

Dalla teoria alla pratica. Ontologie che funzionano - DRAFT

1. Introduzione

Come conseguenza dell'avvento massiccio della digitalizzazione, processo che interessa un numero crescente di settori, dai beni culturali e le pubbliche amministrazioni ai settori privati, le ontologie hanno assunto un ruolo chiave come principio organizzativo dei dati.

Tuttavia, nessuno dei modelli usati per descrivere i dati è autocontenuto e può essere sufficiente da solo: per questo motivo, nel mondo delle ontologie computazionali, è oggi frequente incorrere nell'affermazione "No ontology is an island", parafrasi del celebre verso del poeta inglese John Donne. Secondo il paradigma dei Linked Data, il Web è una rete di *risorse* collegate tra di loro da *relazioni*: risorse e relazioni sono espresse mediante un insieme di vocabolari formalizzati come ontologie, secondo gli standard del Web Semantico. Standardizzazione, sviluppo di strumenti per interrogare dati provenienti da sorgenti diverse e interoperabilità diventano, in quest'ottica, sempre più importanti.

2. Ontologie fatte bene

Le metodologie per la progettazione e lo sviluppo di ontologie (*ontology engineering*) hanno accompagnato dall'inizio lo sviluppo stesso del Web Semantico; scaturite da un'intensa riflessione filosofica relativa al processo di modellazione, hanno acquisito rilevanza sempre maggiore con l'avvento dei Linked Data. Nel mondo aperto dei Linked Data, dove le ontologie prendono vita fuori dalle 'isole' in cui sono state concepite e sviluppate, è di cruciale importanza l'interoperabilità tra le singole ontologie, necessaria non solo per renderle intellegibili e uniformi dal punto di vista della modellazione, ma soprattutto per garantire uniformità e principi comuni nella descrizione dei dati.

OntoClean

Se in questo momento le metodologie di *ontology engineering* si focalizzano sul ciclo di vita delle ontologie e sugli aspetti collaborativi nel loro sviluppo, è necessario menzionare una metodologia che ha esercitato una forte influenza nello sviluppo di ontologie dall'avvento del Web Semantico, denominata OntoClean (Guarino e Welty 2009). Fortemente influenzata dall'ontologia fondazionale Dolce (Gangemi et al. 2002), OntoClean si focalizza sulla natura delle classi che costituiscono la tassonomia di un'ontologia. OntoClean si basa su un insieme di proprietà delle classi (identità, unità, rigidità, dipendenza), intese come *proprietà* della proprietà caratteristica della classe, quella cioè che ne definisce l'appartenenza alla classe. Per questo motivo, tali proprietà sono più correttamente denominate *meta-proprietà*. Ognuna delle meta-proprietà ha un comportamento specifico rispetto all'ereditarietà (può essere ereditabile oppure non ereditabile dalle sottoclassi), con conseguenze significative sulla tassonomia delle classi: per esempio, la proprietà di una classe di essere 'anti-rigida' è ereditata dalle sue sottoclassi, le quali dovranno quindi necessariamente possedere tali proprietà.

Molto sinteticamente, l'*identità* (I) è la meta-proprietà che caratterizza le proprietà che permettono di identificare univocamente gli oggetti di un certo tipo. Per una statua, per esempio, tale proprietà potrebbe essere la forma tridimensionale che essa possiede, mentre il materiale di cui è composta (ad es. l'argilla) non ha tale capacità, preso isolatamente, di identificare la singola statua. L'identità, in quanto tale, viene ereditata dalle sottoclassi, che possono aggiungervi ulteriori caratteristiche: ne consegue che la statua e il materiale di cui è

composta potrebbero essere collegate tra loro solo tramite una relazione non tassonomica (per esempio la proprietà dell'artefatto statua 'essere composto di' un certo materiale). L'*unità* (U) è riferita alla proprietà di un tipo di oggetto di essere unitario, nel senso che è possibile identificare per quell'oggetto una struttura di parti che nel complesso formano l'entità individuale (o *whole*, cioè intero, nella terminologia di OntoClean); tale proprietà viene anch'essa ereditata dalle sottoclassi. Il concetto di persona è un esempio di classe caratterizzata da unità: la persona è formata di parti che sono riconosciute come un tutto organizzato nel concetto di persona e tipi diversi di persona erediterebbero questa proprietà. Di grande rilevanza per la modellazione ontologica è la meta-proprietà denominata *rigidità* (R), che si riferisce alla resistenza ai cambiamenti di un tipo di oggetto. Esempio per eccellenza di questa meta-proprietà è il concetto di studente, contrapposto a quello di persona: mentre il secondo non è soggetto a cambiamenti (perlomeno, non in riferimento all'intervallo che intercorre tra nascita e morte), il secondo è invece tipicamente temporaneo, in quanto si configura come un *ruolo* piuttosto che come una caratteristica permanente¹. La rigidità non è ereditata: Persona può sussumere Studente (secondo un'interpretazione lasca della metodologia), ma non viceversa: infatti l'*anti-rigidità* (intesa come negazione della rigidità e non come assenza di rigidità) viene ereditata dalle sottoclassi (con la conseguenza che sottoclassi di Studente potranno essere solo tipi specifici di studenti, non persone). Infine, la *dipendenza* (D) si riferisce alla proprietà di un tipo di oggetti di dipendere da un altro per la propria definizione, come avviene nel caso del concetto di studente (o ruolo), la cui esistenza dipende da un'istituzione educativa di qualche tipo. In definitiva, le implicazioni delle meta-proprietà per la gerarchia delle classi possono essere definite come segue, date due proprietà, *p* e *q*, di cui *q* sussume *p*:

