Spingersi oltre gli attuali limiti dell'Organizzazione della Conoscenza [1]

Vincenzo Maltese
[1] DISI – Università di Trento, via Sommarive 9 - 38123 POVO, Trento, Italia (fmaltese@disi.unitn.it)

Abstract. All'interno delle comunità di Scienze dell'Informazione, Biblioteconomia ed Organizzazione della Conoscenza sono state sviluppate negli anni soluzioni dimostratesi piuttosto efficaci per la classificazione, l'indicizzazione e la ricerca di grandi moli di documenti. Tuttavia, le soluzioni attualmente impiegate sono limitate in espressività in quanto supportano solo interrogazioni utente basate strettamente sulle proprietà dei documenti quali - ad esempio - il titolo, l'autore ed il soggetto. Parallelamente, nell'ambito dell'Intelligenza Artificiale e le comunità del Web Semantico e della Rappresentazione della Conoscenza, sono state sviluppate più di recente tecniche molto potenti ed espressive che, attraverso l'utilizzo di ontologie, supportano interrogazioni utente basate su qualunque proprietà di qualunque tipologia di entità (ad esempio, le proprietà delle entità che sono soggetto di un documento). Tuttavia, queste ultime discipline non hanno scalato ancora tanto quanto l'Organizzazione della Conoscenza (in termini di diffusione ed efficacia), principalmente a causa della mancanza di una metodologia rigorosa e scalabile per la specifica di tali entità e del relativo linguaggio. In questo articolo viene presentato un nuovo approccio corredato da una nuova metodologia, chiamata DERA, ispirata all'approccio a faccette, che attingendo dalla Rappresentazione della Conoscenza permette di andare oltre le attuali limitazioni dell'Organizzazione della Conoscenza. DERA garantisce allo stesso tempo qualità, estendibilità, scalabilità ed efficacia in ricerca.

1 Introduzione

Le soluzioni sviluppate all'interno delle comunità di Scienze dell'Informazione, Biblioteconomia ed Organizzazione della Conoscenza si sono dimostrate sinora piuttosto efficaci per la classificazione, l'indicizzazione e la ricerca di documenti sia in biblioteche classiche che in archivi digitali. I documenti sono indicizzati e ricercati per le loro proprietà di base, come ad esempio il titolo, l'autore o il soggetto. Vocabolari controllati di varia natura sono impiegati al fine di uniformare la terminologia dei soggetti, garantendo così un'alta precisione in ricerca. L'espansione dei termini indicati dall'utente con sinonimi e termini più specifici estratti dal vocabolario controllato permette l'individuazione di un numero maggiore di risultati rilevanti. Storicamente, tale approccio ha scalato piuttosto bene in quanto consente la classificazione, l'indicizzazione e la ricerca di milioni di libri, anche se a costi molto elevati di formazione e manutenzione (Library of Congress 2007). Diverse metodologie sono state sviluppate per la costruzione e la manutenzione, spesso gestiti in maniera centralizzata, di vocabolari controllati. Tra questi, l'approccio a faccette (Ranganathan 1967) è noto per offrire grandi benefici in termini di qualità e scalabilità delle risorse sviluppate (Broughton 2006).

Tali tecniche sono molto efficaci per quanto riguarda le interrogazioni utente che sfruttano le proprietà di base del documento. Un tipico esempio di interrogazione supportata da queste tecniche è la seguente:

Trova i documenti con autore "Nash, David" e soggetto "scultura in legno"

Tuttavia, tali interrogazioni sono limitate in espressività in quanto falliscono in quelle situazioni in cui gli utenti non conoscono direttamente tali proprietà, ma piuttosto - per esempio - alcune proprietà dell'autore o di qualsiasi altra entità soggetto del documento (ovvero in generale le proprietà di entità correlate al documento). Ad esempio, gli utenti potrebbero formulare la richiesta di qui sopra in modo meno diretto come segue:

Trova i documenti che trattano di sculture in legno che siano stati scritti da un artista nato in Galles

[1] Questo articolo è una rielaborazione e traduzione dell'articolo originale intitolato "From Knowledge Organization to Knowledge Representation" (Giunchiglia et al. 2014) presentato per la prima volta ad ISKO UK 2013 dove ha ricevuto una menzione come uno degli articoli migliori della conferenza.

Che questo tipo di interrogazioni più espressive sia richiesto è dimostrato dal fatto che le comunità dei database e della Rappresentazione della Conoscenza hanno speso decenni nello studio e nello sviluppo di linguaggi di interrogazione altamente espressivi. Possiamo ad esempio citare il linguaggio SQL usato all'interno dei sistemi per la gestione dei database (Ramakrishnan e Gehrke 2000) ed il linguaggio SPARQL per interrogare dati in RDF (Prud'hommeaux e Seaborne 2006). La loro utilità è dimostrata da numerosi studi. Interrogazioni utente con livelli di espressività come quelle suggerite in questo articolo sono d'uso quotidiano e si rivelano efficaci in innumerevoli applicazioni desktop e Web.

Le tecniche attualmente in uso nell'Organizzazione della Conoscenza non sono sufficientemente efficaci per questa tipologia di interrogazioni. Infatti, questo richiederebbe all'operatore esperto la scomposizione manuale della richiesta utente in compiti di ricerca più piccoli e la consultazione di cataloghi esterni per ottenere tutte le informazioni necessarie per riformulare la richiesta in termini di sole proprietà del documento. Nel caso specifico della richiesta data come esempio sopra, si tratterebbe di consultare dei cataloghi per individuare i nomi di tutti gli artisti nati in Galles che hanno scritto su sculture in legno e riscrivere l'interrogazione in termini diretti di autore e soggetto. Questo è in realtà uno dei motivi per cui la ricerca di documenti è ritenuta difficile dagli utenti finali e spesso richiede l'intermediazione di un operatore esperto. Per essere in grado di supportare ed automatizzare queste ricerche occorre impiegare appropriate fonti di conoscenza, la formalizzazione dei autori e dei soggetti in termini di entità, ed un linguaggio di rappresentazione e interrogazione più espressivo.

Grazie all'impiego di ontologie che specificano il linguaggio necessario a descrivere qualunque tipologia di entità e relative proprietà, le tecniche di Rappresentazione della Conoscenza consentono non solo un efficace scambio di informazioni ed il ragionamento automatico (Berners-Lee et al. 2001) (Bouquet et al. 2004), ma anche di essere più espressivi nella ricerca di documenti rispetto all'Organizzazione della Conoscenza. Per entità s'intende ciascun oggetto del mondo che ci circonda che si ritenga tanto importante da attribuirgli un nome proprio, quali ad esempio persone, luoghi, organizzazioni, ed eventi. In effetti, da un punto di vista di Rappresentazione della Conoscenza, i documenti sono solo un particolare tipo di entità con le proprietà tipiche di base che li contraddistinguono, mentre la ricerca dei documenti è solo un caso specifico di ragionamento (basato sulla transitività dei concetti che denotano i soggetti). Tuttavia, da un punto di vista puramente pragmatico, la Rappresentazione della Conoscenza ha finora fallito, dato che a tutt'oggi manca di appropriate metodologie per la specifica delle entità e relativo linguaggio che permettano di scalare tanto quanto ha fatto sinora l'Organizzazione della Conoscenza.

