Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

ONTOLOGIE LEGGERE A FACCETTE

Fausto Giunchiglia,Vincenzo Maltese

January 2010

Technical Report # DISI-10-005

Also: published on AIDA Informazioni. Rivista di Scienze dell'Informazione, n. 3-4/2010

Ontologie leggere a faccette [1]

Fausto Giunchiglia,Vincenzo Maltese
Dipartimento di Ingegneria e Scienza dell'Informazione (DISI)
Università di Trento, Trento, Italy

Abstract. In questo articolo ci concentriamo sull'uso delle ontologie per l'organizzazione di oggetti, quali ad esempio foto, libri e pagine Web. Le *ontologie leggere* sono ontologie con una struttura gerarchica ad albero dove a ciascun nodo è associata un'etichetta in linguaggio naturale. Nelle *ontologie leggere a faccette* le etichette sono organizzate secondo modelli ben definiti, i quali catturano specifici aspetti della conoscenza, ovvero le *faccette*. A tal fine, ci basiamo sull'approccio Analitico-Sintetico, una ben radicata metodologia usata con successo per decenni in biblioteconomia, soprattutto in India, per la classificazione di libri. Le ontologie leggere a faccette hanno una struttura ben definita ed, in quanto tali, risultano più facili da creare, condividere tra gli utenti, e più appropriate in applicazioni semantiche, dove cioè viene automaticamente analizzato e sfruttato il significato ontologico dei termini.

Parole chiave: ontologie**,** ontologie leggere, faccette, classificazioni formali.

1 Introduzione

Le ontologie vengono utilizzate presso molte comunità, per scopi e modalità a volte anche molto differenti. Esistono molteplici tipi di ontologie, a seconda del diverso grado di formalità, complessità della struttura ed espressività del linguaggio usato per rappresentarle [1]. Nella pratica, due sono le applicazioni per le quali vengono più di frequente utilizzate, ovvero per *descrivere* o *organizzare* oggetti. In questo articolo ci concentriamo principalmente sul secondo uso, ovvero, ci concentriamo sul problema della classificazione di oggetti, quali ad esempio foto, libri, pagine Web.

Le *ontologie leggere* sono ontologie con una struttura gerarchica ad albero, dove ciascun nodo è associato ad un'etichetta in linguaggio naturale, ad esempio in italiano. Noi spesso utilizziamo il temine *ontologie leggere formali* quando intendiamo quelle ontologie le quali possono essere ottenute da ontologie leggere trasformando (seguendo un opportuno procedimento) le etichette originarie in formule matematiche espresse in un linguaggio formale, quale ad esempio la logica descrittiva [12]. Tali formule catturano il significato semantico che l'utente intende comunicare attraverso l'etichetta. Ad esempio, con l'etichetta *mela verde* l'utente con molta probabilità intende tutte le mele di colore verde[2]. Nelle ontologie leggere formali, le formule associate a ciascun nodo sono in relazione di sussunzione, ovvero la formula di ciascun nodo è sempre più specifica di quella associata al nodo padre (che si trova nel livello superiore della struttura ad albero) [1, 31]. Ad esempio, se il nodo *mela* è posizionato sotto il nodo *verde*, la formula associata al nodo figlio codificherà il concetto di *mela verde*[3]. Dettagli sul processo di trasformazione, corredati da esempi, possono essere trovati in [2]. Nel resto dell'articolo, noi utilizzeremo indifferentemente i termini ontologia leggera e ontologia leggera formale. Il contesto renderà sempre chiaro cosa intendiamo.

Le ontologie leggere hanno un ruolo fondamentale nella classificazione automatica di documenti [1, 16], in sistemi di ricerca automatici [1, 21] ed anche nella soluzione di problemi di eterogeneità semantica (ad esempio stesso termine ma diverso significato oppure termini diversi ma stesso o simile significato) tra differenti ontologie [15, 18, 19, 20]. Esse

rappresentano uno strumento essenziale ed estremamente potente che può essere sfruttato con profitto per l'automazione del ragionamento per la gestione di dati e della conoscenza.

Nonostante ciò, l'adozione di ontologie (leggere) non è stata finora così diffusa quanto ci si potrebbe aspettare da quando il lavoro intorno al cosiddetto Web semantico [32] è cominciato. I principali motivi riteniamo siano da ricercare nella mancanza di interesse o nella difficoltà a comprendere tale strumento da parte degli utenti, in particolare nel come costruire tali ontologie [4, 5] e soprattutto nel fatto che le ontologie sviluppate per un certo scopo possono essere difficilmente riutilizzate per altri scopi o da altri utenti [5]. In altre parole, la loro costruzione richiede un impegno ed investimento notevoli che non sempre ripagano gli sforzi profusi.

L'obiettivo principale che ci prefiggiamo con questo articolo è quello di introdurre le *ontologie leggere a faccette* quale soluzione molto promettente per i problemi che abbiamo appena sottolineato. Esse sono definite in termini di *faccette*[4], recentemente adottate con grande successo per la progettazione di interfacce di siti Web, in quanto si dimostrano particolarmente utili nell'organizzazione e la navigazione dei dati. Si veda ad esempio lo studio fatto da La Barre [23] ed in particolare il lavoro fatto nel contesto del progetto Flamenco[5] (si veda ad esempio [24]), ma anche il lavoro descritto in [7, 8, 9] come applicazione nel campo della gestione della conoscenza che è in qualche modo affine, in spirito, al nostro lavoro.

Noi costruiamo le faccette seguendo l'approccio originariamente proposto all'inizio del secolo scorso dal bibliotecario indiano Ranganathan [22], ed in modo particolare la metodologia POPSI, originariamente introdotta da Bhattacharyya [26]. Da notare che la teoria di Ranganathan è universalmente riconosciuta quale metodologia fondamentale per guidare il processo di organizzazione della conoscenza in un certo dominio (si veda ad esempio [30]) in termini di soggetti-base e relazioni tra loro.

Le faccette rappresentano differenti *facce (o aspetti) della conoscenza*. Esse formalizzano le principali caratteristiche di ciascun dominio (ad esempio medicina, sport, musica, scienze naturali), in particolare le entità che appartengono a tale dominio (ad esempio in medicina, le parti del corpo), le proprietà di tali entità (ad esempio in medicina, le varie malattie che possono affliggere le parti del corpo) e le azioni che possono essere intraprese (ad esempio in medicina, chirurgia, pervenzione o medicamenti). Più precisamente, una faccetta è una struttura gerachica tassonomica che raggruppa termini omogenei (i nodi), dove ciascun termine denota un concetto primitivo atomico, ovvero una singola unità cognitiva (ad esempio *cellula* o *virus*). Dunque, per ciascun dominio abbiamo tassonomie di entità, proprietà, azioni e così via. Noi usiamo il termine *base di conoscenza di supporto*[6] [17, 14, 21] per denotare uno *schema di rappresentazione a faccette*, ovvero un insieme di faccette che rappresenta/codifica un sistema di conoscenza a priori per i domini di interesse (si veda anche [13] per un primo tentativo di definire uno schema di rappresentazione a faccette non basato sulla teoria di Ranganathan).

Uno schema di rappresentazione a faccette supporta la post-coordinazione, ovvero la costruzione di etichette complesse (i soggetti in biblioteconomia) combinando termini da faccette differenti sia durante l'indicizzazione, la classificazione e la ricerca di documenti. Dunque, le ontologie leggere a faccette sono ontologie leggere dove le etichette, ovvero le corrispondenti formule, contengono solo concetti atomici i quali corrispondono ai concetti primitivi presi dalla base di conoscenza di supporto.