1. se *q* è anti-rigida, *p* deve essere anti-rigida;
2. se *q* è caratterizzata da un criterio d'identità, *p* deve essere caratterizzata dallo stesso criterio;
3. se *q* è caratterizzata da un criterio di unità, *p* deve essere caratterizzata dallo stesso criterio;
4. se *q* è anti-unitaria, anche *p* deve esserlo.

(adattato da Guarino e Welty 2009)

Un'osservazione importante riguarda la relazione tra l'identificazione delle meta-proprietà delle classi dell'ontologia e la realtà esterna: la metodologia, infatti, deve essere applicata alla definizione delle classi così come risultano dalle scelte di modellazione, astraendo dalla natura 'intrinseca' dei concetti a cui esse si riferiscono. Gli autori di OntoClean portano a esempio di questo fatto la definizione di persona, confrontandola con quella che si potrebbe desumere dal romanzo "Il mago di Oz": benché "avere un cervello" sia comunemente ritenuta una proprietà necessaria del concetto di persona, non è tale se tale definizione è desunta dal romanzo in questione, dove lo Spaventapasseri, concettualizzato nel romanzo come persona, non ha un cervello (e si reca quindi dal mago per riceverne uno). Il punto della metodologia, quindi, non è nel guidare l'ontologo nella definizione delle classi sulla base dell'applicazione di categorie filosofiche alle entità del mondo: esso risiede piuttosto nel porre dei vincoli alle relazioni di sussunzione tra le classi a valle delle scelte ontologiche compiute dall'ontologo.

¹ La classe "studente" viene anche definita come *phased sortal*, tipo specifico di *sortal* che attraversa fasi diverse nella sua esistenza pur mantenendo la sua identità, tipologia cui appartiene in senso stretto anche la classe "persona" vista nei suoi cambiamenti biologici.

Qualunque siano le scelte ontologiche, la metodologia suggerisce che l'ossatura dell'ontologia (*backbone ontology*) sia costituita da classi rigide, a cui sono 'agganciate' le altre classi. Infine, un'altra linea guida ampiamente accettata di OntoClean consiste nel tenere separati ruoli e entità, evitando che le entità sussumano i ruoli, di natura tipicamente circoscritta nello spazio-tempo.

Ontology engineering

Un aspetto importante del design di ontologie riguarda l'emergere, anche in questo caso sulla base della spinta propulsiva delle ontologie fondazionali quale DOLCE, di schemi di modellazione standard, come menzionato nella sezione relativa alla modellazione concettuale; tali schemi sono denominati *ontology design pattern* (ODP). I design pattern si configurano come veri e propri cataloghi di configurazioni di classi e proprietà adatte a modellare schemi ricorrenti nella modellazione ontologica. Finalizzati a superare i limiti dei linguaggi delle logiche descrittive, inerentemente classificatorie e basate su relazioni binarie tra classi (sia di sussunzione, sia orizzontali), gli *ontology design pattern* favoriscono l'allineamento tra ontologie di dominio diverse. Un esempio è dato dalla descrizione dei processi: applicando uno stile di descrizione basato sui ruoli dei partecipanti al processo, di facile integrazione con le risorse linguistiche che descrivono i ruoli del verbo, si garantisce che la descrizione degli eventi assuma lo stesso schema di base in ontologie diverse. Elemento essenziale di questo schema è il concetto di *ruolo*, colto da un'apposita classe, *Role*², che si collega (mediante *object properties* in OWL) all'entità che riveste il ruolo in questione, tipicamente un agente. A questo schema, attraverso l'integrazione di altri pattern, è possibile aggiungere ulteriori vincoli: per esempio, associare la partecipazione di un ruolo a un processo a un certo intervallo di tempo.

Anche se con termini diversi per classi e relazioni e con alcune varianti (un possibile schema è riportato in Figura 1), lo schema di descrizione dei processi basato sui ruoli si presenta in numerose ontologie di dominio: per esempio, la Figura 2 riporta un frammento dell'ontologia COMM (Arndt et al. 2007) in cui si descrivono i ruoli di input e output nel processo di annotazione semantica dei media. Nella Figura 3, invece, è presente un adattamento dello schema al contesto delle attività produttive in ambito artistico nell'ontologia CIDOC-CRM³ (Doerr 2003), che descrive il dominio dei beni culturali: in questo caso, i 'ruoli' specifici di 'attore' e 'strumento' di un 'processo produttivo', o di 'modifica', di un bene artistico sono delegati alle relazioni *carriedOut* e *usedSpecificObject* che collegano l'attività (*Activity*) rispettivamente con uno o più attori (*Actor*) e con uno o più oggetti (*Thing*).

