dei casi e soprattutto in presenza di corpora di grandi dimensioni, in maniera semiautomati-

[103] Le dimensioni di un corpus si misurano in *tokens*, ovvero in parole-unità distinte.

[104] Che, quindi, si differenzia dal corpus per il non rispetto della rappresentatività statistica.

ca[105], con l'ausilio di strumenti software dedicati[106]. In quest'ultimo caso, a seguito del processo di estrazione, si ottiene una lista di candidati termini[107] dalla quale selezionare quelli effettivamente rappresentativi del dominio di conoscenza che deve essere strutturato per mezzo del thesaurus.

---

[105] Si parla di semiautomatismo poiché l'intervento umano per la validazione dei candidati termini estratti resta ancora insostituibile.

[106] Tali software si basano sull'assunto che dalla frequenza con la quale un termine occorre all'interno di un documento e dell'intero corpus, dipenda la sua rappresentatività per il dominio oggetto di analisi: quindi maggiore è il valore della frequenza, maggiore è la sua significatività. Per la lingua italiana si può fare riferimento al software T2K (text-to-Knowledge) sviluppato dall'Istituto di Linguistica Computazionale di Pisa, il quale si basa sia su misure statistiche, in particolare sulla funzione tf*idf (*term frequency*inverse document frequency)*, che calcola la frequenza di ogni *termine* all'interno di un documento (*TF = term frequency*), relazionata con la frequenza inversa del *termine* stesso all'interno del *corpus* documentale (IDF= *Inverse Document Frequency*), che su regole linguistiche. L'analisi di un testo prevede ad esempio fasi di segmentazione in parole (tokenizzazione), di attribuzione della categoria grammaticale (*Part-of-Speech Tagging*), di riconoscimento di sintagmi sintattici (*chunking*), di identificazione di parole grammaticali (articoli, congiunzioni, preposizioni, ecc.) da escludere dal processo di estrazione terminologica (stop list). Oltre al glossario terminologico si ottengono in output anche relazioni semantiche tra i termini estratti: gerarchiche e di similarità semantica. Per uno studio più approfondito di tali tematiche si rimanda a FELICE DELL'ORLETTA, ALESSANDRO LENCI, SIMONE MARCHI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, GIULIA VENTURI, *Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio*, in «Aida Informazioni», a. XXVI gennaio-giugno, n. 1-2, 2008, pp. 185-206.

[107] *«Le 'parole' estratte automaticamente per mezzo di strumenti informatici, acquisiscono la dignità d'essere definite 'descrittore' oppure 'termine' solo in seguito ad un processo di validazione da parte di un esperto di dominio»*.
MARIA TAVERNITI, *Fra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della Pubblica Amministrazione*, in «AIDAinformazioni», a. XXVI gennaio-giugno, n. 1-2, 2008, p. 232.

Tale selezione dovrebbe avvenire sulla base della combinazioni di più criteri, tra i quali la frequenza d'uso[108], la conoscenza di esperti di dominio e l'autorevolezza della fonte dalla quale i termini sono stati estratti.

Successivamente alla costituzione del set di termini da inserire nel thesaurus, si passa alla normalizzazione degli stessi secondo quanto previsto dalla norma ISO 25964-1:2011. Ci si riferisce in particolare alla forma del termine, ovvero alla scelta del singolare o del plurale[109], della categoria grammaticale[110], all'inserimento di termini composti[111], ecc.

Riguardo a quest'ultimo punto, è opportuno sottolineare come in un thesaurus a faccette si protenda, laddove possibile, verso la scomposizione dei concetti complessi, soprattutto quando i termini singoli che li costituiscono possono appartenere a più faccette al fine di evitare l'inserimento di termini troppo lunghi, che aumenterebbero la complessità e i livelli gerarchici del thesaurus. La combinazione potrà avvenire, secondo il principio della sinte-

[108] La frequenza d'uso non può da sola determinare l'accettabilità o meno di un termine, soprattutto in presenza di concetti innovativi, che in quanto tali avranno poche occorrenze all'interno di un corpus documentale.