Quest'articolo presenta un nuovo approccio, chiamato DERA, per lo sviluppo di ontologie che descrivono e permettono di ragionare circa le entità che sono rilevanti in un certo dominio. Ad esempio, in musica, le entità possono includere canzoni, cantanti e produttori. DERA è un approccio a faccette. Infatti, la metodologia impiegata per la costruzione e la manutenzione delle ontologie di dominio s'ispira ai principi e canoni dell'approccio a faccette così come originato nell'Organizzazione della Conoscenza. Questo rende DERA in grado di supportare una conoscenza di grandi dimensioni, dinamica, e sempre crescente. DERA è una metodologia basata su principi di Rappresentazione della Conoscenza. Infatti, DERA permette di descrivere il linguaggio necessario per la rappresentazione delle entità rilevanti in un dato dominio (D) in termini di ontologie a faccette dove le categorie fondamentali sono le *classi di entità* (E), le *relazioni* tra entità (R), e loro *attributi* (A). Tali ontologie specificano in modo esplicito e formale i termini e le relazioni tra di essi, e sono direttamente formalizzabili in logica descrittiva (Baader et al. 2002). Questo permette l'automazione di operazioni complesse come la ricerca di documenti in forma molto espressiva che sfruttano le proprietà delle entità, tramite l'utilizzo di strumenti standard di ragionamento automatico su logiche descrittive.

L'argomentazione di questo articolo segue la seguente struttura. La sezione 2 fornisce una motivazione per il lavoro presentato. Essa rende evidente l'utilità di adottare principi di Rappresentazione della Conoscenza per spingersi oltre gli attuali limiti dell'Organizzazione della Conoscenza al fine di

supportare una ricerca di documenti più avanzata ed espressiva. La sezione 3 introduce la nozione di ontologia descrittiva (ovvero un'ontologia costruita allo scopo di descrivere e ragionare sulle entità del mondo reale) che supporta la ricerca di documenti sfruttando le proprietà delle entità ad essi correlate. La sezione 4 presenta la metodologia DERA in quanto approccio innovativo che combina i vantaggi offerti dall'Organizzazione della Conoscenza (in termini di metodologie per lo sviluppo di ontologie scalabili) e quelli offerti dalla rappresentazione della Conoscenza (in termini di espressività e di efficacia della ricerca). Sezione 5 descrive i passi seguiti nella metodologia DERA per la costruzione di ontologie descrittive scalabili. La Sezione 6 presenta in sintesi lo stato dell'arte. La Sezione 7 riassume ciò che è stato presentato nell'articolo e descrive i prossimi passi.

2 Andare oltre ciò che è possibile oggi

Con l'obiettivo di fornire meccanismi efficaci per l'accesso all'informazione, diverse metodologie, sistemi e strumenti sono stati sviluppati negli anni nell'ambito dell'Organizzazione della Conoscenza per supportare la classificazione, l'indicizzazione e la ricerca di documenti. In particolare, i documenti vengono generalmente classificati per soggetto ed indicizzati sulla base delle loro proprietà quali ad esempio il titolo, l'autore ed il soggetto stesso. L'indicizzazione per titolo ed autore è chiaramente facile dato che questi sono estrapolati direttamente dal documento. L'indicizzazione per soggetto è chiaramente molto più complicata dato che essa richiede un analisi del contenuto del documento e l'applicazione di principi e regole ben precisi per la costruzione delle stringhe che denotano il soggetto combinando termini provenienti da un vocabolario controllato.

In ambienti bibliotecari la ricerca è generalmente praticata manualmente tramite la consultazione di cataloghi cartacei o elettronicamente tramite interrogazioni ad un catalogo elettronico che fornisce accesso alle classificazioni ed agli indici. Tali sistemi permettono l'identificazione di quelle entità che corrispondono ad una data interrogazione che l'utente immette attraverso uno strumento di ricerca che restituisce una serie di documenti ritenuti rilevanti per la richiesta fatta. Il tipo di interrogazioni in genere supportate attualmente da strumenti basati sull'Organizzazione della Conoscenza includono esclusivamente proprietà di base del documento, quali ad esempio:

1. Trova i documenti dal *titolo* "Il lago di Garda"
2. Trova i documenti con *soggetto* "Cromford Mill"
3. Trova i documenti con *soggetto* "Michelangelo"
4. Trova i documenti con *autore* "Nash, David" e *soggetto* "scultura in legno"
5. Trova i documenti con *autore* "Clinton, Bill" e con *titolo* contenente "autobiography"

Al fine di favorire la ricerca di un maggior numero di documenti rilevanti, tali sistemi supportano quella che in gergo è talvolta chiamata *ricerca semantica* (Giunchiglia et al. 2009a), ovvero una ricerca dove i termini costituenti il soggetto sono disambiguati ed espansi con sinonimi o termini più specifici ricavati da un vocabolario controllato. Ad esempio, il termine scultura potrebbe essere espanso col termine statua che è più specifico. Purtroppo in pratica ben pochi sistemi supportano effettivamente questa funzionalità (Casson et al. 2009).

In realtà supportare questa tipologia di interrogazioni non è sempre sufficiente. Infatti, ciò presuppone che gli utenti sappiano già in anticipo cosa intendono cercare, in termini delle proprietà di base dei documenti. Invece, quello che potrebbe accadere è che gli utenti abbiano solo una vaga idea di cosa stiano cercando e che magari, piuttosto che conoscere le proprietà dirette dei documenti, essi conoscano una o più proprietà dell'autore o di una qualunque entità che sia soggetto o che sia semplicemente correlata al documento. In tal senso, la ricerca di documenti così come supportata dalla Rappresentazione della Conoscenza va oltre quella supportata dall'Organizzazione della Conoscenza, in quanto la prima è in grado di supportare la ricerca per una qualunque proprietà di qualunque entità correlata a documenti. Esempi di interrogazioni

supportabili dalla Rappresentazione della Conoscenza e non attualmente supportabili dall'Organizzazione della Conoscenza sono forniti di seguito:

1. Trova i documenti scritti da autori di nazionalità italiana che trattino di laghi profondi più di 100 metri
2. Trova i documenti che trattano di uno stabilimento in Inghilterra fondato da Richard Arkwright durante la rivoluzione industriale
3. Trova i documenti che trattano di un artista nato in Italia tra il 1450 ed il 1550
4. Trova i documenti che trattano di sculture in legno e scritti da un artista nato in Galles
5. Trova le autobiografie scritte da presidenti degli Stati Uniti

E' fondamentale osservare che seppur le interrogazioni sopra abbiano una corrispondenza uno ad uno con le interrogazioni indicate precedentemente per l'Organizzazione della Conoscenza, quest'ultima fallirebbe nel trattarle in tale forma. Infatti, anche se queste interrogazioni possono essere approcciate da un operatore esperto andando a consultare liste, cataloghi e risorse simili, questo non è facilmente automatizzabile. Infatti, questo richiederebbe la scomposizione delle interrogazioni in interrogazioni più piccole ciascuna delle quali dovrebbe fare uso di risorse sparse per ottenere le informazioni rilevanti che sono necessarie per riformulare le interrogazioni sopra in interrogazioni basate esclusivamente su proprietà dei documenti, ovvero trattabili nell'Organizzazione della Conoscenza. Ad esempio, rispondere alla terza interrogazione sopra richiederebbe l'individuazione dei nomi di artisti italiani nati nell'intervallo di date specificate. Questo è uno dei motivi per cui la ricerca spesso risulta complicata per gli utenti finali, ragione per cui viene mediata da un esperto.

Un ulteriore ostacolo per l'Organizzazione della Conoscenza è rappresentato dal fatto che le entità indicate nei soggetti sono spesso date come stringhe informali in linguaggio naturale. Ad esempio, nelle stringhe di soggetto (1) *Buonarroti, Michelangelo* e (2) *scultura - Rinascimento*, non viene specificato che Michelangelo va inteso come l'artista italiano, che scultura è un termine che indica una forma d'arte, e che Rinascimento denota un periodo storico. La disambiguazione dei termini presenti nel soggetto è infatti possibile nell'Organizzazione della Conoscenza solo se per ciascun termine esiste un'unica voce come termine preferito in un vocabolario controllato. Questo approccio è di solito seguito per i nomi comuni, ma non sempre (dato il loro potenziale alto numero) per i nomi propri. Anche quando questo viene fatto, ad esempio nel caso dei tesauri, spesso ci si limita ad indicate una generica relazione gerarchica tra i termini, ad esempio piazzando *Buonarroti Michelangelo* come termine più specifico di *artisti italiani*. Questo però è ancora una specifica informale e piuttosto limitata in quanto non è in grado di abilitare ragionamenti complessi basati su descrizioni articolate di entità. In effetti, tale relazione gerarchica dice solo che i documenti su Michelangelo Buonarroti sono anche documenti su artisti italiani. Essa non specifica espressamente che Michelangelo sia un'artista italiano. Inoltre, lo specificare solo il nome potrebbe causare problemi nella ricerca (ad esempio, una perdita di precisione nel caso di omonimia o di copertura quando un nome alternativo viene fornito dall'utente).