² <http://ontologydesignpatterns.org/wiki/Submissions:AgentRole>

³ <http://www.cidoc-crm.org>

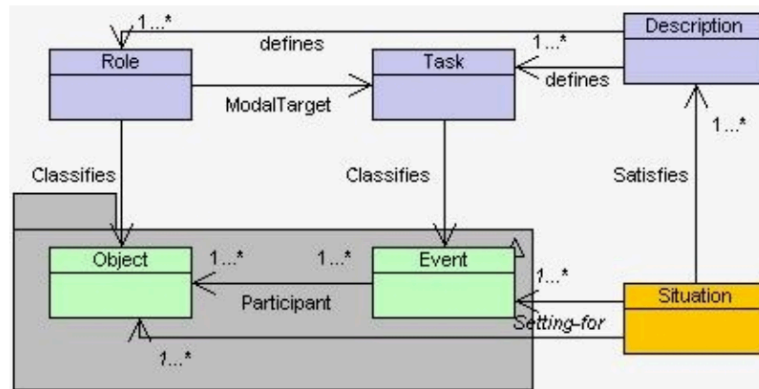
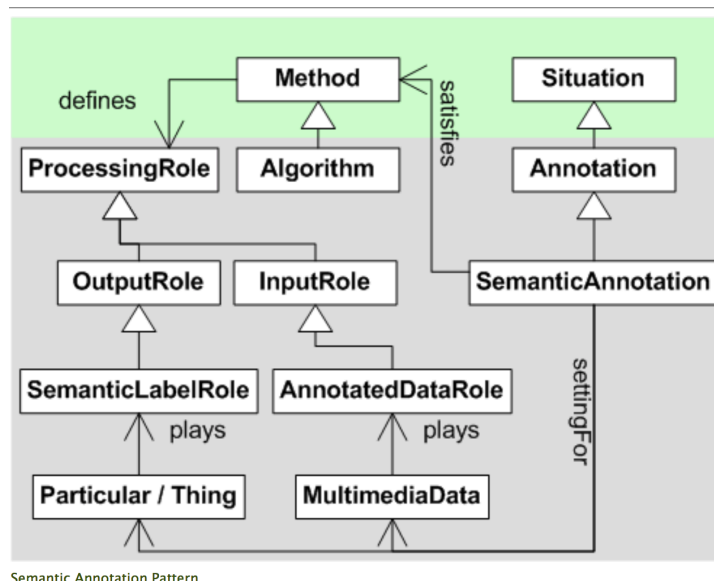


Figura 1: ODP per ruolo e task. Tratto da Gangemi (2005).



Semantic Annotation Pattern

Figura 2: rappresentazione di ruoli nel al processo di annotazione semantica di dati multimediali nell'ontologia COMM (Ontology of Multimedia).