In resto dell'articolo è organizzato nel modo seguente: la sezione 2 introduce e definisce formalmente il concetto di ontologia leggera (da classificazione); la sezione 3 introduce le faccette; la sezione 4 introduce i soggetti a faccette e la nozione di ontologia leggera a

faccette; infine, la sezione 5 mostra, attraverso un esempio, come un soggetto a faccette viene costruito secondo il sistema di indicizzazione POPSI; la sezione 6 riassume e conclude l'articolo.

2 Ontologie leggere da classificazione

Le ontologie sono state usate per secoli presso comunità differenti, per scopi e con modalità differenti. Il concetto ha origine remota e si fa risalire a circa duemila anni fa dalla filosofia, e più precisamente dalla teoria metafisica delle categorie di Aristotele[7]. Lo scopo originario era lo studio dell'essere, ovvero quello di fornire una categorizzazione di tutte le cose esistenti nel mondo. Le ontologie sono state successivamente adottate in molti altri campi, quali la biblioteconomia, l'intelligenza artificiale, e più di recente l'informatica come principale mezzo per descrivere come le classi di oggetti sono tra loro correlate, o per organizzare quello che gli archivisti chiamano genericamente documenti (ossia qualunque oggetto che possa essere archiviato).

Sono state fornite molteplici definizioni di ontologie. Secondo la più citata, un ontologia è "*una esplicita e dettagliata descrizione di una concettualizzazione*" [10]. Il loro scopo principale è quello di favorire l'interoperabilità, fornendo una comune terminologia e la piena comprensione di un certo dominio di interesse, che a sua volta supporti l'assegnazione di un significato chiaro ed univoco a tutte le componenti dell'informazione. Questo viene fatto attraverso la nozione di concetto. Un concetto è un'unità cognitiva che rappresenta l'intensione, ovvero l'insieme delle proprietà che lo contraddistinguono dagli altri e riassume la sua estenzione, ovvero l'insieme degli oggetti che presentano tali proprietà. Basti pensare ad esempio alle proprietà che distinguono i rettili dai mammiferi, o i vertebrati dagli invertebrati.

È importante però sottolineare che ci sono diversi tipi di ontologie, distinguibili a seconda del grado di formalità ed espressività del linguaggio usato per descriverle (si veda [2] per una dissertazione sull'argomento). Esse spaziano da una rappresentazione del tutto informale, come le classificazioni utente (ad esempio la struttura delle cartelle in un personal computer) e le Web directories (ad esempio le ben note DMOZ, Yahoo! and Google[8]), a rappresentazioni progressivamente più formali come i tradizionali schemi di classificazione enumerativi (la Dewey Decimal Classification[9] (DDC) e la Library of Congress Classification[10] (LCC)), gli schemi di organizzazione della conoscenza quali thesauri (ad esempio AGROVOC, NALT, AOD e HBS) e schemi di classificazione a faccette (la Colon Classification) e infine le ontologie cosiddette formali le quali sono espresse in un linguaggio logico formale e sono rappresentate usando specifici formalismi ben noti nel campo del Web semantico, quali RDF e OWL.

Ai fini del nostro lavoro e seguendo la terminologia usata in [1], noi piuttosto distinguiamo fondamentalmente tra:

1. ontologie principalmente usate per descrivere oggetti, chiamate anche *ontologie descrittive*, e
2. ontologie principalmente usate per organizzare oggetti, chiamate anche *ontologie da classificazione*.

Questa distinzione si riflette nella semantica di riferimento, ovvero nella *semantica del mondo reale*[11] e la *semantica classificatoria*[12] descritte di seguito. Basandoci su tale distinzione noi poi raffiniamo il concetto di ontologia da classificazione nella nozione di

ontologia leggera da classificazione, la quale è la nozione chiave proposta in questo articolo. Analizziamo dunque queste nozioni in dettaglio.

2.1 Ontologie descrittive

Nelle ontologie descrittive, *i concetti rappresentano entità del mondo reale*, ad esempio l'estensione del concetto "animale" è l'insieme costituito da tutti gli animali del mondo reale. Essi possono essere connessi attraverso relazioni di diverso tipo. Lo scopo delle ontologie descrittive è quello di specificare i termini usati nel loro significato originario, secondo la natura e la struttura del dominio che esse modellano [11]. Due sono i principali tipi di relazioni che vengono utilizzati per costruire le tassonomie, le quali forniscono la struttura portante di queste ontologie. Esse sono le cosiddette relazioni di *genere-specie*, denotate in inglese con *is-a* (letteralmente *è un*), e quelle *mereologiche* o di *tutto-e-parte*, denotate in inglese con *part-of* (letteralmente *parte di*). In Fig. 1 (a) e (b) forniamo due esempi di ontologie descrittive, basate su questi due tipi di relazione. Ciascun nodo rappresenta un concetto e ciascuna freccia rappresenta una relazione tra concetti. La direzione della freccia rappresenta la direzione della relazione. Situazioni in cui i due tipi di relazioni sono entrambi presenti sono inoltre possibili.

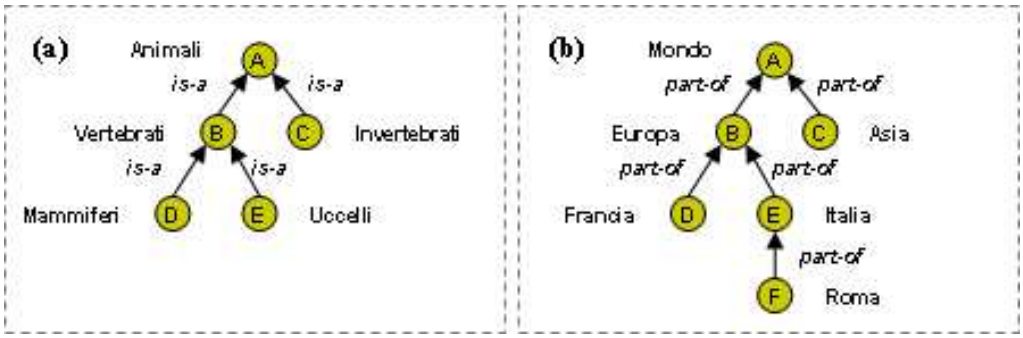

Fig. 1. (a) Una ontologia di genere-specie; (b) Una ontologia mereologica

Ci sembra utile notare che, quando queste ontologie vendono trasformate in logica descrittiva, le relazioni di tipo *is-a* vengono trasformate in sussunzione (denotata col simbolo $\sqsubseteq$), o più precisamente si assume che implichi sussunzione, mentre questo non può essere fatto per le *part-of*. In altre parole, mentre è naturale poter dire che tutti i mammiferi sono una fattispecie di animali, non ha senso dire che l'Italia è una fattispecie di Europa. Di conseguenza, la relazione di *is-a* costituisce l'ossatura, ovvero la struttura portante, della gerarchica di un dominio basata su sussunzione.

2.2 Ontologie da classificazione

Le ontologie in semantica classificatoria sono costruite con lo scopo di indicizzare documenti. Di sconseguenza, *l'estensione di ciascun concetto (espresso dall'etichetta del nodo) è l'insieme dei documenti circa le entità o individui descritti dall'etichetta del concetto* [1, 2]. Ad esempio, l'estensione del concetto "animale" è l'insieme dei documenti il cui soggetto sono gli animali (di qualunque tipo, del presente e del futuro, vivi o morti). Questo ha tre principali conseguenze.