E' dunque di fondamentale importanza rendere esplicito e non ambiguo il significato del soggetto, in ogni sua parte. Tra l'altro, la mancanza di formalità rende la costruzione, la manutenzione e l'impego dei soggetti piuttosto difficile e costoso. Infatti, durante la loro costruzione gli esperti devono selezionare in modo appropriato i termini dal vocabolario controllato ed incasellarli nel giusto ordine di citazione, durante la manutenzione gli esperti devono prendersi cura dell'aggiornamento di quei termini che diventano obsoleti, mentre durante la ricerca devono assistere gli utenti poco esperti che non hanno familiarità con la terminologia di dominio e col modo in cui tali termini debbano essere combinati seguendo la sintassi e le regole del linguaggio di indicizzazione utilizzato (Library of Congress 2007).

Inoltre, i soggetti ed i vocabolari da soli non dicono niente, in modo esplicito, circa Michelangelo e le sue proprietà, ad esempio la sua data di nascita o i suoi lavori, in modo tale da poter essere direttamente

impiegabili da strumenti di ragionamento automatico. Ad esempio, rispondere alla terza interrogazione sopra richiederebbe specificare nel soggetto, ad esempio tramite identificatori univoci che puntano ad una risorsa esterna, che Buonarroti Michelangelo fa riferimento proprio a quell'artista nato in Italia nel 1475.

Fig. 1. La ricerca basata sulle proprietà dei documenti è un caso particolare della ricerca basata sulle proprietà di una qualunque entità

Come esemplificato in Fig. 1, la ricerca basata sulle proprietà delle entità (tipica della Rappresentazione della Conoscenza) in effetti include la ricerca basata sulle proprietà dei documenti (tipica dell'Organizzazione della Conoscenza). Un'importante differenza è che, mentre l'Organizzazione della Conoscenza si poggia solo su vocabolari controllati ed indici di vario tipo, la Rappresentazione della Conoscenza fa uso di risorse supplementari, quali ontologie, che forniscono una descrizione esplicita degli attributi delle entità, ad esempio persone (e.g. la loro data di nascita), organizzazioni ed infrastrutture (e.g. la loro data di fondazione), eventi (e.g. quando sono avvenuti), e le relazioni tra di esse (ad esempio il fatto che una certa persona è nata in un certo paese). La Rappresentazione della Conoscenza fornisce una rappresentazione ed un linguaggio di interrogazione più espressivi, ed è in grado di codificare e interrogare automaticamente tale conoscenza.

La Scienza dell'Informazione sembra riconoscere l'esigenza di tali risorse. Possiamo ad esempio menzionare iniziative come RDA[2], FRBR[3], FRAD[4] ed il lavoro promosso da OCLC che mira ad allineare i modelli BIBFRAME e Schema.org (Godby 2013). C'è da dire che invece la Rappresentazione della Conoscenza già offre in tal senso tecniche per la rappresentazione e l'utilizzo automatico di tali risorse.

3 Ontologie da Classificazione ed Ontologie Descrittive

Le ontologie costituiscono descrizioni di alto livello di un dominio, che possono essere utilizzate da applicazioni intelligenti per trarre conseguenze implicite da conoscenza rappresentata esplicitamente (Baader et al. 2002). Questo risultato è ottenuto attraverso una qualche forma di ragionamento automatico. È stato osservato che l'Organizzazione della Conoscenza e la Rappresentazione della Conoscenza, aventi scopi differenti, utilizzano tipi diversi di risorse che corrispondono a tipi diversi di ontologie (Giunchiglia et al. 2006) (Giunchiglia et al. 2009b). Infatti, Giunchiglia et al. (2006) hanno introdotto la distinzione fondamentale tra ontologie da classificazione ed ontologie descrittive.

[2] http://metadataregistry.org/
[3] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[4] http://www.ifla.org/publications/functional-requirements-for-authority-data

L'Organizzazione della Conoscenza impiega diversi Sistemi di Organizzazione della Conoscenza (in inglese "Knowledge Organization Systems", KOS). Tali sistemi corrispondono comunemente a ciò che in Rappresentazione della Conoscenza vengono chiamate ontologie da classificazione, cioè ontologie principalmente utilizzate per classificare e ricercare documenti. In queste ontologie, il focus principale è dunque sui documenti, i termini che costituiscono le etichette dei nodi indicano insiemi di documenti, le relazioni gerarchiche tra termini denotano relazioni di insieme/sottoinsieme, e gli individui (l'estensione dei termini) sono i documenti stessi. Un esempio di tali ontologie è dato in **Error! Reference source not found.**.

Ad esempio, il termine *cavalli* denota documenti su cavalli (nel senso di animali), mentre il fatto che tale termine sia posto sotto *mezzi di trasporto* indica che i documenti su cavalli sono anche documenti su mezzi di trasporto (almeno nel contesto nel quale viene utilizzata la classificazione). Questa viene chiamata semantica da classificazione (Giunchiglia et al. 2009b). L'unica forma semplice di ragionamento effettuata nell'ambito dell'Organizzazione della Conoscenza è per supportare la ricerca di documenti e si basa sulla transitività delle relazioni gerarchiche. In realtà, questo è esattamente ciò che è necessario per poter abilitare la ricerca semantica (Giunchiglia et al. 2009a). Ad esempio, per transitività i documenti su *cavalli* possono essere restituiti per la ricerca di documenti su *infrastrutture*.

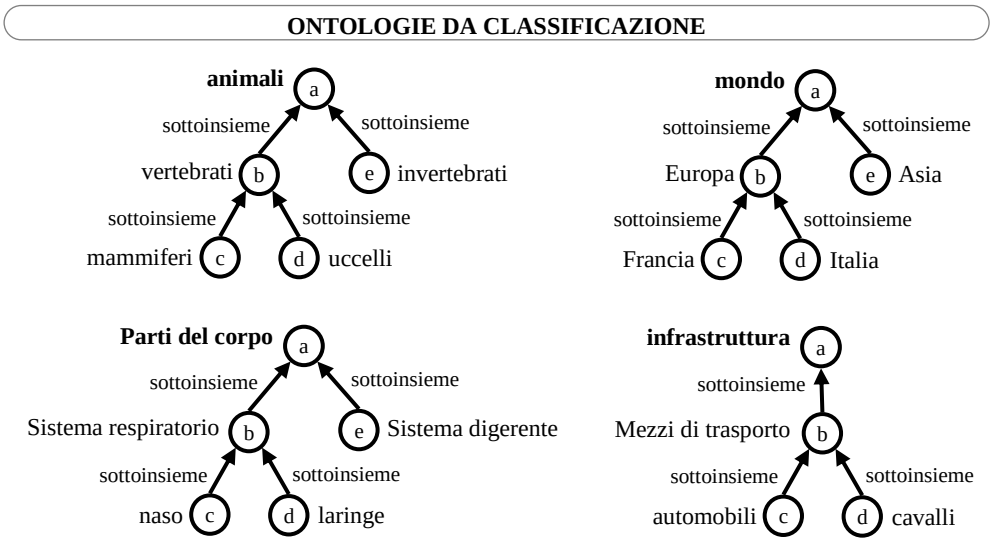

Fig. 2. Esempi di ontologie da classificazione

La Rappresentazione della Conoscenza fa uso di ontologie descrittive, ovvero di ontologie costruite allo scopo di descrivere e ragionare su entità del mondo reale. In queste ontologie, i termini denotano insiemi di entità del mondo reale, le relazioni gerarchiche forniscono la struttura portante di queste ontologie ed indicano principalmente una relazione di sottoinsieme, mentre gli individui possono essere di qualunque tipo. Per esempio, la relazione *cavallo is-a animale* indica che i cavalli sono un sottoinsieme di tutti gli animali. Questa viene chiamata semantica del mondo reale (Giunchiglia et al. 2009b).