[109] In genere si adotta il plurale per quei termini che rappresentano entità concrete e quindi numerabili (es. Edifici) e il singolare in presenza di concetti astratti e non numerabili (es. Architettura).

[110] Si preferisce ad esempio il verbo sostantivato alla forma dell'infinito presente.

[111] In presenza di concetti complessi, la norma distingue tra casi in cui è preferibile mantenere e inserire nel thesaurus i termini nella loro forma composta e casi in cui è preferibile inserire il risultato della loro scomposizione. A favore della prima opzione rientrano ad esempio le seguenti motivazioni: frequenza d'uso; rischio di ambiguità; perdita o mutamento di significato; presenza di metafore. In generale, soprattutto per finalità di indicizzazione, si tiene conto del grado di specificità che si intende raggiungere, tenendo presente che i termini composti aumentano tale livello, e della capacità di discriminazione dei concetti in base al contenuto della collezione da indicizzare.

si, nell'attribuzione del soggetto al documento o, in una logica di post-coordinazione, nella formulazione di un'interrogazione.

L'organizzazione di tali termini prevede la definizione delle relazioni semantiche, così come illustrate nel paragrafo 3.1 e l'assegnazione, nel caso di thesauri a faccette, di ciascun termine alla o alle categorie di appartenenza.

La definizione della struttura del thesaurus, soprattutto in quest'ultimo caso può seguire un approccio induttivo o deduttivo: nel primo caso si procede dall'alto verso il basso definendo preliminarmente la struttura gerarchica e quindi l'insieme delle faccette e subito dopo dei principi di suddivisione per poi organizzare i termini al loro interno nella misura in cui questi vengono identificati e selezionati, mentre nel secondo si parte dal basso, ovvero dai termini scelti, la cui analisi permette di definire la struttura per mezzo della quale organizzarli. Nella maggior parte dei casi i due approcci sono complementari.

Le modalità di visualizzazione della struttura del thesaurus illustrate nella ISO 2788:1986 sono: alfabetica, sistematica, distinguendo ulteriormente tra organizzazione in settori o discipline e organizzazione a faccette, e grafica; la complementarità tra di esse veniva garantita dalla presenza della notazione. La norma, si ribadisce, si riferiva ai thesauri a stampa, per cui alcune delle indicazioni fornite e delle problematiche evidenziate non sono più valide in ambiente digitale.

Nel layout alfabetico i termini sono elencati secondo quest'ordine e per ciascuno di essi vengono evidenziate le relazioni nelle quali è coinvolto; in quello sistematico i termini sono organizzati generalmente in gerarchie (ad esempio a partire da *top term* che rappresentano i concetti con il più alto livello di genericità, quali le discipline) o in categorie nel caso di thesaurus a faccette (è possibile anche un'organizzazione mista)[112].

[112] La visualizzazione è possibile anche tramite liste permutate nelle quali i

Gli accorgimenti nella costruzione della versione a stampa riguardano ad esempio i casi di poligerarchia, per cui un termine con più di una collocazione veniva inserito completo delle relazioni nelle quali era coinvolto solo nella gerarchia principale alla quale apparteneva, prevedendo per le altre solo una forma di rinvio: il tutto per esigenze di spazio e di semplificazione della presentazione del thesaurus. Simili indicazioni non hanno ovviamente senso di esistere nei contesti digitali, per i quali già la norma britannica introduce il concetto di display informatico e di strumenti software per la costruzione dei thesauri. Ciascun termine è ad esempio inserito una volta per tutte pur moltiplicandone le possibili collocazioni: è l'utente infatti ad avere la possibilità di navigare ed esplodere le relazioni thesaurali secondo le proprie esigenze, scegliendo gli estratti che vuole rendere visibili e muovendosi, attraverso link ipertestuali, non solo da un concetto ad un altro, ma anche da un concetto ai documenti ad esso associati. Egli può navigare tramite le relazioni e passare da una forma di presentazione all'altra (multilivello, singolo termine, per categoria, per top term, per faccette, ecc.) superando la linearità delle presentazioni a stampa. Questo tipo di display viene definito *classificato esteso* (Calvitti, Viti, 2009)[113].