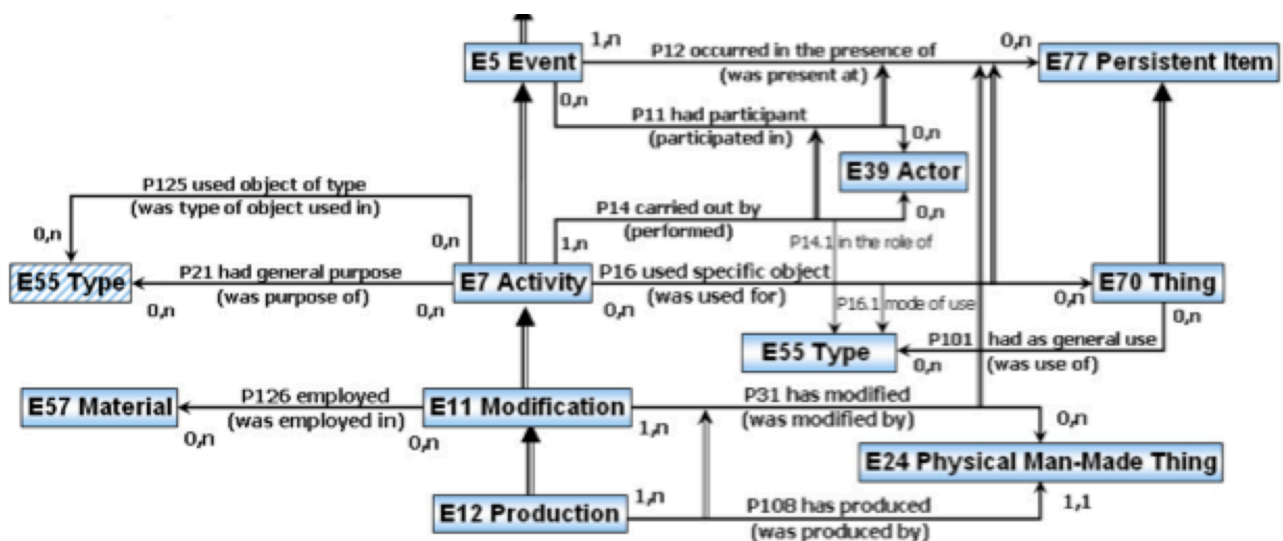


Figura 3: un frammento dell'ontologia CIDOC-CRM. Si noti la relazione tra le classi Actor e Thing, rispettivamente agente e strumento di un'attività. Le linee doppie indicano relazioni di sussunzione tra classi.

Come esempio di integrazione con le risorse linguistiche, si consideri FrameNet⁴, risorsa linguistica sviluppata dall'Università di Berkeley (Baker et al. 98), che consiste in un catalogo di frame⁵ linguistici sviluppato per l'analisi del linguaggio naturale. In FrameNet, un frame rappresenta uno schema di ruoli (o *elements*) associato a un insieme di forme verbali; alcuni ruoli costituiscono il nocciolo (*core elements*) del significato del frame e devono comparire necessariamente nell'applicazione del frame, altri sono facoltativi. Per esempio, il frame per il verbo 'mangiare' (*ingestion*) comprende due ruoli come core elements, *ingestor*, cioè l'entità senziente che effettua l'azione di ingerire il cibo, e *ingestible*, cioè il cibo stesso, inteso come entità commestibile. Eventuali altri elementi, come lo strumento usato (*instrument*) o la durata dell'azione di mangiare (*duration*) sono *non-core elements* e possono comparire o essere assenti nelle manifestazioni del frame nel linguaggio naturale.

Ingestion

[Lexical Unit In](#)

Definition:

An **Ingestor** consumes food or drink (**Ingestibles**), which entails putting the **Ingestibles** in the mouth for delivery to the digestive system. This may include the use of an **Instrument**. Sentences that describe the provision of food to others are NOT included in this frame.

The wolves DEVoured the carcass completely.

FEs:

Core:

Ingestibles [Ingible] Ingestor [Ing] Semantic Type: Sentient	The Ingestibles are the entities that are being consumed by the Ingestor . The Ingestor is the person eating or drinking.
--	---

Non-Core:

Degree [Degr] Semantic Type: Degree	The extent to which the Ingestibles are consumed by the Ingestor . The wolves DEVoured the carcass completely
---	--

Figura 4: interfaccia di consultazione di FrameNet (il frame Ingestion)

Gli ODP sono ingrediente fondamentale di una metodologia di ontology engineering denominata NeOn (Suárez-Figueroa et al. 2012), acronimo di *networked ontologies* in riferimento alla visione delle ontologie come rete di risorse collegate da relazioni di ri-uso, *merging*, allineamento.

Orientata agli aspetti collaborativi nello sviluppo e nel mantenimento di network di ontologie, la metodologia NeOn si articola in un insieme di 9 *scenari*, associati a specifiche attività e documenti. Per esempio, lo scenario 1 "Dalle specifiche all'implementazione" ("From Specification to Implementation") prevede che si produca un documento di requisiti per l'ontologia ("Ontology Requirements Specification Document") e include attività quali la raccolta dei requisiti ("Ontology requirements specification"), la programmazione delle attività ("Scheduling"), la formalizzazione dell'ontologia ("Ontology formalization") e la sua implementazione ("Implementation"). Lo Scenario 3 "Reusing Ontological Resources" comporta il fatto di cercare ("Searching") le risorse ontologiche potenzialmente idonee al

⁴ <https://framenet.icsi.berkeley.edu>

⁵ Si veda la dispensa sulla Modellazione concettuale.

riuso, valutarle ("Assessing"), confrontarle ("Comparing") e selezionarne ("Selecting") un sottoinsieme.

Elemento importante della metodologia è il modello del ciclo di vita dell'ontologia: secondo tale modello, dalla fase di inizializzazione ("Initiation phase") l'ontologia passa alla fase di progettazione ("Design phase"), quindi di implementazione ("Implementation phase") e infine di mantenimento ("Maintenance"). Per far fronte alla necessità di adeguare l'ontologia all'evoluzione del contesto di utilizzo e delle risorse collegate può essere necessario ripetere l'intero processo in maniera iterativa per più volte dopo la fase iniziale.

3. Ontologie, tassonomie, tesauroi

Le risorse semantiche strutturate in forma di ontologie e codificate nei linguaggi del Web Semantico costituiscono solo la punta dell'iceberg di un molto più vasto insieme, che include risorse di vario tipo, caratterizzate dal fatto di essere internamente strutturate secondo un modello gerarchico.