Le ontologie descrittive forniscono conoscenza circa le entità in termini di classi, attributi e relazioni. Ad esempio, esse possono indicare che gli animali sono affetti da determinati tipi di malattie e che determinate cure sono necessarie per sconfiggerli. Un esempio di ragionamento complesso è la ricerca di cure per una malattia che colpisce un certo animale. In sostanza, lo scopo della Rappresentazione della Conoscenza è molto più ampio di quello dell'Organizzazione della Conoscenza. In realtà, preso da un punto di vista della Rappresentazione della Conoscenza, i documenti sono solo un particolare tipo di entità, con le loro proprietà peculiari (ad esempio titolo, autore e soggetto) e la ricerca dei documenti è un caso specifico di ragionamento.

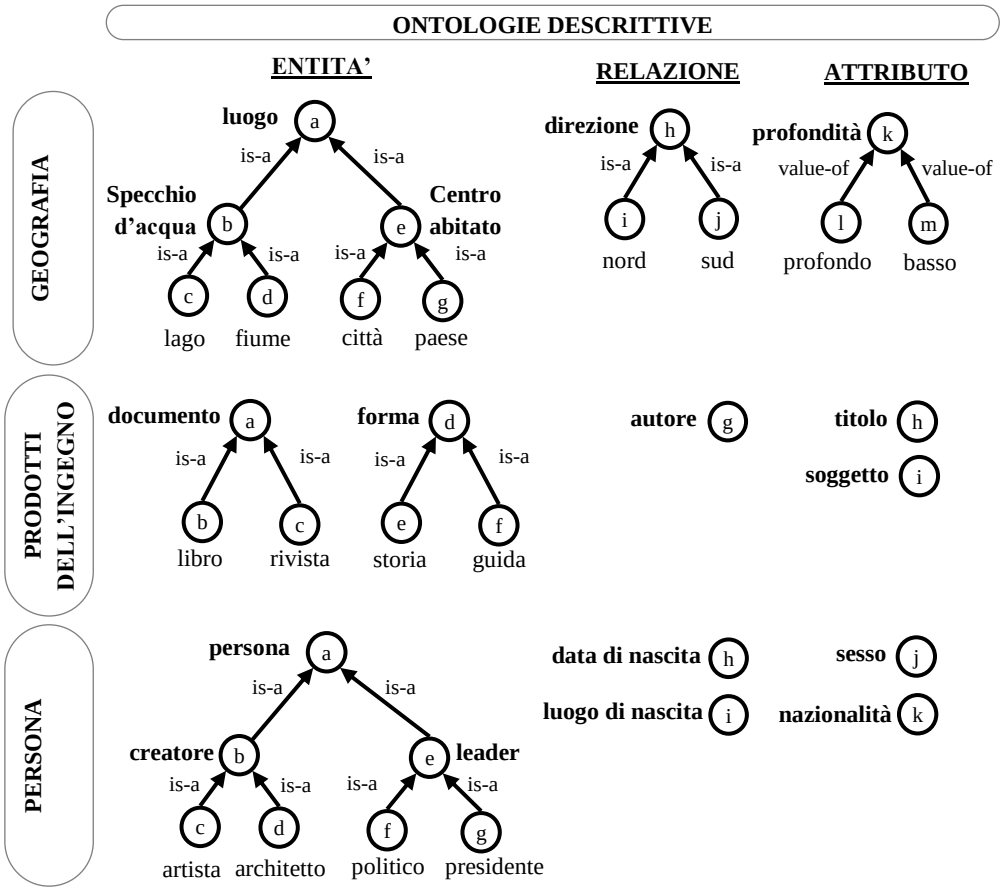

Fig. 3. Esempi di ontologie descrittive in diversi domini

Un esempio di ontologie descrittive riguardanti la geografia, i prodotti dell'ingegno ed il dominio delle persone è dato in **Error! Reference source not found.**. Nella figura, ciascun nodo denota una diversa classe di entità, relazione o attributo. Le entità che sono rilevanti nel dominio geografia sono i luoghi e sue entità più specifiche, come fiumi e laghi; le entità rilevanti nel dominio persona sono ovviamente le persone stesse; i documenti sono modellati come quelle entità che sono l'oggetto del dominio prodotti dell'ingegno, con titolo, autore e soggetto loro proprietà fondamentali. In particolare, mentre il titolo ed il soggetto sono attributi, l'autore è rappresentato come una relazione tra un documento e una persona.

Le ontologie descrittive sono popolate con entità e coi valori delle loro proprietà. Ad esempio, in Fig. 4 il dominio geografico comprende le entità *Lago di Garda* (come istanza di lago) e *Italia* (come esempio di paese), il dominio prodotti dell'ingegno comprende l'entità *Libro#1* (come istanza di libro, che a sua volta è più specifico di documento) avente il corrispondente titolo, autore e soggetto. Da notare come la stringa di soggetto *Lago di Garda - storia - guida* sia rappresentata come tre diversi valori di soggetto.

In Rappresentazione della Conoscenza, la ricerca di documenti rappresenta un tipo di ragionamento standard su ontologie descrittive. Ad esempio, rispondere all'interrogazione:

Trova i documenti scritti da italiani su qualsiasi lago con profondità superiore a 100 metri

Basandosi sulle ontologie descrittive in **Error! Reference source not found.** e corrispondenti entità in Fig. 4, corrisponde ad individuare tutte le entità che (a) sono istanze della classe *documento* e (b) con

soggetto impostato su entità che sono istanze della classe lago aventi *profondità* superiore a 100 metri e (c) con *autore* impostato su entità che hanno *nazionalità* uguali ad *Italia*. Questo restituisce Libro#1, perché (a) si tratta di un'istanza della classe *libro* che è più specifico di *documento*, (b) ha per *soggetto* il *Lago di Garda* che è un'istanza di *lago* ed ha una *profondità* di 346 metri, che è superiore a 100 metri e (c) il suo *autore* è *Solitro Giuseppe* che ha *nazionalità* impostata ad *Italia*.