Da un punto di vista meramente tecnico, è opportuno accennare ai software oggi esistenti per la costruzione dei thesauri in formato elettronico, al fine di evidenziarne le caratteristiche principali che garantiscono una corretta e coerente realizzazione di tale strumento:

– Non imporre limiti al numero di termini e/o relazioni che è possibile definire;

termini composti vengono elencati in ordine alfabetico e ciascun descrittore di cui si compongono può o meno mantenere la propria posizione nella stringa (KWIC - *Keyword in context;* KWOC - *Keyword out of Context*).

[113] Calvitti, T., Viti, E., *op. cit.*, p. 314.

- Inserire in maniera automatica le relazioni inverse se le stesse sono simmetriche (es. BT/NT);
- Esportare il thesaurus in formati standard, quali XML e, nel contesto del Web semantico, anche SKOS;
- Prevedere diversi layout;
- Impedire la definizione di relazioni non corrette (che, ad esempio, coinvolgono termini non preferiti);
- Permettere l'organizzazione a faccette e quindi la gestione delle stesse e dei *node label*;
- Garantire funzioni di ricerca e di navigazione all'interno del thesaurus;
- Gestire il multilinguismo.

4 Conclusioni

Come più volte ribadito nel corso del capitolo, l'evoluzione delle tecnologie per il recupero dell'informazione e, in particolare, l'ampia diffusione delle ontologie come strumenti per eccellenza del Web Semantico, hanno determinato un ripensamento dei concetti stessi di thesaurus e tassonomia. Una simile rivalutazione ha fatto sì che gli stessi non continuassero ad essere considerati come strumenti ormai obsoleti, determinando la valorizzazione e il potenziamento delle loro funzionalità in ambiente digitale.

È importante sottolineare, quindi, come le differenze esistenti tra thesauri e ontologie, alle quali si è fatto accenno, non debbano indurre a considerare il thesaurus come uno strumento semanticamente meno ricco e per questo meno valido di un'ontologia: le due tipologie di KOS, infatti, sono nate in risposta ad esigenze diverse e la predisposizione dell'uno o dell'altro dipende dalle caratteristiche e dagli obiettivi dei contesti di applicazione.

Proprio in virtù di tali differenze, quindi, esistono dei casi in

cui un'ontologia[114] assolverebbe meglio di un thesaurus a determinate funzionalità[115]: le metodologie di passaggio da un thesaurus ad un'ontologia sono oggetto di interesse scientifico e in letteratura esistono proposte di approcci sperimentati su thesauri esistenti. Si tratta essenzialmente di processi di arricchimento semantico orientati all'identificazione della natura delle relazioni esistenti tra i concetti. L'approccio proposto da (Chrisment et alii, 2006) per la trasformazione di un thesaurus in un'ontologia leggera di dominio può essere applicata ai thesauri monolingue con struttura gerarchica realizzati conformemente alle norme ISO 2788:1986 e ANSI Z39.19:2005. La metodologia è stata concretamente sperimentata sul thesaurus di astronomia *dell'International Astronomical Union* (IAU) e la sua originalità rispetto ad altri studi aventi i medesimi obiettivi risiede nel fatto che le operazioni di trasformazione sono state condotte non solo a partire dal thesaurus, ma anche da un corpus documentale appositamente costituito, al fine di estrarre le informazioni implicite o non modellizzate nel thesaurus stesso. In particolare, la relazione gerarchica di tipo genere-specie nell'ontologia è definita a partire dalle relazioni di tipo BTG-NTG presenti nel thesaurus. Questa fase prevede, inoltre, l'inserimento di classi più generi-

[114] Per ontologia si intende la rappresentazione formale ed esplicita di una concettualizzazione condivisa, interpretabile tanto da un operatore umano che da una macchina.
Cfr. NICOLA CAPUANO, *Ontologie OWL: Teoria e Pratica*, in «Computer Programming», n 148, luglio-agosto 2005, pp. 59-64.