La prima è che la relazione semantica che sussiste tra i nodi che si trovano l'uno sotto l'altro è *sempre* quella di *sottoinsieme*. In altre parole, l'insieme di documenti che può essere

classificato in un certo nodo è sempre un sottoinsieme dei documenti che possono essere classificati nel nodo sopra di esso. Questo motiva alcune tecniche per la minimizzazione del numero dei nodi in cui un certo documento viene classificato. Basti pensare, ad esempio, al ben noto *principio di gradazione della specificità*[13], secondo il quale è opportuno cercare di classificare un documento nella locazione più specifica per esso (e quindi più in profondità possibile nella classificazione). Rimandiamo alla lettura di [16] per una formalizzazione di tale principio ed il suo uso per la classificazione automatica di documenti.

Le Fig. 2 (a) e (b) forniscono la versione in semantica classificatoria delle due ontologie riportate in Fig. 1 (a) e (b). Come può essere notato dalla Fig. 2, le relazioni delle ontologie descrittive sono convertite in relazioni tra insiemi. Dunque le relazioni di tipo *is-a*, ma anche quelle di *part-of* quando transitive, e quelle dette di *instance-of* (la quale denota un individuo di una specifica classe, ad esempio Roma in qualità di instanza della classe città), vengono convertite in relazioni di sottoinsieme, mentre tutte le altre relazioni corrispondono alla *sovrapposizione* di insiemi (denotata col simbolo ⊓). La Fig. 2 (b) mostra un caso dove le relazioni di *part-of* della Fig. 1 (b) sono convertite in relazioni di sottoinsieme secondo la semantica classificatoria. Un esempio in cui questo non è possibile è la catena di relazioni: maniglia *part-of* porta *part-of* scuola *part-of* sistema scolastico. Infatti, qui le relazioni di *part-of* sono di tipi diversi. Si veda ad esempio [33].

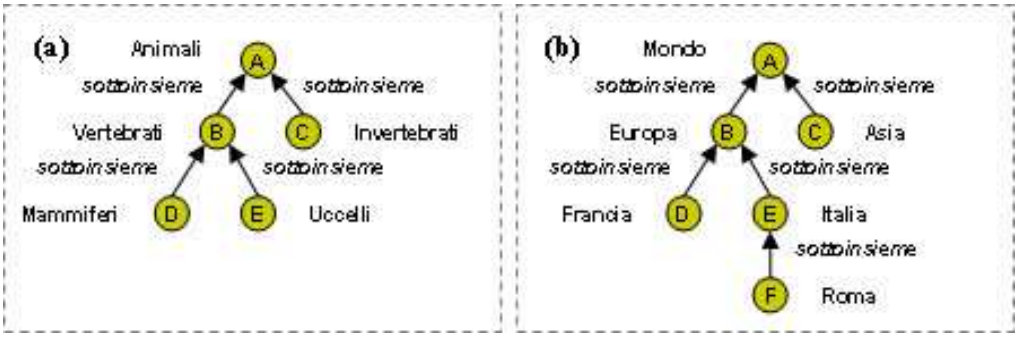

Fig. 2. Due ontologie in semantica classificatoria

La seconda conseguenza è che, con la trasformazione in logica descrittiva delle ontologie da classificazione, la relazione di sottoinsieme viene tradotta in sussunzione tra le formule dei nodi che sono uno sotto l'altro. È importante osservare che la trasformazione della stessa ontologia, a seconda se interpretata con la semantica del mondo reale o con quella classificatoria, porta a due differenti teorie (si confrontino ancora una volta le Fig. 1(b) and Fig. 2(b)).

Nota che le etichette nelle due ontologie in Fig. 2 sono tali che ciascuna rappresenta un sottoinsieme proprio dell'etichetta del nodo sopra. Ad esempio, *Vertebrati* rappresenta un sottoinsieme proprio di *Aminali*. C'è da dire però che, e questa è la terza conseguenza, nelle ontologie classificatorie la situazione di cui sopra può essere generalizzata al caso in cui le etichette che denotano gli insiemi non siano in relazione di sottoinsieme, ma piuttosto in relazione di sovrapposizione. D'altra parte, questo è ciò che normalmente avviene nella maggior parte delle ontologie da classificazione [2].

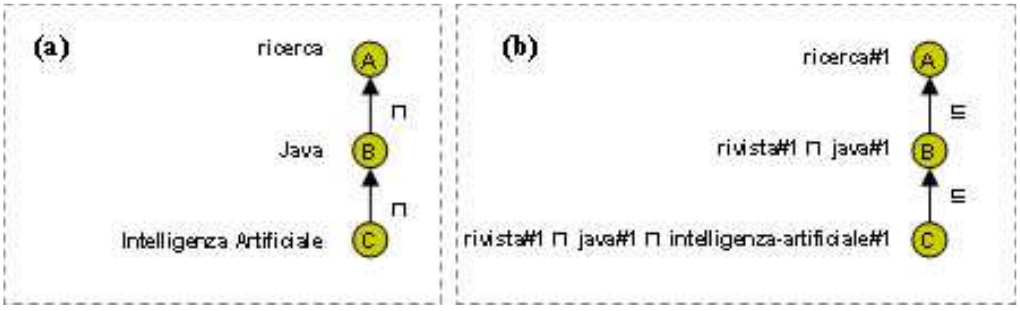

Fig. 3. (a) Una ontologia da classificazione in cui non c'è relazione di sottoinsieme tra le etichette, (b) la corrispondente ontologia formale.

Prendiamo ad esempio l'ontologia da classificazione in Fig. 3 (a). L'intuizione è che il nodo B debba contenere tutti i documenti circa "*ricerca in Java*". In altre parole, il significato del nodo (chiamato *concetto al nodo* in [1, 14]) può essere costruito prendendo la *congiunzione* in logica descrittiva (denotata con ⊓ e semanticamente corrispondente all'intersezione dei due insiemi) dei concetti di tutte le etichette lungo il percorso dalla radice della classificazione al nodo stesso. L'applicazione di questa regola all'esempio in Fig. 3 (a) produce l'ontologia mostrata in Fig. 3 (b). Come può essere notato, il concetto associato a ciascun nodo è in relazione di sussunzione con ciascuno dei nodi sopra di esso e questo è ottenuto applicando l'operatore logico di congiunzione lungo il percorso. I numeri indicati accanto a ciascuna etichetta denotano il concetto ottenuto dall'applicazione del processo di disambiguazione semantica. Infatti, ciascuna parola può corrispondere a più di un concetto, ad esempio Java può denotare un linguaggio di programmazione (#1), un'isola (#2), o un tipo di caffè (#3). Non è difficile notare come la situazione in Fig. 3 (a) collassi nella situazione in Fig. 2 (a) una volta ritornati alla relazione di sottoinsieme. Infatti, tutte le congiunzioni diventano ridondanti dovute al fatto che se B ⊑ A (B sussume A) allora A ⊓ B è equivalente ad A.

2.3 Ontologie leggere da classificazione

Quello che sosteniamo è che tutta la teoria sulla classificazione in biblioteconomia - e di conseguenza la teoria delle faccette, così come originariamente concepita da Ranganathan e successivamente raffinata nella metodologia POPSI – è basata sulla semantica classificatoria. E ciò è assolutamente corretto, poichè queste metodologie vennero inventate al fine di classificare libri e posizionarli su scaffali. Dunque, nel resto dell'articolo anche noi ci concentriamo sulle ontologie da classificazione e la semantica classificatoria. La motivazione è piuttosto simile a quella della biblioteconomia. D'altronde, è un dato di fatto che strumenti quali, ad esempio, i cataloghi online, i file system sui PC, le Web directory e le classificazioni biblioteconomiche vengono tutti utilizzati per classificare oggetti. Queste classificazioni possono essere tutte convertite, in modo preciso o con un certo grado di approssimazione, in ontologie classificatorie.