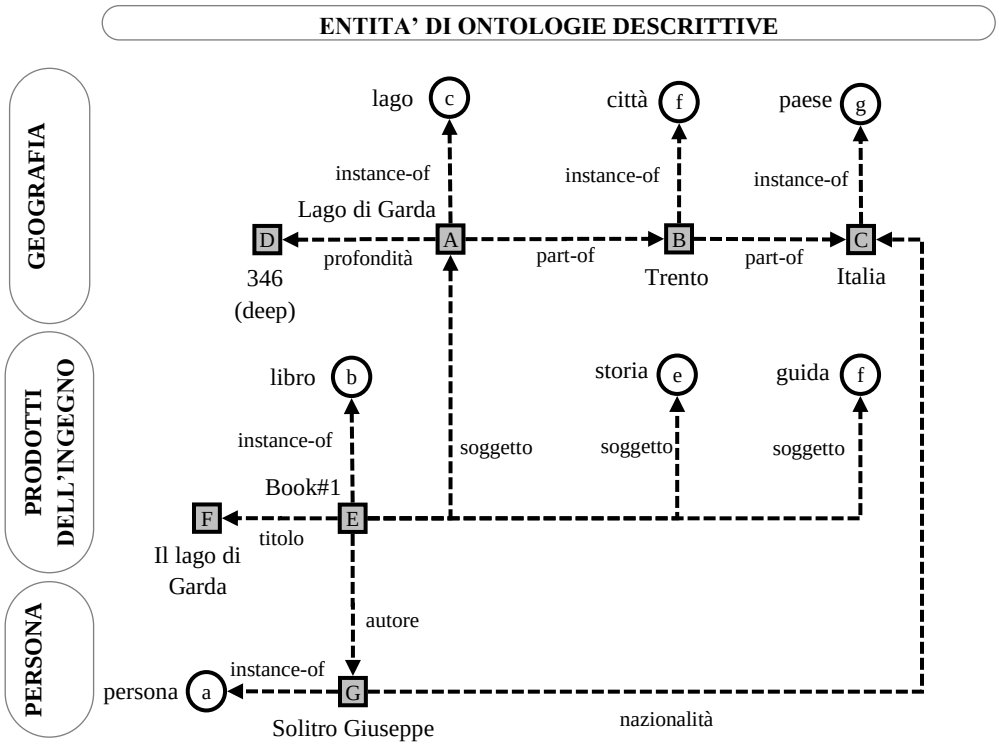

Fig. 4. Entità e rispettive proprietà che popolano l'ontologia descrittiva data in **Error! Reference source not found.**

4 L'approccio DERA

DERA fornisce una risposta concreta alla necessità di un approccio ed una metodologia adeguati per lo sviluppo di ontologie descrittive che permettano di scalare per la produzione e manutenzione di una conoscenza sempre crescente ed in evoluzione, e il loro conseguente impiego per una ricerca di documenti più espressiva. Questo a sua volta permette di costruire, a richiesta, sulla base dell'interrogazione utente, una corrispondente teoria in logica descrittiva, come descritto in dettaglio in (Giunchiglia et al. 2014).

Nel senso più ampio, DERA è un nuovo approccio a faccette fondato sia su principi di Organizzazione della Conoscenza che di Rappresentazione della Conoscenza per lo sviluppo di ontologie descrittive e il loro impiego per il ragionamento automatico. DERA è a faccette in quanto prende ispirazione dai sistemi basati su categorie fondamentali e in particolare dall'approccio introdotto da Ranganathan (1967) e successivamente semplificato da Bhattacharyya (1975), mirando così ad analoghi benefici di qualità e scalabilità. Tuttavia, si differenzia nettamente da essi in quanto l'approccio originale mira allo sviluppo di ontologie da classificazione. DERA è centrato sulla nozione di entità, piuttosto che di documento. Per entità si intende un qualunque oggetto ritenuto così importante da essere indicato con un nome proprio. Essi comprendono entità del mondo reale, come luoghi, persone, organizzazioni ed eventi, così come gli stessi documenti, qualsiasi prodotto dell'ingegno e opera d'arte. Una conseguenza immediata di adottare un approccio basato sulla Rappresentazione della Conoscenza è che DERA è un sistema di *categorie*

semantiche, vale a dire che le categorie fondamentali a supporto della specifica della terminologia di dominio mirano alla rappresentazione (piuttosto che all'organizzazione) delle entità (anziché solo dei documenti) attraverso le loro proprietà di base (quindi, non solo il soggetto).

L'approccio adotta ed estende il concetto di dominio come originariamente indicato nelle Scienze dell'Informazione. In DERA, un dominio è una qualsiasi area della conoscenza o campo di interesse che si occupa di specifici tipi di entità. Essi comprendono campi di studio convenzionali (ad esempio, la fisica, la matematica), applicazioni di discipline pure (ad esempio l'ingegneria, l'agricoltura), qualsiasi aggregazione di tali campi (ad esempio, le scienze fisiche, le scienze sociali), oppure possono anche catturare la conoscenza utile alla vita di tutti i giorni (ad esempio la musica, la cinematografia, lo sport, la cucina, il turismo). I domini offrono una vista di alto livello su tutto il campo della conoscenza, offrono un contesto globale all'interno del quale la classificazione e la ricerca possono essere supportati (Mills 2004) ed i termini disambiguati (Ciaramita e Altun 2006). I domini hanno due proprietà fondamentali (Giunchiglia et al. 2012a): essi appresentano il principale mezzo con cui la diversità viene catturata, in termini di lingua, conoscenza ed esperienza personale (ad esempio, secondo le consuetudini locali il dominio cibo può o meno includere gli insetti); inoltre, i domini permettono di scalare in quanto tengono conto dell'evoluzione della conoscenza (ad esempio, in evolvere il dominio dei trasporti potremmo estendere i mezzi di trasporto con le automobili elettriche).

Nel contesto di ciascun dominio, le entità sono descritte in termini di proprietà di base e in particolare delle loro classi, relazioni ed attributi che quindi diventano le categorie fondamentali del nostro sistema di categorizzazione. Sotto ogni categoria fondamentale, i termini sono disposti in faccette, ognuna delle quali copre un diverso aspetto del dominio. Più precisamente, definiamo una ***faccetta*** come *una gerarchia di termini omogenei che descrivono un aspetto del dominio, dove ogni termine nella gerarchia denota un concetto atomico diverso* (Giunchiglia et al. 2009b). Le faccette sono ulteriormente suddivise in sotto-faccette. Le faccette (e loro suddivisioni) sono tra loro disgiunte.

Ciascun dominio in DERA è denotato da una tripla D = <E, R, A> dove:

- E (per entità) è un insieme di faccette che raggruppano termini che denotano classi di entità, le cui istanze (le entità appunto) possono esse concrete o percepite come tali. I termini in queste gerarchie sono esplicitamente connessi attraverso relazioni *is-a* (di genere) o *part-of* (di parte).
- R (per Relation) è un insieme di faccette che raggruppano termini che denotano relazioni tra entità. I termini in queste gerarchie sono collegati da relazioni *is-a*.
- A (per attributo) è un insieme di faccette che raggruppano termini che denotano attributi qualitativi/quantitativi o descrittivi delle entità. Noi distinguiamo tra nomi di attributi e valori di attributi in modo tale che ogni nome di attributo sia associato ai valori corrispondenti. I nomi degli attributi sono collegati da relazioni *is-a*, mentre i valori degli attributi sono collegati ai corrispondenti nomi degli attributi da relazioni *value-of*.

Le faccette descritte sopra hanno una diretta traduzione in logica descrittiva. Le relazioni is-a, part-of e value-of costituiscono la spina dorsale delle faccette, si presume siano transitive ed asimmetriche, e quindi si dicono *gerarchiche*. Tutte le altre relazioni eventualmente definite che non abbiano tali proprietà, si dicono *associative* e collegano termini presenti in faccette diverse.

Le faccette costituiscono la TBox (ovvero la parte terminologica di una teoria in logica descrittiva) di una ontologia descrittiva. Ad esempio, all'interno del dominio geografico le entità rilevanti sono i *luoghi* (una faccetta di tipo E), che possono includere tra l'altro le *terre emerse* (ad esempio continenti ed isole), gli *specchi d'acqua* (ad esempio mari e fiumi), le *formazioni geologiche* (ad esempio montagne e valli), le *divisioni amministrative* (ad esempio comuni e province) e i *centri abitati* (ad esempio città e villaggi). Ciascuno di essi genera una diversa sotto-faccetta di classi di entità. Le relazioni spaziali tra luoghi possono includere *vicino*, *adiacente*, *di fronte*. Tali entità possono essere descritte in termini di *lunghezza* (ad esempio di un fiume, con valori qualitativi *lungo* e *corto*) o la *profondità* (ad esempio di un lago, con

valori qualitativi *profondo* e *basso*). Ciascuno di essi genera una sotto-faccetta di attributi. Si veda l'esempio in Fig. 5.