[115] Si pensi ad esempio all'utilizzo congiunto di ontologie e di sistemi capaci di effettuare delle inferenze e di fare dei ragionamenti a partire dalle informazioni modellizzate, estraendo da queste conoscenza nuova ed implicita e sfruttando le restrizioni espresse sulle relazioni e le proprietà attribuite alle relazioni stesse (transitività, simmetria, ecc.). Esse inoltre rappresentano lo strumento per eccellenza del Semantic Web e consentono di sviluppare sistemi di IR con funzionalità avanzate e che consentono una maggiore interazione con gli utenti.

che, poiché spesso i thesauri al livello gerarchico più elevato contengono un gran numero di termini, al fine di agevolare le operazioni di navigazione. L'identificazione di tali classi, che rappresentano o concetti di dominio o concetti generici per strutturare il dominio stesso, avviene attraverso due processi automatici: il raggruppamento dei concetti le cui etichette presentano la stessa testa lessicale per creare delle classi ad un primo livello della gerarchia e la definizione di categorie astratte (proprietà, fenomeni, eventi, strumenti, oggetti, ecc.) alle quali associare i concetti di dominio a partire da un'ontologia generica esistente, ovvero WordNet[116]. La specificazione delle relazioni associative avviene a partire dall'analisi sintattica del corpus documentale che permette di estratte il contesto linguistico delle etichette relative a ciascun concetto. Tali contesti sono poi raggruppati sulla base delle categorie astratte alle quali i concetti appartengono, le relazioni sono inizialmente definite a livello di categorie astratte, ma vengono poi inserite tra i concetti a queste appartenenti a partire dai termini interessati da una relazione RT all'interno del thesaurus. Ulteriori relazioni associative non presenti nel thesaurus sono identificate a partire dall'analisi dei documenti del corpus, basandosi sulla frequenza dei termini che cooccorrono con le etichette dei concetti e sui risultati di un'analisi distribuzionale che tiene conto della similitudine dei contesti in cui occorrono i sintagmi.

(Soergel et alii, 2004)[117] descrive invece l'approccio definito per la trasformazione in ontologia del thesaurus multilingue AGROVOC. Il modello proposto prevede una netta distinzione tra il livello concettuale, che si riferisce al significato, il livello terminologico, relativo ai termini utilizzati per rappresentare i

[116] <http://wordnet.princeton.edu/>.

[117] Cfr. DAGOBERT SOERGEL, BORIS LAUSER, ANITA LIANG, FREHIWOT FISSEHA, JOHANNES KEIZER, STEPHEN KATZ, *Reengineering Thesauri for New Applications: the AGROVOC Example*, in «Digit. Inf.», vol. 4, n. 4, 2004.

concetti, e il livello di stringa, ovvero le varianti lessicali possibili per ciascun termine. I concetti devono essere assegnati a categorie generiche (processi, sostanze, ecc.) che vincolano le tipologie di relazioni dalle quali gli stessi possono essere interessati. L'inserimento di nuove informazioni e la precisazione di quelle esistenti avviene con l'ausilio di un *ontology editor* e attraverso il riconoscimento di pattern ricorrenti sui quali l'editor formula delle regole applicabili a casi identificati come simili.

Di notevole interesse anche le più attuali ricerche nell'ambito del Web Semantico o Web di dati, che riguardano lo sviluppo di *linked data*[118] e l'integrazione/allineamento dei sistemi di organizzazione della conoscenza. Esemplificativa in tal senso la Figura 7 che riporta i *dataset open* disponibili sulla rete e collega-