WordNet

In primis, si considerino le risorse linguistiche, tra le quali la risorsa per eccellenza è WordNet⁶ (Miller 95). Sviluppata all'università di Princeton a partire dagli anni '80, WordNet si presenta come un dizionario strutturato secondo relazioni di tipo gerarchico. WordNet è una risorsa linguistica che consiste in un lessico organizzato secondo relazioni di significato. In WordNet, gli elementi del lessico (sostantivi, verbi, aggettivi, ecc.), sono raggruppati in insiemi denominati *synset*. I termini (o lessemi) che fanno parte di uno stesso *synset* siano sinonimi, quindi ogni *synset* corrisponde a un determinato significato, rappresentato appunto da termini diversi e sinonimi tra di loro. Si noti che lo stesso termine può comparire in *synset* diversi, anzi questa costituisce di gran lunga la situazione più comune, a causa della polisemia, cioè molteplicità di significati, che caratterizza le parole nel linguaggio naturale. Per esempio, il sostantivo inglese "road" appartiene a due *synset* diversi, uno letterale, come 'strada' e l'altro metaforico, dove la parola in questione acquisisce il significato di 'mezzo per' un determinato fine ("la strada per la fama"):

- S: (n) road, route (an open way (generally public) for travel or transportation)
- S: (n) road (a way or means to achieve something) "the road to fame"

Le relazioni semantiche in WordNet, tuttavia, non sono limitate al raggruppamento dei sinonimi, ma includono anche una componente tassonomica. I *synset*, infatti, sono organizzati secondo relazioni di iperonimia e iponimia, cioè rispettivamente secondo relazioni di tipo 'termine più generale' e 'termine più specifico'. Per esempio, il primo *synset* nell'esempio sopra riportato ha come *synset* più generale, cioè come iperonimo, il *synset* che include il termine "way" ("artifact consisting of a road or path affording passage from one place to another", cioè "artefatto che consiste di una strada o percorso che fornisce un passaggio da un posto all'altro") e come *synset* più specifici un insieme di *synset* che includono termini specifici per indicare una strada in senso concreto: per esempio, quello che include "highway" e "main road" (autostrada), o quello che include "railway line" e "rail line" (linea ferroviaria). Oltre alle relazioni tassonomiche (iperonimia e iponimia) in WordNet sono presenti altre relazioni semantiche, quali *mereonimia* (per i sostantivi, "window" è parte di "building"), *implicazione*

⁶ <https://wordnet.princeton.edu>

(per i verbi, "snore" implica "sleep"), *troponimia* (per i verbi, "lisp", cioè "chiacchierare", è un troponimo di "talk", parlare, cioè una modalità specifica di quest'ultimo). Le relazioni verticali (iperonimia e iponimia) e orizzontali (mereonimia, implicazione, ecc.) tra i termini fanno di WordNet una vera e propria 'rete di parole', tale che non solo ogni termine acquisisce significato dalla sua posizione rispetto alla rete, ma le relazioni tra termini forniscono informazioni essenziali per il trattamento del linguaggio naturale: per esempio, per la disambiguazione del significato di un termine, oppure per stabilire la vicinanza semantica tra termini. Per questo motivo, WordNet è definita anche un'ontologia "leggera" (light ontology), dato che ha la struttura di un'ontologia senza avere l'apparato formale proprio delle vere ontologie. Per la lingua italiana è disponibile una versione specifica di WordNet, parte di un progetto denominato MultiWordNet⁷, caratterizzato dal tentativo di creare una corrispondenza tra i *synset* di più lingue nazionali (tra cui l'italiano).⁸

La parola cercata (*shout*) è presente in quattro accezioni diverse (o *synset*) in Wordnet. Ogni *synset* è identificato da un numero ed è accompagnato da una descrizione. Ad ogni *synset* corrisponde un diverso concetto di SUMO. Per esempio, *shout* inteso come *gridare sgarbatamente all'indirizzo di qualcuno* (200845299) corrisponde al concetto *Communication* in SUMO, mentre *gridare* nel senso di *parlare a voce molto alta* (200912473) corrisponde al concetto *Vocalizing* in SUMO

Risultati della ricerca

SUMO Search Tool

This tool relates English terms to concepts from the [SUMO](#) ontology by means of mappings to [WordNet](#) synsets.

English Word: Verb Submit

According to WordNet, the verb "shout" has 4 sense(s).

[200845299](#) use foul or abusive language towards; "The actress abused the policeman who gave her a parking ticket"; "The angry mother shouted at the teacher".

- SUMO Mappings: [Communication](#) (subsuming mapping)

[200912048](#) utter aloud; often with surprise, horror, or joy; "I won!" he exclaimed; "Help!" she cried; "I'm here," the mother shouted when she saw her child looking lost".

- SUMO Mappings: [Communication](#) (subsuming mapping)

[200912473](#) utter in a loud voice; talk in a loud voice (usually denoting characteristic manner of speaking); "My grandmother is hard of hearing--you'll have to shout".