ENTITA'	RELAZIONE	ATTRIBUTO
Luogo	**Direzione**	**Nome**
Terra emersa	(is-a) Est	**Latitudine**
(is-a) Rilievo naturale	(is-a) Nord	**Longitudine**
(is-a) Rilievo continentale	(is-a) Sud	**Altitudine**
(is-a) Montagna	(is-a) Est	**Area**
(is-a) Collina		**Popolazione**
(is-a) Rilievo oceanico	**Relativo**	
(is-a) Monte sottomarino	(is-a) Sopra	**Profondità**
(is-a) Collina sottomarina	(is-a) Sotto	(value-of) profondo
(is-a) Depressione naturale		(value-of) basso
(is-a) Depressione continentale	**Contenimento**	
(is-a) Valle	(is-a) parte di	**Lunghezza**
(is-a) Canale		(value-of) lungo
(is-a) Depressione oceanica		(value-of) corto
(is-a) Valle oceanica		
(is-a) Canale oceanico		
Specchio d'acqua		
(is-a) Specchio d'acqua dinamico		
(is-a) Corso d'acqua		
(is-a) Fiume		
(is-a) Ruscello		
(is-a) Specchio d'acqua statico		
(is-a) Lago		
(is-a) Stagno		

Fig. 5. Versione semplificata della terminologia del dominio geografico secondo DERA

Quando le faccette vengono popolate con entità specifiche di un dominio, le relazioni di *instance-of* (di istanza) connettono le entità vere e proprie alle loro rispettive classi in E. Le entità sono descritte in termini di relazioni (R) ed attributi (A), ciascuno dei quali è a sua volta una coppia <n , v> dove n è il nome dell'attributo o della relazione e v è il suo valore coerente con quanto definito rispettivamente in A per gli attributi e in R per le relazioni.

Le entità e le loro proprietà che popolano le faccette costituiscono la ABox (ovvero la parte assertiva di una teoria in logica descrittiva) di una ontologia descrittiva. Ad esempio, il *Lago di Garda* (un'entità) può essere descritto come un'istanza di *lago* (che è una classe nella sotto-faccetta *specchio d'acqua*), che si trova in *Italia* (indicato attraverso la relazione *parte di*) con la *profondità* (che è un nome di attributo) di 346 m (che è un valore quantitativo) che può essere considerato *profondo* (che è un valore qualitativo).

5 La metodologia DERA per la costruzione e manutenzione di ontologie descrittive

La metodologia impegnata in DERA segue un insieme semplificato e minimale di principi guida, descritti in (Giunchiglia et al. 2012b), che si ispirano agli innumerevoli canoni e principi descritti da Ranganathan in (Ranganathan 1967). Essa guida l'intero processo di costruzione e manutenzione delle ontologie a faccette, ognuna delle quali copre un diverso aspetto del dominio. Tuttavia, in contrasto con l'approccio originale, DERA mira allo sviluppo delle faccette in quanto ontologie descrittive (anziché ontologie da classificazione). Le fasi principali del metodo sono le seguenti:

Fase 1: Identificazione dei concetti atomici. I termini rilevanti per il dominio in una certa lingua (ad esempio, in inglese o in italiano) sono raccolti, esaminati e disambiguati in concetti atomici. I termini sono raccolti principalmente intervistando esperti di dominio e consultando la letteratura disponibile per quel dominio inclusi indici, compendi, glossari, opere di riferimento. L'analisi di log delle ricerche fatte dagli utenti, quando disponibili, possono essere estremamente utili per determinare gli interessi degli utenti stessi. I termini con lo stesso significato (sinonimi) sono raggruppati insieme ed associati ad una descrizione in linguaggio naturale che rende esplicito il loro significato. Questo corrisponde a ciò che nell'approccio a faccette è chiamato il *piano verbale* ed a ciò che in (Giunchiglia et al. 2012a) è chiamato il *livello di linguaggio naturale*. Ciascun gruppo di termini denota un concetto atomico diverso che viene successivamente classificato ed a seconda dei casi come classe di entità (E), relazione (R) o attributo (A). Questo corrisponde a ciò che nell'approccio a faccette è chiamato il *piano delle idee* e ciò che in (Giunchiglia et al. 2012a) è chiamato il *livello di linguaggio formale*. Ad esempio, nel dominio geografico i termini *stream* e *watercourse* sono sinonimi in inglese, il cui significato può essere descritto in italiano come "un corpo d'acqua naturale che scorre sopra o sotto terra" (linguaggio naturale). Il gruppo denota una classe di entità (ovvero un concetto atomico a livello di linguaggio formale). In altre parole lo stesso concetto può essere espresso in inglese ed italiano come segue:

(E) ENG - watercourse, stream: "a natural body of running water flowing on or under the earth"
ITA - corso d'acqua: "un corpo d'acqua naturale che scorre sopra o sotto terra"

Questo è diverso dalla metodologia a faccette originale, non solo in termini di categorie fondamentali, ma soprattutto perché l'approccio di Ranganathan non contempla l'individuazione di sinonimi e relative definizioni nelle varie lingue. Meccanismi basati su vocabolari controllati sono invece stati considerati a partire da Battacharyya (1982).

Fase 2: Analisi. I concetti atomici vengono analizzati per *genus et differentia*, ovvero al fine di identificare le similarità e le differenze. L'obiettivo principale è quello di identificare il maggior numero di proprietà distintive - chiamate caratteristiche - degli oggetti del mondo reale rappresentati dai concetti. Questo permette di essere il più specifici possibile nel differenziare tra i concetti. Ad esempio, si può riconoscere che nel dominio geografico possiamo identificare le seguenti caratteristiche per il concetto di *fiume*:

- Uno specchio d'acqua
- Uno specchio d'acqua dinamico
- Nessun margine fisso
- Confinato all'interno di un letto
- Più grande di un ruscello

Questa fase è molto simile all'approccio a faccette originario.

Fase 3: Sintesi. I concetti atomici raccolti in precedenza sono disposti in faccette in modo tale che ad ogni livello della gerarchia, ciascuno dei quali rappresenta un diverso livello di astrazione, i concetti siano raggruppati per una caratteristica comune. I concetti che condividono la stessa caratteristica formano un insieme, o matrice, di concetti omogenei. I concetti in ogni matrice possono essere via via ulteriormente organizzati in sotto-gruppi (o sotto-faccette), generando così un nuovo livello della gerarchia. Ciascun concetto è collegato al concetto genitore attraverso una relazione esplicita *is-a* (di genere), *part-of* (di parte) oppure *value-of* (di valore) a seconda dei casi. Così, possiamo progressivamente ottenere il seguente risultato:

Specchio d'acqua
(is-a) Specchio d'acqua dinamico
  (is-a) Corso d'acqua
    (is-a) Ruscello
    (is-a) Fiume
(is-a) Specchio d'acqua statico
  (is-a) Stagno
  (is-a) Lago

Questa fase è in parte diversa dalla metodologia a faccette originale, dove le relazioni di genere e parte sono lasciate implicite. Infatti, mirando alla creazione di ontologie da classificazione, i termini sono disposti in faccette secondo relazioni gerarchiche generiche. Inoltre, nell'approccio originario, ed in genere nell'Organizzazione della conoscenza, non si distingue tra classi ed entità. Tra l'altro, le relazioni esplicite rendono la manutenzione più rigorosa. Ad esempio, questo facilita la distinzione tra relazioni transitive e non transitive (Maltese and Farazi 2011).

Fase 4: Standardizzazione. Ciascun concetto atomico può essere potenzialmente indicato con uno qualsiasi dei termini del gruppo di sinonimi. Quando il gruppo contiene più di un termine, un termine standard (o preferito) dovrebbe essere scelto tra i sinonimi. Questo è solitamente fatto identificando il termine più comunemente utilizzato nel dominio e che minimizza l'ambiguità. Questo è simile all'approccio seguito dalla risorsa linguistica WordNet, dove i termini nell'insieme di sinonimi sono ordinati. Ad esempio, è possibile specificare che il termine preferito in inglese tra *stream* e *watercourse* è il primo.