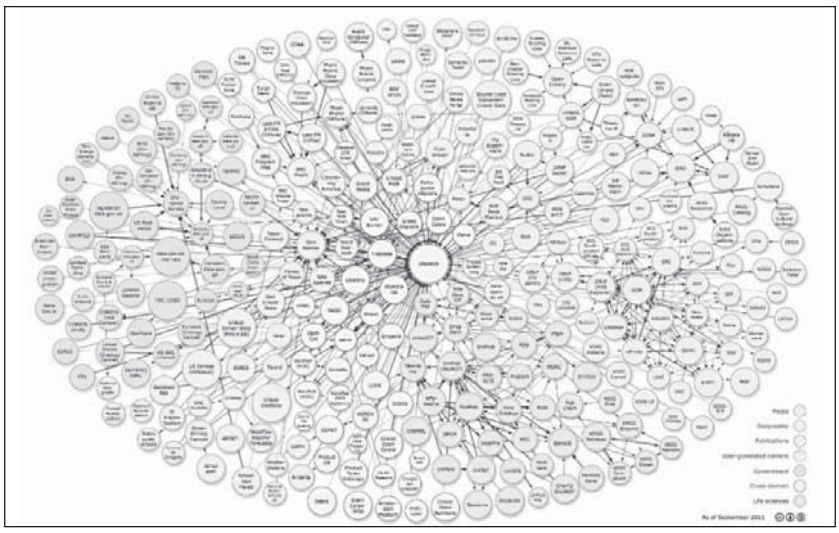

Figura 7. Linked Open Data Cloud.

[118] *«Si costruisce così un reticolo di dati collegati (linked data, appunto) appartenenti a un dominio (che costituisce il contesto di partenza), collega-*

ti tra di loro. La sfida che l'integrazione/condivisione dei dati porta con sé è rappresentata dalla possibilità di accedere a tutte le risorse informative indicizzate tramite i concetti appartenenti a vocabolari controllati interconnessi e allineati tra di loro, per cui tali studi risultano di notevole importanza per la definizione di nuovi strumenti di organizzazione della conoscenza e per l'integrazione di quelli esistenti al fine di garantire un accesso federato e al tempo stesso controllato all'informazione e ai dati presenti nel Web.

Bibliografia

ANSI/NISO Z39-19:2005, *Guidelines for the construction, format, and management of monolingual controlled vocabularies*

AITCHISON, J., BAWDEN, D., GILCHRIST, A., *Thesaurus Construction and use: a practical manual*, ed. 4, Londra, ASLIB, 2000

BIBLIOTECA NAZIONALE CENTRALE DI FIRENZE, *Nuovo Soggettario*, Milano, Editrice Bibliografica, 2006

BROUGHTON, V., (a), *Costruire Thesauri: strumenti per indicizzazione e metadati semantici*, Cavaleri, P. (a cura di), Ballestra L., Venuti L. (traduzione di), Milano, Editrice Bibliografica, 2008

BROUGHTON, V., (b), *A faceted classification as the basis of a faceted terminology: conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification*, ed. 2, in «Axiomathes», vol. 18, Springer, 2008, pp. 193-210

BS 8723:2004-2008, *Structured vocabularies for information retrieval – Guide*

CABRÉ, M.T., *Terminology: theory, methods and applications*, Sager J.C. (ed.), DeCesaris J.A. (traduzione di), Philadelphia PA, John Benjamins, 1998

CALVITTI, T., VITI, E., *Da ISO 2788 ai nuovi standard per la costruzione e*

to a sua volta ad altri set di dati esterni, ovvero fuori dal dominio, in un contesto di relazioni sempre più estese».

MAURO GUERRINI, TIZIANA POSSEMATO, *Linked data: un nuovo alfabeto del Web Semantico*, in «Biblioteche oggi», aprile 2012, p. 7.

l'interoperabilità dei vocabolari controllati: un'analisi comparativa, in «Bollettino AIB», vol. 49, n. 3, settembre 2009, pp. 307-322

CAPUANO, N., *Ontologie OWL: Teoria e Pratica*, in «Computer Programming», n 148, luglio-agosto 2005, pp. 59-64

CASSON, E., *Dai thesauri ai vocabolari controllati: alcune novità introdotte nell'ultima edizione dello standard ANSI/NISO Z39.19-2005*, in «AIDAinformazioni», a. 24, n. 1-2, gennaio-giugno 2006, pp. 69-77