- SUMO Mappings: [Vocalizing](#) (equivalent mapping)

[200913065](#) utter a sudden loud cry; "She cried with pain when the doctor inserted the needle"; "I yelled to her from the window but she couldn't hear me".

- SUMO Mappings: [Vocalizing](#) (equivalent mapping)

Explore the word [shout](#) on the WordNet web site.

Synset diversi possono corrispondere allo stesso concetto!

Figura 5: un esempio di corrispondenza tra i concetti dell'ontologia fondazionale SUMO e WordNet

Grazie alla sua natura strutturata, WordNet è stato incorporato in molte ontologie, al fine di fornire una corrispondenza tra i concetti dell'ontologia e il linguaggio naturale. Per esempio, l'ontologia fondazionale SUMO (Niles e Pease, 2001 e 2013), è stata allineata con WordNet,

⁷ <http://multiwordnet.fbk.eu/english/home.php>

⁸ Si noti che l'allineamento non può che essere parziale, dato che un determinato concetto non è necessariamente espresso nel lessico di tutte le lingue.

cioè a ognuna delle sue classi è stato associato un *synset* di WordNet. Ci sono due modi in cui un *synset* di WordNet può corrispondere a un concetto SUMO: il concetto di SUMO corrisponde esattamente al significato del *synset* (corrispondenza denominata "equivalent mapping") oppure il concetto di SUMO include al suo interno come concetto più specifico ('*sussume*') anche il significato del *synset* (corrispondenza denominata "subsumption mapping").

L'esempio in Figura 5 mostra la corrispondenza tra il termine inglese "shout" (urlare) e i concetti dell'ontologia SUMO. Mentre le prime due accezioni del termine corrispondono entrambe (per sussunzione) al concetto di comunicazione in generale ("Communication") in SUMO, le ultime due accezioni del termine (rispettivamente "parlare a voce molto alta" [200912473] e "fare un urlo" [200913065]), corrispondono entrambe al concetto Vocalizing in SUMO, cioè il loro significato corrisponde al concetto di 'vocalizzare', secondo una corrispondenza di tipo equivalente ("equivalent mapping"). Si noti quindi come una parte del significato presente nel linguaggio naturale (e rispecchiato nella distinzione tra i due *synset* in WordNet) non viene colta dall'ontologia, che ha per necessità una granularità maggiore.

Thesauri

Il termine *thesauro* si riferisce a un insieme di termini strutturati tra di loro secondo relazioni gerarchiche. Utilizzato in ambiti disciplinari diversi (tra cui, per esempio, la filologia classica e la linguistica), il suo utilizzo nella rappresentazione della conoscenza deriva principalmente dalla biblioteconomia e dall'archivistica ed è finalizzato a rendere univoco e stabile l'uso dei termini in un determinato ambito di conoscenza. Per questo motivo, i thesauri sono normalmente accessibili al pubblico - anche se talvolta dietro pagamento di abbonamenti - in vari formati, normalmente accompagnati da strumenti di consultazione che permettono all'utente di cercare termini nel thesauro e di navigarne la struttura.

Accanto ai thesauri propriamente detti è possibile reperire nell'ambito della rappresentazione della conoscenza molte risorse che, seppur non definibili come thesauri in senso stretto, sono finalizzati a normare l'uso dei termini in un determinato dominio e sono caratterizzati da una struttura tassonomica. Per enfatizzare la dimensione terminologica, tali risorse sono spesso denominate *vocabolari* o *dizionari*, con un'accezione che sottintende la loro validità rispetto a un ambito specifico di conoscenza.

Il sistema Iconclass⁹ per la classificazione dei soggetti iconografici (cioè i motivi ricorrenti nelle arti figurative), il sistema dell'Association for Computing Machinery (ACM) per la classificazione degli argomenti dei lavori di ambito informatico¹⁰ sono esempi di risorse strutturate caratterizzate da una forte componente gerarchica, la prima maggiormente consona alla definizione di tesoro, la seconda più scarna e più simile a un sistema di parole chiave. Hanno forma tassonomica, seppure meno rilevante, anche gli *authority file*, raccolte di termini finalizzate a normare l'uso dei nomi propri e dei termini specialistici nella compilazione di schede d'archivio. Appartengono a questa tipologia VIAF, sistema internazionale di authority file, e i numerosi dizionari curati dalla Getty Foundation¹¹ nell'ambito dei beni culturali: tra i più noti, il dizionario dei nomi di artisti (ULAN) e il tesoro dei nomi geografici (TGN).

La Figura 6 illustra la porzione di Iconclass relativa al soggetto iconografico dato dal 'paesaggio' ("landscapes"). Come si può osservare nella figura, la tassonomia di Iconclass è formata da 10 concetti di base, che si declinano in specificazioni sempre maggiori: per

⁹ <http://www.iconclass.org>

¹⁰ <https://www.acm.org/publications/class-2012>

¹¹ <http://www.getty.edu/research/tools/vocabularies/>

esempio, il percorso dalla radice della tassonomia ("Nature") alla nozione di Paesaggio ("landscapes") passa da un concetto di Terra intesa come corpo celeste: Natura --> Terra --> Paesaggio. Il Paesaggio, come si può osservare nella parte destra della figura, si divide ulteriormente in categorie più specifiche: 'paesaggio nella zona temperata', 'paesaggio con acqua', ecc.