Questo è diverso dalla metodologia a faccette originale, in cui un solo termine viene mantenuto nello schema di classificazione, mentre gli altri vengono scartati. Eventualmente venga richiesto, risorse esterne sono necessarie per identificare eventuali sinonimi e per ottenere definizioni. Sinonimi e definizioni sono invece generalmente previsti in schemi a faccette più recenti.

***Fase 5: Ordinamento*.** I concetti atomici all'interno di ciascuna matrice vengono ordinati. Ranganathan suggerisce diversi criteri d'ordinamento. Essi comprendono ad esempio: ordine cronologico, ordine spaziale, per aumento o diminuzione di quantità, per aumento di complessità, per ordine canonico (ovvero secondo l'ordine tradizionalmente seguito), ordine alfabetico. Ad esempio, nel dominio geografico si può seguire l'ordine canonico.

Questa fase è molto simile all'approccio tradizionale. L'ordinamento non è considerato essenziale nella Rappresentazione della Conoscenza, ma risulta molto utile per scopi di manutenzione, ad esempio per controllare il livello di copertura di una faccetta o per facilitare l'identificazione di una posizione adatta per un nuovo concetto.

Passo 6: Formalizzazione. Le categorie fondamentali E/R/A sono disegnate per consentire una naturale formalizzazione delle faccette in logica descrittiva. Questo passaggio viene solo in parte ed implicitamente eseguito nell'Organizzazione della Conoscenza. Infatti, la formalizzazione comprende ciò che nell'approccio a faccette è chiamato il *piano di notazione*, cioè il livello in cui una notazione univoca viene utilizzata per collegare sinteticamente il significato e per dare ordine ai termini. Tuttavia, il modo in cui questo viene fatto in DERA rende possibile l'automazione di compiti non banali, come la ricerca avanzata di documenti altamente espressiva e basata su proprietà di entità. In effetti, la ricerca di documenti può essere ricondotta ad un problema di *instance retrieval* in logica descrittiva (Baader et al. 2002).

6 Stato dell'arte

In Scienze dell'Informazione nel tempo sono state sviluppate diverse metodologie per la costruzione e la manutenzione di sistemi per l'organizzazione della conoscenza. In particolare, nei sistemi di indicizzazione per soggetto basati su categorie i termini rilevanti per un certo dominio sono organizzati in uno schema di classificazione costituito da poche categorie fondamentali. Poiché il fine ultimo è la costruzione di *soggetti* di documenti, tali sistemi sono fondati su *categorie sintattiche*, vale a dire le categorie che svolgono un ruolo nella sintassi del linguaggio di indicizzazione dei soggetti, cioè il linguaggio usato per costruire le stringhe di soggetto memorizzate negli indici di soggetto. Le gerarchie sotto ogni categoria fondamentale codificano i diversi aspetti o faccette della conoscenza di dominio.

I vari approcci si differenziano per tipo e numero di categorie fondamentali. Kaiser (1911) ha proposto solo tre categorie, ovvero *Concreto*, *Processo* e *Paese*; Vickery (1960) ha adottato tredici categorie. Ranganathan (1967) indicava cinque categorie, ovvero *Personaggi*, *Materia*, *Energia*, *Spazio* e *Tempo*. Bhattacharyya (1975) ha semplificato le categorie proposte da Ranganathan proponendo solo *Disciplina*, *Entità*, *Proprietà* ed *Azione*. In questi approcci, le faccette di applicabilità generale sono chiamati isolati o modificatori comuni (ad esempio, *lingua* e *forma* del documento). Tuttavia, Ranganathan è stato il primo ad aver proposto e formalizzato una teoria di analisi a faccette, oggi ampiamente riconosciuta come una metodologia fondamentale che guida la creazione di schemi di classificazione di alta qualità, in termini di robustezza, estensibilità, riusabilità, compattezza e flessibilità (Broughton 2006). L'approccio di Ranganathan consente di scalare in quanto permette l'aggiunta di nuova conoscenza in qualsiasi momento, secondo le esigenze.

Viceversa, in Rappresentazione della Conoscenza finora non sono state proposte metodologie per lo sviluppo di ontologie descrittive che permettano di scalare tanto quanto è stato fatto in Scienze dell'Informazione e nell'Organizzazione della Conoscenza. In Rappresentazione della Conoscenza gli attuali approcci per la costruzione e manutenzione di ontologie si concentrano su temi di validazione di ontologie (Guarino e Welty 2002), strumenti di supporto (Corcho et al. 2004), criteri generali di progettazione (Gruber 2003), o sul processo di costruzione dell'ontologia stessa (Fernandez-Lopez 1999).

In particolare, OntoClean (Guarino e Welty 2002) suggerisce una serie di meta-proprietà che impongono una serie di vincoli per la valutazione ed il miglioramento della struttura tassonomica delle ontologie che risultano essere molto utili durante il processo di costruzione (Welty et al. 2004). Welty e Jenkins (1999) hanno proposto un'ontologia specifica per la descrizione di documenti e loro soggetti, ma essi non affrontano alcuna questione metodologica né forniscono un'implementazione esplicita. Poiché lo sviluppo di ontologie è un compito estremamente oneroso e soggetto ad errori, molti approcci hanno tentato di riutilizzare risorse esistenti (Stuckenschmidt et al. 2004). Si va da risorse lessicali (ad esempio WordNet) a risorse di domini specifici (come UMLS per la medicina e AGROVOC per l'agricoltura). Tutti questi approcci sottolineano l'utilità della conoscenza specifica di dominio (Laursen et al. 2008).

7 Conclusioni

In questo articolo abbiamo mostrato come, nonostante le numerose soluzioni introdotte, gli approcci impiegati attualmente in Organizzazione della Conoscenza per la classificazione, l'indicizzazione e la ricerca di documenti, impiegando sistemi di organizzazione della conoscenza che corrispondono ad ontologie da classificazione, siano limitate in espressività in quanto supportano solo interrogazioni utente basate sull'uso diretto delle proprietà di base dei documenti.

Da questo punto di vista la Rappresentazione della Conoscenza è molto potente e potenzialmente illimitata in quanto, impiegando ontologie descrittive, è in grado di supportare interrogazioni utente

basate su qualunque proprietà di qualunque entità. Questo motiva l'utilità di andare oltre i limiti attuali dell'Organizzazione della Conoscenza e di estenderli fino ad incorporare principi di Rappresentazione della Conoscenza applicati alla definizione della terminologia necessaria per rappresentare i documenti e le entità ad essi correlate al fine ultimo di abilitare una ricerca di documenti più espressiva.

In questo articolo abbiamo presentato in particolare un nuovo approccio a faccette, chiamato DERA, ispirato da principi di Rappresentazione della Conoscenza, ed una corrispondente metodologia per lo sviluppo di ontologie descrittive scalabili e di alta qualità. DERA permette la definizione della terminologia necessaria a descrivere le entità rilevanti per un dominio (inclusi i documenti) in termini di classi, relazioni ed attributi e consente il ragionamento automatico. In particolare, l'approccio abilita una ricerca altamente espressiva di documenti che sfrutta le proprietà delle entità.

Colmando il gap tra l'Organizzazione della Conoscenza e la Rappresentazione della Conoscenza, DERA supera le limitazioni e fa leva sui rispettivi punti di forza di questi due approcci. In realtà, DERA eredita le proprietà di qualità e scalabilità proprie dell'approccio a faccette e l'espressività e l'efficacia della ricerca resa possibile dalla Rappresentazione della Conoscenza. Grazie alla metodologia proposta, i domini in DERA sono flessibili, riutilizzabili, e permettono di scalare e di tenere conto della diversità del mondo e dell'evoluzione della conoscenza. Il ragionamento automatico è reso possibile perché le faccette associate alle categorie fondamentali E/R/A hanno una diretta formalizzazione in logica descrittiva.