D'altra parte, in tutte le applicazioni di biblioteconomia, ma anche nelle nostre applicazioni di riferimento, le ontologie da classificazione che sono richieste sono piuttosto semplici e consistono di strutture ad albero, eventualmente con radici multiple, dove la maggior parte dei nodi in relazione di padre-figlio *non hanno* alcuna esplicita etichetta la quale denoti una relazione di sottoinsieme. L'etichetta di ciascun nodo può dunque essere convertita in una formula logica (tipicamente costruita come combinazione di congiunzioni

e, più raramente, disgiunzioni di concetti atomici) la quale rappresenta il significato del nodo tenuto conto del suo contesto, ovvero il percorso dalla radice al nodo stesso [3]. Questo ci porta alla definizione di ontologia leggera da classificazione, come originariamente definita in [25] (la parola "classificazione" non compare nella definizione originaria):

*Un'**ontologia leggera da classificazione** O è un albero con radice <N,E,L^F> dove:*

a) N è un insieme finito di nodi;

b) E è un insieme di archi su N;

c) L^F è un insieme finito di etichette esperesse in logica descrittiva proposizionale tale che per ciascun nodo $n_i \in N$, esiste una ed una sola etichetta $l_i^F \in L^F$;

d) $l_{i+1}{}^F \sqsubseteq l_i^F$ con n_i padre di n_{i+1}.

3 Faccette

Secondo l'approccio Analitico-Sintetico [14], le faccette vengono definite in due passi successivi:

1. Esamina il campo (dominio) per identificare i termini rilevanti. Questi possono essere identificati consultando esperti del dominio ed ogni sorta di sorgente di informazioni relativa al dominio. Questo processo comincia nel cosiddetto *piano delle idee*, il livello concettuale indipendente dallo specifico linguaggio usato, dove i concetti primitivi vengono identificati. Ciascun concetto così identificato viene espresso nel cosiddetto *piano verbale* in un dato linguaggio, per esempio in italiano, cercando di articolare l'idea in modo coestensivo (avente stessa estensione), ovvero identificando un termine che in maniera esatta e non ambigua esprima il concetto;
2. Raggruppa i termini identificati (anche chiamati *singole idee*) secondo le loro proprietà o caratteristiche comuni, e ordinali (in gerarchie) in una sequenza significativa. Un insieme di termini omogenei forma una *faccetta*. Ad esempio, *Naso*, *Laringe*, *Trachea*, *Bronchi*, *Polmone*, *Sacco Pleurale* e *Parete Mediale* potrebbero formare una faccetta chiamata *Sistema Respiratorio* (queste entità sono in relazione di *part-of*). A loro volta, i termini *Naso Esterno* e *Osso Nasale*, i quali rappresentano parti di *Naso*, potrebbero formare una faccetta chiamata *Naso* la quale verrà trattata come sotto-faccetta della faccetta *Sistema Respiratorio*.

Questi due passi portano alla costruzione di uno *schema di rappresentazione a faccette* e corrisponondono a quello che nella nostra precedente ricerca chiamavamo la definizione e costruzione della *base di conoscenza di supporto* [17, 14, 21], ovvero una conoscenza a priori che deve esistere per poter rendere efficace la semantica.

Nota che le faccette create al passo 2 seguono la semantica del mondo reale, ovvero, sono ontologie descrittive formate utilizzando relazioni di *part-of*, *is-a* ed *instance-of* . Le faccette hanno in particolare le seguenti proprietà:

1. Esse sono organizzate in insiemi di domini indipendenti i quali sono completamente modulari e possono essere sviluppati indipendentemente;
2. Per ciascun dominio, le faccette sono raggruppate in categorie elementari ben definite. In origine, Ranganathan postulò cinque categorie fondamentali: *Personalità*, *Materia*, *Energia*, *Spazio* e *Tempo* (PMEST). In seguito, Bhattacharyya propose un raffinamento che consiste in quattro categorie principali, chiamate sinteticamente

DEPA: *Disciplina* (D) (quello che noi chiamiamo dominio), *Entità* (E), *Proprietà* (P) e *Azione* (A), più una speciale categoria, chiamata *Modificatore* (m).

Nel nostro approaccio noi organizziamo le faccette secondo le categorie DEPA, descritte qui di seguito:

- **Disciplina** (o **dominio**): essa include le branche di studio tradizionali (ad esempio, *Biblioteconomia*, *Matematica* e *Fisica*), applicazioni delle tradizionali discipline pure (ad esempio *Ingegneria* ed *Agricoltura*), qualunque aggregato di tali campi (ad esempio *Scienze fisiche* e *Scienze sociali*) o anche, in termini più moderni, campi quali la *Musica*, lo *Sport*, l'*Informatica* e così via.
- **Entità:** questa categoria elementare si manifesta in entità fisiche o astratte. Si distingue nettamente dalle proprietà ed azioni effettuate da esse o su esse. In pratica, essa contiene i concetti che rappresentano le idee chiave che sono alla base del dominio trattato. Ad esempio, *Insegnanti*, *Studenti* e *Corsi* sono i concetti chiave del dominio *Istruzione*.
- **Proprietà:** essa include i concetti che denotano attributi quantitativi e qualitativi delle entità. Ad esempio, Qualità, Quantità, Misura, Peso, ecc.
- **Azione**: questa categoria include i concetti che denotano la nozione di "*fare*". Include cioè i "*processi*" e le "*fasi*" del fare. Una azione si può manifestare in qualità di "*azione propria*" oppure "*azione esterna*". Quelle del primo tipo sono azioni fatte da un qualche agente (esplicito o implicito) su o in se stesso. Ad esempio, *Immaginazione*, *Interazione*, *Reazione*, *Ragionamento*, *Pensiero*, ecc. Quelle del secondo tipo invece sono azioni fatte da un agente (esplicito o implicito) su un concetto di una qualunque delle categorie elementari discusse sopra. Ad esempio, *Organizzazione*, *Cooperazione*, *Classificazione*, *Catalogazione*, *Calcolo*, *Progettazione* ecc.
- **Modificatore**: questa categoria include i concetti usati oppure intesi per qualificare un altro concetto. Con l'aiuto di un modificatore, l'estensione di un concetto viene decrementata e l'intensione aumentata senza alterare la propria integrità concettuale. Per essere concreti, si consideri ad esempio il concetto "Estrazione mineraria in India". Qui il termine "India" è appositamente utilizzato al fine di modificare il concetto di "Estrazione Mineraria" restringendo l'estensione alla sola zona geografica indiana. Ciò implica che, in linea di principio, qualunque concetto o combinazione di due o più concetti elementari appartenenti ad una qualsiasi delle categorie elementari di sui sopra può fungere da base per derivare un modificatore. Esistono diverse tipologie di modificatori. In particolare, si usa distinguere tra *modificatori universali*[14] (ad esempio *modificatore spaziale*, *modificatore temporale*, *modificatore ambientale*, *modificatore di forma*, *modificatore di linguaggio*) e *modificatori speciali* (ad esempio, *Infettiva*, *Batterica*, *Virale*, ecc … sono modificatori per il concetto di *Malattia* nel dominio *Medicina*). I modificatori universali sono comuni a tutte le discipline e sono usati per modificare le manifestazioni (ovvero i singoli termini che compaiono in un soggetto) di uno o più categorie elementari, sia che occorrano singolarmente che in combinazione. I modificatori speciali modificano le manifestazioni di una o più categorie elementari. Tuttavia, seguendo il principio di riuso (descritto in seguito), alcuni modificatori possono essere condivisi da un insieme (ma non tutti) di domini (ad esempio, le

sostanze chimiche sono usate sia in *Chimica* che in *Agricoltura*, eventualmente sotto diverse categorie).