CHETI, A., *Manuale ipertestuale di analisi concettuale*, 1996 <http://biblioteche.unibo.it/manuals/html_1/HOME.HTML>

DELL'ORLETTA, F., LENCI, A., MARCHI, S., MONTEMAGNI, S., PIRRELLI, V., VENTURI, G., *Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio*, in «Aida Informazioni», a. XXVI gennaio-giugno, n. 1-2, 2008, pp. 185-206

DEXTRE CLARKE, S.G., LEI ZENG, M., *From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling*, in «Information Standards Quarterly», vol. 24, n. 1, 2012, pp. 20-26

ERTZSCHEID, O., GALLEZOT, G., *Etude exploratoire des pratiques d'indexation sociale comme une renégociation des espaces documentaires. Vers un nouveau big bang documentaire?*, in Document numérique et société, Charton G., Broudoux E. (a cura di), ADBS Éditions, 2006

FLORIO, I., *La normativa standardizzata per la gestione delle documentazione tra Italia e Francia*, Rubbettino Editore, 2011

FOLINO, A., IOZZI, F., TAVERNITI, M., *Gestione documentale in ambiente digitale*, in «Archivistica e Documentazione», Guarasci R. (a cura di), vol. 7, Marzi, Cosenza, Comet Editor Press, 2012

FRANCIS, É., QUESNEL, O., *Indéxation collaborative et folksonomies*, in «Documentaliste – Sciences de l'Information», vol. 44, n. 1, 2007, pp. 58-63

GNOLI, C., *Coordinazione, ordine di citazione e livelli integrativi in ambiente digitale*, in «Bibliotime», a. 6, n. 1, marzo 2003 <http://www.spbo.unibo.it/bibliotime/num-vi-1/gnoli.htm>

GNOLI, C., MARINO, V., ROSATI, L., *Organizzare la conoscenza: dalle biblioteche all'architettura dell'informazione per il web*, Milano, Tecniche Nuove, 2006

GROUPE LANGAGES DOCUMENTAIRES DE L'ADBS, *Les normes de conception, gestion et maintenance de thésaurus: évolution récentes et perspectives*, in «Documentaliste-Sciences de l'Information», vol. 44, n. 1, 2007, pp. 66-74

GUERRINI, M., POSSEMATO, T., *Linked data: un nuovo alfabeto del Web Semantico*, in «Biblioteche oggi», aprile 2012, pp. 7-15

HODGE, G., *Systems of Knowledge Organization for Digital libraries. Beyond traditional authority files*, 2000 <http://www.clir.org/pubs/reports/pub91/contents.html>

IFLA, WORKING GROUP ON GUIDELINES FOR MULTILINGUAL THESAURI, *Guidelines for multilingual thesauri*, IFLA, 2005

ISO 25964-1:2011, Information and documentation – *Thesauri and interoperability with other vocabularies*, Part 1: *Thesauri for information retrieval*

ISO 25964-2:2013, Information and Documentation – *Thesauri and interoperability with other vocabularies*, Part 2: *Interoperability with other vocabularies*

ISO 2788:1986, Documentation – *Guidelines for the establishment and development of monolingual thesauri*

ISO 5964:1985, Documentation – *Guidelines for the establishment and development of multilingual thesauri*

IYER, H., *Classificatory Structures: Concepts, Relations and Representation*, Würzburg, Ergon Verlag, 2012

LEI ZENG, M., SALABA, A., *Toward an International Sharing and Use of Subject Authority Data*, FRBR Workshop, OCLC, 2005

LENCI, A., MONTEMAGNI, S., PIRRELLI, V., *Testo e computer: Elementi di Linguistica Computazionale*, Roma, Carocci Editore, 2005

MARINO, V., *Classificazioni per il Web. I vantaggi dell'adozione di schemi a faccette*, Associazione Italiana Biblioteche (AIB) - WEB, 2004 <http://www.aib.it/aib/contr/marino1.htm>

MEDELYAN, O., WITTEN, I. H., *Thesaurus Based Authomatic Keyphase Indexing*, in Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA, ACM, 2006, pp. 296-297