Nell'immediato futuro intendiamo continuare a sperimentare DERA in domini verticali supportati da piattaforme collaborative. Finora, la metodologia si è già dimostrata efficace in esperimenti condotti nel dominio geografico, ad esempio per la codifica della terminologia (Giunchiglia et al. 2012b) e per il supporto alla ricerca di mappe in un geo-catalogo semantico (Farazi et al. 2012). Più in dettaglio, in (Giunchiglia et al. 2012b) si descrive lo sviluppo di un'ontologia descrittiva per il dominio geografico la quale comprende attualmente più di 1000 concetti e circa 7 milioni di entità spaziali principalmente tratte da GeoNames ed il Getty Thesaurus of Geographic Names (TGN). In (Farazi et al. 2012) si descrive come l'utilizzo di un'ontologia descrittiva a faccette in combinazione con strumenti standard di Intelligenza Artificiale porti ad un miglioramento significativo dei risultati della ricerca delle mappe. Inoltre, negli ultimi anni i nostri sforzi sono stati diretti allo sviluppo di una nuova risorsa che abbiamo chiamato Universal Knowledge Core (UKC), e una piattaforma collaborativa per l'impiego di esperti per la costruzione e la manutenzione di ontologie descrittive (Tawfik et al. 2015).

Intendiamo inoltre valutare i costi di tali attività, anche se la nostra ipotesi è che esso sia paragonabile ai costi necessari per un sistema per l'organizzazione della conoscenza, con la differenza che l'ontologia descrittiva prodotta avrebbe un'applicabilità più ampia rispetto ad un sistema per l'organizzazione della conoscenza (Maltese and Farazi 2011).

Bibliografia

Baader, Franz and Calvanese, Diego and McGuinness, Deborah and Nardi, Daniele and Patel-Schneider, Peter. 2002. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press.

Battacharyya, G. 1975. POPSI: its fundamentals and procedure based on a general theory of Subject Indexing Languages. Library Science with a slant to documentation 16 (1): 1-34.

Battacharyya, G. 1982. Classaurus: its fundamentals, design and use. Universal classification: subject analysis and ordering systems. Proceedings of the fourth International Study Conference on Classification Research, 6th Annual Conference of Gesellshaft für Klassifikation, 1: 139-148.

Berners-Lee, Tim and Hendler, James and Lassila, Ora. 2001. The semantic web. Scientific American, 284 (5): 28-27.

Bouquet et al. (2004). Bouquet, Paolo and Giunchiglia, Fausto and van Harmelen, Frank and Serafini, Luciano and Stuckenschmidt, Heiner. 2004. Contextualizing ontologies. Journal of Web Semantics, 1(4): 325-343.

Broughton, Vanda. 2006. The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings, 58(1/2): 49-72.

Casson, Emanuela and Fabbrizzi, Andrea and Slavic Aida. 2009. Subject search in Italian OPACs: An opportunity in waiting? IFLA 2009 Satellite Meetings.

Ciaramita, Massimiliano and Altun, Yasemin. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. Conference on Empirical Methods in Natural Language Processing, 594-602.

Corcho et al. (2004). Corcho, Oscar and Gomez-Perez, Asuncion and Gonzalez-Cabero, Rafael and Suarez-Figueroa, Carmen. 2004. ODEVAL: a tool for evaluating RDF(S), DAML + OIL, and OWL Concept Taxonomies. First Conference on AI Applications and Innovations, 369–382.

Farazi, Feroz, and Maltese, Vincenzo and Dutta, Biwanath Ivanyukovich, Alexander and Rizzi, Veronica. 2012. A semantic geo-catalogue for a local administration. Artificial Intelligence Review: 1-20.

Fernandez-Lopez, M. 1999. Overview of methodologies for building ontologies. Workshop on ontologies and problem-solving methods at IJCAI.

Giunchiglia, Fausto and Marchese, Maurizio and Zaihrayeu, Ilya. 2006. Encoding Classifications into Lightweight Ontologies. Journal of Data Semantics, 8: 57-81.

Giunchiglia, Fausto and Kharkevich, Uladzimir and Zaihrayeu, Ilya. 2009. Concept search. European Semantic Web Conference (ESWC).

Giunchiglia, Fausto, Dutta, Biswanath and Maltese, Vincenzo. 2009. Faceted lightweight ontologies. In "Conceptual Modeling: Foundations and Applications", LNCS Springer.

Giunchiglia, Fausto and Maltese, Vincenzo and Dutta, Biswanath. 2012. Domains and context: first steps towards managing diversity in knowledge. Journal of Web Semantics, 12-13: 53-63.

Giunchiglia, Fausto and Dutta, Biswanath and Maltese, Vincenzo and Farazi, Feroz. 2012. A facet-based methodology for the construction of a large-scale geospatial ontology. Journal on Data Semantics, 1(1): 57-73

Giunchiglia, Fausto and Dutta, Biswanath and Maltese, Vincenzo, 2014. From Knowledge Organization to Knowledge Representation. Knowledge Organization Journal, 41(1), 44-56

Godby, Carol Jean. 2013. The Relationship between BIBFRAME and the Schema.org 'Bib Extensions' Model: A Working Paper. Dublin, Ohio: OCLC Research.

Gruber, Thomas R. 2003. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poll (Eds.), Formal ontology in conceptual analysis and knowledge representation, Padova, Italy.

Guarino, Nicola and Welty, Christopher (2002). Evaluating Ontological Decisions with OntoClean. Communications of the ACM, 45(2): 61-65.

Kaiser, J. 1911. Systematic indexing. London: Isaac Pitman & Sons.

Lauser, Boris and Johannsen, Gudrun and Caracciolo, Caterina and Keizer, Johannes and Van Hage, Willem Robert and Mayr, Philipp. 2008. Comparing Human and Automatic Thesaurus Mapping Approaches in the Agricultural Domain. International Conference on Dublin Core and Metadata Applications.

Library of Congress. 2007. Library of Congress Subject Headings: Pre- vs. Post-Coordination and Related Issues. Library of Congress Cataloging Policy and Support Office Technical Report.

Maltese, Vincenzo and Farazi, Feroz. 2011. Towards the Integration of Knowledge Organization Systems with the Linked Data Cloud. UDC seminar.

Mills, Jack. 2004. Faceted classification and logical division in information retrieval. Library Trends, 52(3): 541-570.

Prud'hommeaux, Eric and Seaborne, Andy (2006). SPARQL Query Language for RDF. W3C Working Draft. http://www.w3.org/TR/2006/WD-rdf-sparql-query-20061004/

Stuckenschmidt, Heiner and Van Harmelen, Frank and Serafini, Luciano and Bouquet, Paolo and Giunchiglia, Fausto. 2004. Using C-OWL for the Alignment and Merging of Medical Ontologies. First International workshop on Formal Biomedical Knowledge Representation.

Ramakrishnan, Raghu and Gehrke, Johannes. 2000. Database Management Systems. McGraw-Hill.

Ranganathan, S. R. (1967). Prolegomena to library classification. London: Asia Pub. House.

Tawfik, Ahmed and Giunchiglia, Fausto and Maltese, Vincenzo (2015). A Collaborative Platform for Multilingual Ontology Development. World Academy of Science, Engineering and Technology, 8 (12).

Vickery, Brian Campbell. 1960. Faceted classification: A guide to the construction and use of special schemes. London: ASLIB.

Welty, Christopher and Jenkins, Jessica. 1999. Formal ontology for subjects. Journal on Data and Knowledge Engineering, 32 (1): 155-181.

Welty, Christopher and Mahindru, Ruchi Chu-Carroll, Jennifer. 2004. Evaluating ontology cleaning. AAAI: 311-316.