La regola base per formare un soggetto ai fini dell'indicizzazione è quello di porre prima la *Disciplina* (la base), seguita dalla *Entità* (il concetto chiave), a sua volta seguita dalla *Proprietà* e/o dall'*Azione*. Le ultime due possono a loro volta essere seguite da altre *Proprietà* e/o dall'*Azioni* ed eventualmente dei *modificatori universali.* Le *specie*/*tipi* e/o i *modificatori* e/o le *parti* e/o i *costituenti* di ciascuna delle categorie elementari seguono immediatamente la manifestazione per la quale sono rispettivamente *specie*/*tipi* oppure *modificatori* oppure *parti* oppure *costituenti*. Ad esempio, un possibile soggetto potrebbe essere:

Medicina : Fegato : Infezione : Batterica : Prevenzione

In **Fig. 4** viene riportato un esempio di faccette raggruppate secondo la categorizzazione DEPA per il dominio *Medicina*. Nota che, anche se non è il caso nell'esempio specifico, in ciascuna categoria è possibile, in linea di principio, avere una o più faccette.

Le faccette possiedono alcune proprietà essenziali, elencate di seguito:

- **Ospitalità**: le faccette sono facilmente estendibili. Nuovi termini, i quali rappresentano nuova conoscenza, possono essere inseriti senza alcuna difficoltà nella struttura gerarchica. I termini nelle tassonomie sono definiti in maniera chiara, sono mutulamente esclusivi e collettivamente esaustivi.
- **Compattezza**: i sistemi basati sulle faccette richiedono meno spazio rispetto agli altri sistemi di indicizzazione. Essi permettono di evitare l'esplosione di soggetti dovuta alle possibili combinazioni dei soggetti-base (contenuti nelle faccette). In altre parole, le faccette supportano l'approccio post-coordinato in contrasto rispetto a quello puramente enumerativo.
- **Flessibilità**: i sistemi per l'organizzazione gerarchica della conoscenza sono per lo più rigidi nella loro struttura, mentre i sistemi basati su faccette sono flessibili per natura.
- **Riusabilità**: una ontologia basata su faccette sviluppata per uno specifico dominio può essere parzialmente utilizzata nel contesto di un altro dominio correlato.
- **Chiara, ma rigorosa, struttura**: l'approccio basato sulle faccette mira all'identificazione delle relazioni logiche tra i concetti ed i gruppi di concetti. I concetti sullo stesso livello della gerarchia condividono una comune caratteristica o proprietà.
- **Uso della metodologia**: viene proposta una robusta metodologia per l'analisi e la categorizzazione dei concetti, accompagnata da una serie di regole affidabili per la sintesi.
- **Omogeneità**: una faccetta rappresenta un gruppo omogeneo di concetti, secondo la o le caratteristiche comuni specificate.

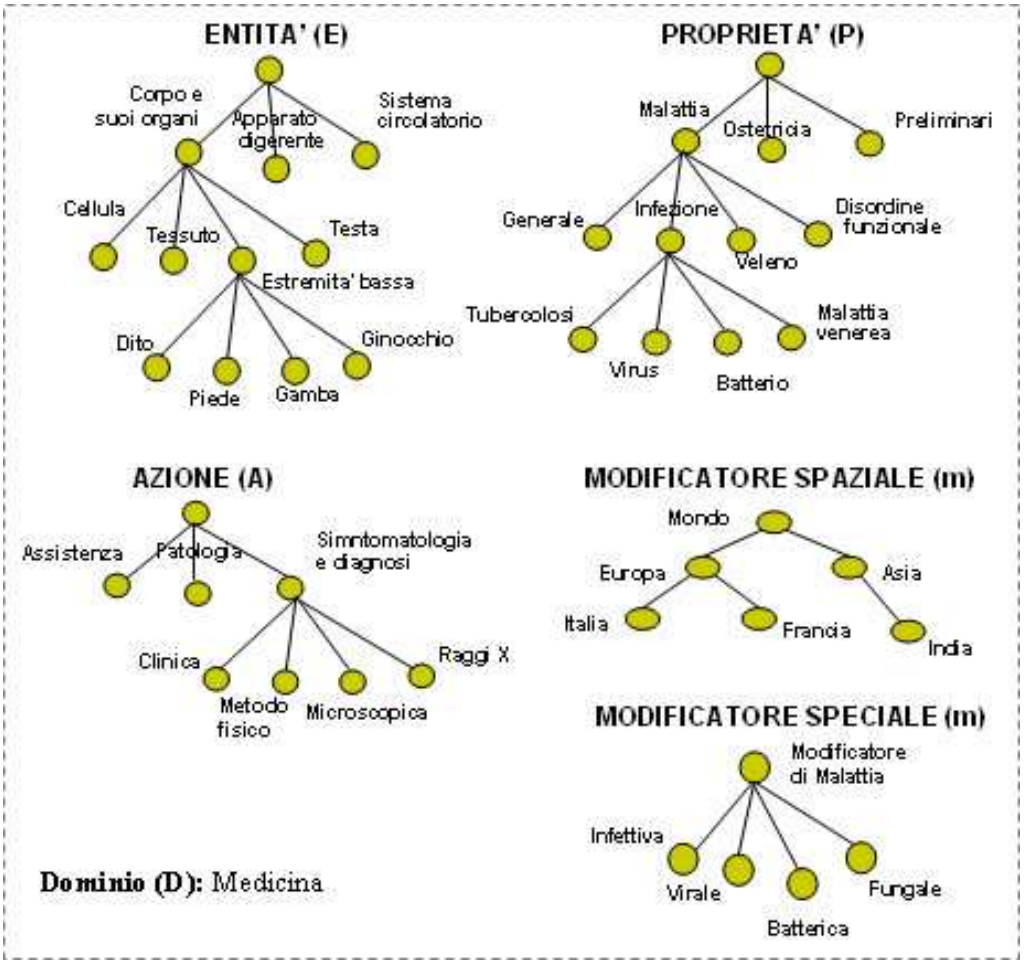

Fig. 4. L'insieme di faccette per il dominio *Medicina*

4 Ontologie leggere a faccette

Una volta costruita la base di conoscenza di supporto, il passo successivo è quello di utilizzare le faccette per indicizzare o classificare documenti (nel nostro caso in ontologie leggere). Come spiegato precedentemente, questo per noi corrisponde all'associare a ciascun documento e nodo della classificazione una formula in logica descrittiva [1, 2]. Questo passo avviene in quello che Ranganathan ha chiamato il *piano notazionale*. In tale piano una notazione non ambigua viene usata per associare, in maniera sintetica, il significato semantico e porre ordine tra gli oggetti gestiti, tipicamente libri sugli scaffali. Seguendo Bhattacharyya [26], l'idea chiave è quella di associare a ciascun nodo o documento un *soggetto*, ovvero "*un pezzo di informazione non discorsiva che sintetizza indicativamente di cosa parla un libro oppure un documento (ovvero un qualsiasi oggetto contenente informazione)*". Un soggetto, nella nostra accettazione, è una etichetta e corrisponde al concetto associato al documento o al nodo nella ontologia leggera. Poichè nelle ontologie leggere noi usiamo la semantica classificatoria, un documento verrà classificato in tutti i nodi (tenendo presente il principio di gradazione di specificità) il cui soggetto è più generale del soggetto del documento [1, 16].