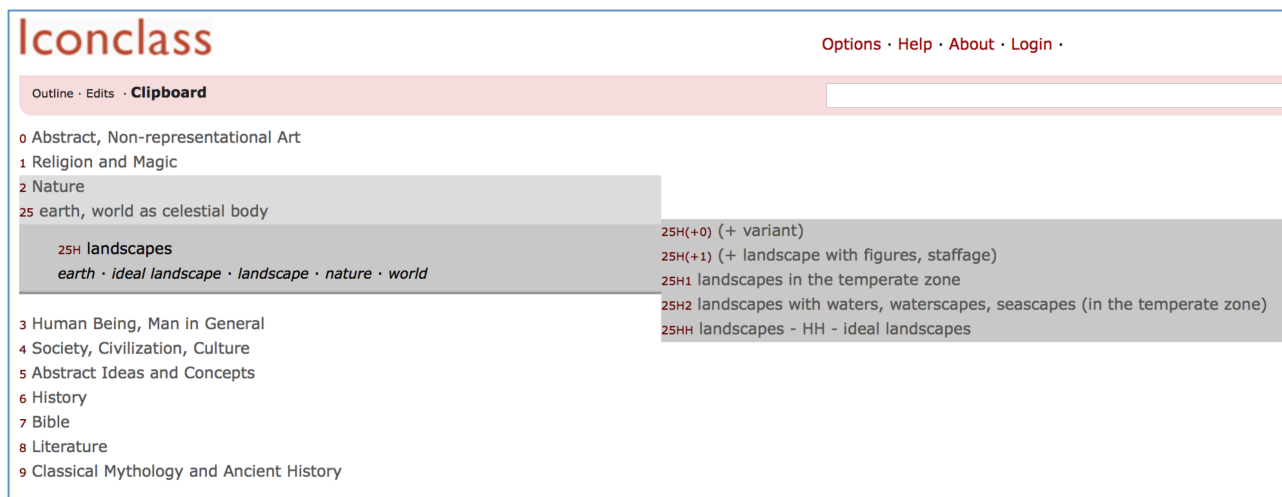


Figura 6: il concetto di paesaggio (landscapes) nel sistema Iconclass.

La figura permette anche di osservare come, per agevolare l'utilizzo dei tesauri nel trattamento automatico dell'informazione, ai termini del tesoro siano normalmente associato un sistema di indici la cui struttura permette di collocare l'indice al livello corrispondente della tassonomia. Per esempio, in Iconclass i livelli di base sono identificati dalle cifre da 0 a 9 (2 per "Nature"); il livello successivo corrisponde all'aggiunta di una seconda cifra (5 per "earth"), seguita da una lettera (H per "landscapes"), fino a comporre il l'identificativo univoco di ogni concetto. A ogni concetto, inoltre, si associa una definizione in linguaggio naturale e un insieme di parole chiave che descrivono sinteticamente il soggetto.

Con l'avvento della digitalizzazione, molti tesauri sono stati resi compatibili con il paradigma dei Linked Data: pubblicati in formato RDF, sono interrogabili mediante servizi web.

SKOS, Simple Knowledge Organization Systems

Data la pluralità di formati e strutture con cui si presentano le risorse semantiche sopra elencate (vocabolari, tesauri, sistemi di parole chiave, ecc.) il W3C, in quanto responsabile della standardizzazione del Web Semantico e dei Linked Data, ha concepito e pubblicato uno standard volto a facilitare e uniformare la pubblicazione di quelli che vengono genericamente definiti *vocabolari RDF*. Centrato sulla nozione di "concetto" e sulle sue relazioni con altri concetti e con il linguaggio, il sistema SKOS (Simple Knowledge Organization System) è rivolto - secondo le specifiche stesse del sistema - all'uniformazione di "thesauri, classification schemes, subject heading lists, taxonomies, folksonomies", con la finalità di rendere tali risorse interoperabili: "As an application of the Resource Description Framework (RDF), SKOS allows concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into other concept schemes."¹²

¹² <https://www.w3.org/TR/skos-primer>

Per esempio, il pittore italiano Giotto da Bondone è descritto come segue in VIAF¹³ (*authority file* internazionale di nomi di autori in ambito culturale):

```
<rdf:Description
rdf:about="http://viaf.org/viaf/sourceID/NKC%7Cjn19990002696#skos:Concept">

  <foaf:focus rdf:resource="http://viaf.org/viaf/27073355"/>
  <skos:altLabel>Giotto, asi 1266-1337</skos:altLabel>
  <skos:altLabel>Bondone, Giotto di, asi 1266-1337</skos:altLabel>
  <skos:altLabel>Bondone, Ambrogio di, asi 1266-1337</skos:altLabel>
  <skos:prefLabel>Giotto di Bondone, asi 1266-1337</skos:prefLabel>
  <skos:inScheme rdf:resource="http://viaf.org/authorityScheme/NKC"/>
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>

</rdf:Description>
```