RANGANATHAN, S.R., *Colon Classification, I: Schedules for Classification*, ed. 7, Gopinath M.A. (a cura di), Sarada Ranganathan Endowment for Library Science, 1989, (ed. 1, 1933)

RAYBURN, G., *Taxonomies and Thesauri*, 2011 <http://www.llrx.com/system/files?file=taxonomiesthesauri.pdf>

RIEDIGER, H., *Cos'è la terminologia e come si fa un glossario*, 2012 <http://www.term-minator.it/corso/doc/mod3_termino_glossa.pdf>

ROSATI, L., *La classificazione a faccette fra Knowledge Management et Information Architecture (parte I)*, It Consult, 2003 <http://www.itconsult.it/knowledge/articoli/pdf/itc_rosati_faccette_e_KM.pdf>

ROSENFELD, L., MORVILLE, P., *Information Architecture for the World Wide Web*, ed. 2, O'Reilly, 2002

SERRAI, A., *Le classificazioni: idee e materiali per una teoria e per una storia*, Firenze, Leo S. Olschki Editore, 1970

SINCLAIR, J., *Trust the text: language, corpus and discourse*, Londra, Routledge, 2004

SOERGEL, D., LAUSER, B., LIANG, A., FISSEHA, F., KEIZER, J., KATZ, S., *Reen-*

gineering Thesauri for New Applications: the AGROVOC Example, in «Digit. Inf.», vol. 4, n. 4, 2004

SPINELLI, S., *Introduzione ai thesauri*, 2005 <http://biocfarm.unibo.it/~spinelli/indicizzazione/thesauri.htm>

SPINELLI, S., *Introduzione all'indicizzazione*, 2006 <http://biocfarm.unibo.it/~spinelli/indicizzazione/>

SPITERI, L.F., *The Essential Element of Faceted Thesauri*, in «Cataloging & Classification Quarterly», vol. 28, n. 4, The Haworth Press, Inc, 1999, pp. 31-52

TAVERNITI, M., *Fra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della Pubblica Amministrazione*, in «AIDAinformazioni», a. XXVI gennaio-giugno, n. 1-2, 2008, pp. 239-250

TUDHOPE, D., BINDING, C., *Faceted Thesauri*, in «Axiomathes», vol. 18, n. 2, giugno 2008, pp. 211-222

UNI ISO 5963:1985, Documentazione - *Metodi per l'analisi dei documenti, la determinazione del loro soggetto e la selezione dei termini di indicizzazione*, 1985

ZACKLAD, M., *Classification, thésaurus, ontologies, folksonomies : comparaison du point de vue de la recherche ouverte d'information (ROI)*, in CAIS/ACSI 2007, 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information.Partage de l'information dans un monde fragmenté: Franchir les frontières, Montréal, 10-12 maggio 2007, Arsenault C., Dalkir, K. (a cura di), 2007

Sitografia

<http://adlib.alzheimers.org/adear/alzdb/thesaurus.aspx>
<http://aims.fao.org/agrotagger>
<http://aims.fao.org/standards/agrovoc/about>
<http://aims.fao.org/standards/agrovoc/about>
<http://data.nasa.gov/nasa-thesaurus/>
<http://databases.unesco.org/thesaurus/>
<http://dublincore.org/>
<http://eurovoc.europa.eu/drupal/?q=it>
<http://thesaurus.com/Roget-Alpha-Index.html>
<http://wordnet.princeton.edu/>
<http://www.csa.com/factsheets/supplements/LISAguide.pdf>
<http://www.freethesaurus.info/redined/en/index.php>
<http://www.getty.edu/research/tools/vocabularies/aat/index.html>
<http://www.iso.org/iso/home.html>

<http://www.nlm.nih.gov/mesh/>
<http://www.w3.org/>
<http://www.w3.org/2004/02/skos/>
<http://www.w3.org/RDF/>
<http://www.w3.org/TR/owl-features/>
<http://www.w3.org/XML/>
<http://zthes.z3950.org/>