Noi definiamo i soggetti in termini di faccette. Le intuizioni chiave sono tre:

1. Noi associamo a ciascun termine del soggetto un'etichetta ed il concetto corrispondente preso dallo schema di classificazione a faccette (nota che in POPSI il concetto è dato dal termine standard e dal suo contesto);
2. Per ciascun termine nella faccetta, il contesto è costruito associando ad esso tutti i termini lungo il percorso dalla radice della faccetta al termine in questione, dunque disambiguando il concetto inteso. Questo significa che, nel passaggio dalla base di conoscenza di supporto al concetto del soggetto, occorre tradurre dalla semantica del modo reale (usata nella base di conoscenza di supporto) alla semantica classificatoria (usata nelle ontologie leggere).
3. Ciascun soggetto contiene termini (dunque concetti) da potenzialmente tutte le categorie DEPA, dunque permettendo una completa disambiguazione del soggetto. Tuttavia, si suppone che l'utente fornisca, esplicitamente o implicitamente, almeno la *disciplina* e l'*entità* principale.

In POPSI, per poter costruire il contesto, ciascun *termine standard o termine guida*[15] è seguito dalla *dicitura di contesto*, ovvero dall'insieme di termini ausiliari che preservano il contesto (in termini di disciplina e l'intero percorso dalla radice della faccetta al termine stesso). Ad esempio, il contesto per il termine *Cellula* è:

Cellula (termine guida)
Medicina, Corpo e suoi organi > Cellula (dicitura di contesto)

In questo esempio, "," separa quelle che vengono chiamate le *singole idee*[16] (ovvero i concetti) corrispondenti alle differenti categorie fondamentali, mentre ">" indica l'incremento dell'intensione ed il decremento dell'estensione della singola idea all'interno della faccetta, ovvero il passaggio da un termine generico ad uno più specifico. Nota che, come da Fig. 4, Medicina è il nome del dominio, mentre la seconda parte è il percorso completo nella faccetta delle entità. Si consideri inoltre il soggetto "*Diagnosi microscopica su cellule colpite da virus batterici in India*". I termini contenuti sono completamente contestualizzati in POPSI nel modo seguente (la sequenza dei passi concreti necessari alla loro identificazione è descritta nella sezione successiva):

(Dominio):	Medicina,
(Entità):	Corpo e suoi organi > Cellula,
(Proprietà):	Malattia > Infezione > Virus,
(Modificatore di P.)	Batterica,
(Azione):	Sintomatologia e diagnosi > Microscopica,
(Modificatore spaziale):	Asia > India

Il principale vantaggio dell'approccio basato su faccette è che esso rende esplicite le relazioni semantiche tra concetti e gruppi di concetti e rimuove le limitazioni delle gerarchie tradizionali. Esso permette di vedere una voce (dell'indice) complessa da diverse prospettive o diverse angolazioni. Ad esempio, una *mucca* può essere descritta in quanto animale, in quanto animale domestico, in quanto cibo, in quanto merce, come Dio presso alcune comunità particolari (ad esempio in India), e così via a seconda del dominio. Di

conseguenza, ogni volta, fornendo il contesto, l'approccio basato sulle faccette permette la rappresentazione di concetti differenti.

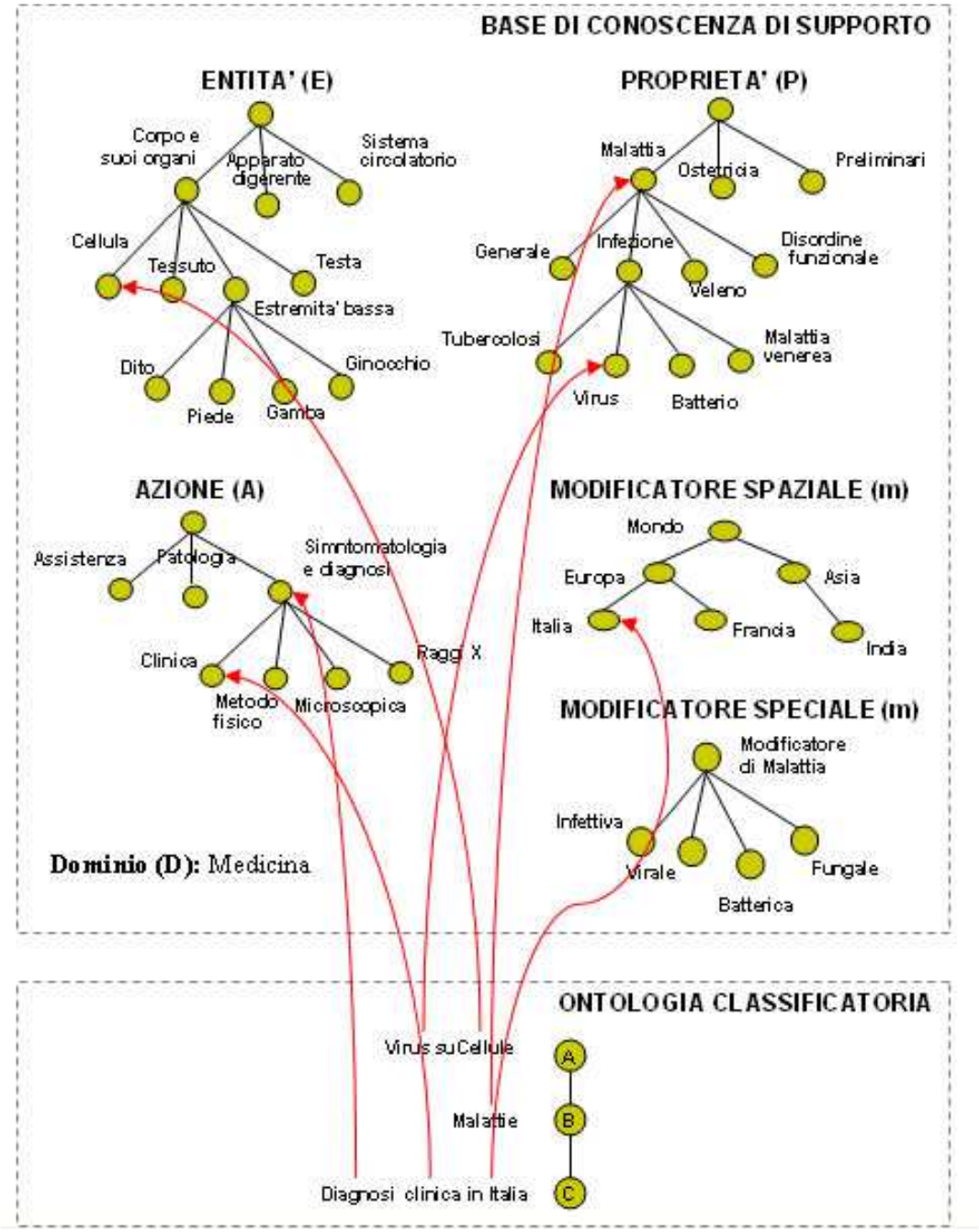

Fig. 5. Una ontologia leggera a faccette.

Basandoci sulla nozione di soggetto, possiamo ora definire un'*ontologia (classificatoria) leggera a faccette*:

*Un'**ontologia (classificatoria) leggera a faccette** è un'ontologia leggera dove ciascun termine e corrispondente concetto che occorre nelle etichette dei suoi nodi devono*

corrispondere ad un termine e corrispondente concetto nella base di conoscenza di supporto, modellata come schema di classificazione a faccette.

In Fig. 5 è fornito un esempio di come questo possa essere fatto. Nota che in una ontologia leggera a faccette vi possono essere nodi, come nell'esempio, le cui etichette contengono termini da più categorie DEPA. Più termini e corrispondenti categorie DEPA avremo, più l'ontologia leggera sarà specifica.

5 Indicizzazione dei soggetti

Come dunque usare in pratica gli schemi di classificazione a faccette? Come menzionato nella sezione precedente, i documenti vengono classificati sotto tutti quei nodi il cui soggetto è più generale di quello dei documenti stessi. Ma la vera sfida è che in molti casi il soggetto è solo parzialmente specificato. A tal proposito, POPSI fornisce una metodologia per l'identificazione delle *informazioni contestuali mancanti*. La soluzione risiede principalmente in una appropriata rappresentazione dell'estensione del soggetto dei documenti indicizzati.

Di seguito discutiamo i singoli passi della metodologia POPSI per derivare i soggetti partendo dai titoli (ovvero i soggetti) associati ai documenti da indicizzare. Si consideri il seguente esempio di soggetto, dato nella precedente sezione:

"*Diagnosi microscopica su cellule colpite da virus batterici in India*".

L'analisi è organizzata in otto passi successivi, come descritto di seguito:

Passo 1 (Analisi dell'espressione indicativa del soggetto): esso consiste nell'analisi dell'espressione indicativa del soggetto relativa alla sorgente d'informazione. Questi può essere il titolo del libro, dell'articolo e così via. Dall'esempio di cui sopra deriviamo i seguenti termini:

D = Medicina (implicito nel titolo di cui sopra)
E = Cellule (esplicito)
P = Virus (esplicito)
m di P = Batterici (esplicito)
A = Diagnosi microscopica (esplicito)
m = India (esplicito) (Modificatore spaziale)

Nel nostro approaccio, questo passo è eseguito analogamente. Nota che le categorie implicite devono essere manualmente fornite dall'utente o calcolate automaticamente dal sistema.

Passo 2 (Formalizzazione del soggetto): in questa fase viene formalizzata la sequenza dei termini che compaiono nel soggetto derivato al passo 1 (Analisi). Secondo i principi di sequenza, le componenti vengono indicate in successione nel modo seguente:

Medicina (D), Cellule (E), Virus (P), Batterici (m di P), Diagnosi microscopica (A), India (m)

Nel nostro approaccio questo non è richiesto.

Passo 3 (Standardizzazione del soggetto): Esso consiste nell'identificazione dei termini standard, quando ad esempio sono disponibili sinonimi per lo stesso termine, che denotano i concetti atomici presenti nel soggetto. Nel nostro esempio si applica solo la riduzione alla forma singolare:

> Medicina, Cellula, Virus, Batterico, Microscopica, India

Nel nostro approaccio, questo passo è eseguito analogamente. Queste informazioni sono codificate nella base di conoscenza di supporto.

Passo 4 (Modulazione del soggetto): Esso consiste nell'aggiungere al soggetto standardizzato i concetti padre di ciaascun concetto, utilizzando i termini standard con indicazione degli eventuali sinonimi. In pratica, esso corrisponde all'identificazione dei corrispondenti termini contestuali, ovvero la corretta disambiguazione di ciascun termine usato, fornendo la giusta quantità di concetti gerarchicamente correlati:

> Medicina, Corpo e suoi organi > Cellula, Malattia > Infezione > Virus, Batterico, Sintomatologia e diagnosi > Microscopica, Asia > India

Nel nostro approaccio, questo passo è eseguito analogamente: noi estraiamo dalla base di conoscenza di supporto il concetto corrispondente a ciascuno dei termini in linguaggio naturale che compaiono nel soggetto.

Passo 5 (Preparazione della voce dell'indice per la classificazione): Questo passo consiste nel preparare le principali voci del cosiddetto indice associativo, secondo un ordine alfabetico. Come descritto in [26], questo viene fatto assegnando un insieme sistematico di numeri che indicano le categorie e la posizione del soggetto. In pratica, tali numeri codificano il significato semantico del soggetto. Nel nostro esempio:

> Medicina 8 Corpo e suoi organi 8.3 Cellula 8.2 Malattia 8.2.4 Infezione 8.2.4.4 Virus 8.2.4.4.6 Batterico 8.2.1 Sintomatologia e diagnosi 8.2.1.4 Microscopica 4 Asia 4.4 India

Nel nostro approaccio questo non è necessario.

Passo 6 (Decisione circa i termini da usare come voci dell'indice): Esso consiste nello stabilire i termini che verranno usati come voci standard dell'indice al fine di generare gli indici associativi, e per il controllo dei sinonimi. A tal fine, ciascun termine standard deve poter essere identificato a partire da ciascuno dei suoi sinonimi. Ad esempio (questo non è parte dell'esempio che stiamo utilizzando):

> Trattamento chimico (Medicina)
> *vedi* Chemioterapia

Nel nostro approaccio questo non è necessario.

Passo 7 (Preparazione delle voci per l'indice associativo): Esso consiste nel preparare le voci sotto ciascun termine standard attraverso una permutazione ciclica. Ad esempio (tutte le voci possono essere trattate in modo similare):

Corpo e suoi organi
Medicina, Corpo e suoi organi > Cellula, Malattia > Infezione > Virus, Batterico, Sintomatologia e diagnosi > Microscopica, Asia > India

Cellula
Medicina, Corpo e suoi organi > Cellula, Malattia > Infezione > Virus, Batterico, Sintomatologia e diagnosi > Microscopica, Asia > India

Nel nostro approaccio questo non è necessario.

Passo 8 (Ordinamento alfabetico delle voci): esso consiste nell'ordinare tutte le voci presenti nell'indice in ordine alfabetico, secondo delle regole standard che ignorano segni di punteggiatura e simboli:

Asia
Medicina, Corpo e suoi organi > Cellula, Malattia > Infezione > Virus, Batterico, Sintomatologia e diagnosi > Microscopica, Asia > India

Batterico
Medicina, Corpo e suoi organi > Cellula, Malattia > Infezione > Virus, Batterico, Sintomatologia e diagnosi > Microscopica, Asia > India

…

Virus
Medicina, Corpo e suoi organi > Cellula, Malattia > Infezione > Virus, Batterico, Sintomatologia e diagnosi > Microscopica, Asia > India

Nel nostro approccio questo corrisponde all'indicizzazione o classificazione in una ontologia leggera a faccette utilizzando i concetti dei nodi e dei documenti.

6 Conclusioni

In questo articolo, noi abbiamo introdotto la nozione di ontologia leggera a faccette quale ontologia leggera I cui termini sono estratti da una base di conoscenza di supporto organizzata secondo uno schema a faccette. L'utilizzo delle faccette permette un maggiore controllo sul linguaggio ed i concetti usati per costruire le ontologie, ed anche sulla loro organizzazione, che in generale trae vantaggio dalla struttura e dai termini presenti nelle quattro ategorie fondamentali DEPA.

Note

1 Questo articolo e' una traduzione e rielaborazione di [34].
2 In Italia esiste una ben nota trasmissione televisiva chiamata *Mela Verde*, percui non e' detto che il significato che l'utente ha in mente sia effettivamente quello del frutto.
3 In questo caso senza alcuna ambiguita' possiamo concludere che si intende il frutto.
4 Il termine inglese e' *facet* (letteralmente sfaccettatura, aspetto), metaforicamente indicante i diversi lati o aspetti della conoscenza.
5 http://flamenco.berkeley.edu

6 Il termine in inglese e' *background knowledge*.
7 http://plato.stanford.edu/entries/aristotle-categories/
8 http://dmoz.org/; http://dir.yahoo.com/; http://directory.google.com/
9 http://www.oclc.org/dewey/
10 http://www.loc.gov
11 In inglese *real world semantics.*
12 In inglese *classification semantics*.
13 In inglese *get-specific*.
14 In inglese *common modifiers*.
15 In Inglese *leading heading* (o *lead term* oppure *term-of approach*)
16 In Inglese *isolate ideas.*

Citazioni

1. F. Giunchiglia, M. Marchese, I. Zaihrayeu, *Encoding Classifications into Lightweight Ontologies*. Journal of Data Semantics 8, pp. 57-81, 2006. Short version in: Proceedings of the 3rd European Semantic Web Conference (ESWC), 2006,
2. F. Giunchiglia, I. Zaihrayeu, *Lightweight ontologies.* In S. LNCS, editor, Encyclopedia of Database Systems, 2008.
3. I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, and X. Huang, *From web directories to ontologies: Natural language processing challenges*. In the 6th International Semantic Web Conference (ISWC 2007), 2007.
4. J.-E. Mai, *Classification in Context: Relativity, Reality, and Representation.* Knowledge Organization. 31(1), 39-48, 2004.
5. E. Duval, W. Hodgins, S. Sutton, S, L.Weibel, *Metadata Principles and Practicalities.* DLib Magazine, 8(4), 2002.
6. D. Nicholson, S. Neill, S. Currier, L. Will, A. Gilchrist, R. Russell, M. Day, *HILT: High Level Thesaurus Project – Final Report to RSLP & JISC.* Centre for Digital Library Research, Glasgow, UK, 2001.
7. Y. Tzitzikas, N. Armenatzoglou, P. Papadakos, *FleXplorer: A Framework for Providing Faceted and Dynamic Taxonomy-Based Information Exploration.* DEXA Workshops: pp. 392-396, 2008.
8. Y. Tzitzikas, A. Analyti, N. Spyratos, P. Constantopoulos, *An algebra for specifying valid compound terms in faceted taxonomies*. Data Knowl. Eng. (DKE) 62(1):1-40, 2007.
9. Y. Tzitzikas, N. Spyratos, P. Constantopoulos, A. Analyti, *Extended Faceted Ontologies*. CaiSE: pp.778-781, 2002.
10. T. R. Gruber. *A translation approach to portable ontology specifications.* Knowledge Aquisition, 5(2): pp.199–220, 1993.
11. N. Guarino. *Helping people (and machines) understanding each other*: The role of formal ontology. In CoopIS/DOA/ODBASE, p. 599, 2004.
12. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. F. Patel-Schneider, *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, 2002.
13. D. Soergel, *A Universal Source Thesaurus as a Classification Generator*. Journal of the American Society for Information Science 23(5), pp. 299–305, 1972.
14. F. Giunchiglia, M. Yatskevich, P. Shvaiko, *Semantic Matching: algorithms and implementation.* Journal on Data Semantics, IX, 2007.

15. F. Giunchiglia, F. McNeill, M. Yatskevich, J. Pane, P. Besana, P. Shvaiko, *Approximate Structure-Preserving Semantic Matching*. In the 7th International Conference on Ontologies, Databases and Applications of Semantics (ODBASE), 2008.
16. F. Giunchiglia, I. Zaihrayeu, U. Kharkevich, *Formalizing the get-specific document classification algorithm.* In the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), 2007. LNCS Springer Verlag.
17. F. Giunchiglia, P. Shvaiko, M. Yatskevich, *Discovering Missing Background Knowledge in Ontology Matching*. In: vol.141, Leipzig University, Germany: Brewka et al., 2006. p. 382-386. Proceedings of the 17th European Conference on Artificial Intelligence - ECAI 2006.
18. F. Giunchiglia, P. Shvaiko, M. Yatskevich, Semantic schema matching. In On the move to meaningful internet systems 2005: COOPIS, DOA, and ODBASE: OTM Confederated International Conferences, vol. 1. Proceedings: CoopIS, DOA, and ODBASE, Agia Napa, Cyprus, 2005 Editors: Meersman R., Tari Z., Berlin Heidelberg: Springer, LNCS, Vol. 3760/2005, p. 347-365, 2005.
19. F. Giunchiglia, M. Yatskevich, E. Giunchiglia, *Efficient semantic matching*. In: Proceedings of the 2nd European semantic web conference (ESWC'05). Editors: Gomez-perez A., Euzenat J., Heidelberg: Springer, 2005. Lecture Notes in Computer Science, Vol. 3532/2005, p. 272-289, Proceedings: Second European Semantic Web Conference, ESWC, Heraklion, Crete, Greece, 29 May - 1 June 2005, Note: ISBN: 3-540-26124-9
20. F. Giunchiglia, M. Yatskevich, *Element Level Semantic Matching*. Workshop on Meaning Coordination and Negotiation. ISWC04, Hiroshima, Japan, November 2004
21. F. Giunchiglia, U. Kharkevich, I. Zaihrayeu, *Concept Search: Semantics Enabled Syntactic Search*. In Semantic Search (SemSearch) 2008 workshop at the 5th European Semantic Web Conference (ESWC), 2008.
22. S. R. Ranganathan, *The Colon Classification*. In S. Artandi, editor, Vol IV of the Rutgers Series on Systems for the Intellectual Organization of Information, 1965. New Brunswick, NJ: Graduate School of Library Science, Rutgers University.
23. K. La Barre, *Adventures in faceted classification: A brave new world or a world of confusion?* Knowledge organization and the global information society: proceedings 8th ISKO conference, London, 13-16 July 2004.
24. M. Hearst, *Design Recommendations for Hierarchical Faceted Search Interfaces.* In ACM SIGIR Workshop on Faceted Search, Seattle, WA, 2006.
25. F. Giunchiglia, V. Maltese, A. Autayeu, *Computing minimal mappings*. 4th Ontology Matching Workshop at the ISWC 2009.
26. G. Bhattacharyya, *POPSI: its fundamentals and procedure based on a general theory of subject indexing languages*, Library Science with a Slant to Documentation, Vol. 16 No. 1, pp. 1-34, 1979.
27. S. R. Ranganathan, *Prolegomena to library classification*. London: Asia Publishing House, 1967.
28. S.R. Ranganathan, *Elements of library classification*. Bombay: Asia Publishing House. pp. 3, 1960.
29. D. V. Aptagiri, M.A Gopinath, A.R.D. Prasad, *A frame-based knowledge representation paradigm for automating POPSI*. Knowledge Organization, 1995.
30. V. Broughton, *The need for a faceted classification as the basis of all methods of information retrieval*. Aslib Proceedings, 58(1/2) pp. 49-72, 2006.

31. I. Zaihrayeu, M. Marchese, F. Giunchiglia, *Encoding Classifications into Lightweight Ontologies*. Proceedings of the 3rd European Semantic Web Conference (ESWC), 2006. Lecture Notes in Computer Science, Vol. 4011, p. 80-94.
32. T. Berners-Lee, J. Hendler, O. Lassila, *The semantic web*. Sci. Am. 284(5): pp. 34-43, 2001.
33. A.C. Varzi, *A note on the transitivity of parthood*. Applied Ontology, 1(2), pp. 141-146, 2006.
34. F. Giunchiglia, B. Dutta, V. Maltese, *Faceted lightweight ontologies.* Conceptual Modeling: Foundations and Applications, A. Borgida, V. Chaudhri, P. Giorgini, Eric Yu (Eds.) LNCS 5600 Springer