Si noti in particolare, nel frammento precedente, l'utilizzo di SKOS, evidenziato dall'utilizzo di un insieme di termini propri di SKOS, evidenziati dal *namespace* corrispondente (skos:). Come sarà illustrato più avanti, tale frammento associa alla entry che rappresenta il pittore in questione un insieme di etichette, o *label*, alternative ("Bondone, Giotto di" e "Giotto di Bondone" tra le altre), indicandone una in particolare come scelta preferenziale ("Giotto di Bondone"), e ponendo in relazione la entry a un determinato schema di concetti (<http://viaf.org/authorityScheme/NKC>).

In SKOS, come previsto da RDF, i concetti sono identificati con gli URI e a essi sono associate delle etichette lessicali, introdotte dai predicati `prefLabel`, `altLabel`: scopo di queste etichette è consentire l'indicizzazione dei concetti in linguaggio naturale. Il frammento che segue, per esempio, associa al concetto `ex:animals` l'etichetta "animals" come etichetta preferita (`prefLabel`) e l'etichetta "animaux" come etichetta alternativa in lingua francese (`altLabel`), come specificato dall'uso del suffisso `@fr`.

```
ex:animals rdf:type skos:Concept;
  skos:prefLabel "animals";
  skos:altLabel "animaux"@fr.
```

Al fine di codificare le relazioni tra concetti, SKOS dispone di proprietà apposite per esprimere generiche relazioni semantiche (`related`) e relazioni di generalità/specificità (`narrower/broader`). Il frammento successivo esprime il fatto che il concetto `ex:animals` ha come concetto più specifico quello di mammiferi (`ex:mammals`) e viceversa:

```
ex:animals rdf:type skos:Concept;
  skos:prefLabel "animals"@en;
  skos:narrower ex:mammals.

ex:mammals rdf:type skos:Concept;
  skos:prefLabel "mammals"@en;
  skos:broader ex:animals.
```

¹³ http://viaf.org/viaf/27073355/#Giotto_1266?-1337

Si noti che le relazioni semantiche tra concetti codificate da SKOS forniscono una rappresentazione alternativa (ma non equivalente!) alle relazioni tassonomiche tra classi esprimibili in RDF/OWL tramite `rdf:subClass` (SKOS stesso è disponibile in formato OWL). Altra importante funzione di SKOS consiste nell'associare i concetti alla loro documentazione, associando ogni concetto alla sua definizione (`definition`), a un esempio (`example`) e ai cambiamenti occorsi al concetto (`historyNote` e `changeNote`).

Alcune proprietà molto importanti permettono di collegare un concetto singolo a uno schema di concetti, o vocabolario (tramite la proprietà `inScheme`) e di richiamare concetti appartenenti a schemi diversi all'interno di una stessa descrizione (si veda per esempio, nel frammento di VIAF relativo al pittore Giotto, l'uso del termine `skos:inScheme` per asserire l'appartenenza del concetto allo schema identificato dall'URI <http://viaf.org/authorityScheme/NKC>). L'importanza della nozione di schema in SKOS è data dal fatto che, nel mondo dei Linked Data, un dato e la sua descrizione si trovano spesso separati dalla definizione del vocabolario o dei vocabolari su cui si basa la descrizione. Qualora schemi diversi, elaborati in momenti differenti o per finalità differenti, contengano la definizione con termini diversi di uno stesso concetto, in SKOS la corrispondenza tra concetti può essere stabilita mediante le proprietà `broadMatch` (corrispondenza lasca) e `narrowMatch` (corrispondenza stretta).

4. Linked Data

In seguito alla diffusione sempre maggiore di dati in formato digitale, provenienti da sorgenti diverse che spaziano dall'attività di istituzioni pubbliche quali ministeri e enti culturali, e dalla digitalizzazione di archivi (molto diffusa per esempio tra i canali televisivi e radiofonici pubblici e privati), all'attività dei social media e delle industrie private, il numero e la qualità dei *dataset* disponibili in rete e descritti secondo criteri semantici è cresciuto nell'ultimo decennio fino a richiedere un cambio di paradigma nelle tecnologie semantiche. L'iniziativa nota come Linked Data costituisce una risposta a questa tendenza, in cui le ontologie formali retrocedono per dare maggiore spazio all'interoperabilità dei dati. Linked Data è un metodo per pubblicare dati strutturati in modo che possano essere collegati tra loro e diventare più utili attraverso interrogazioni semantiche (Heath e Bizer 2011). Si basa su tecnologie Web standard come HTTP, RDF e URI, normalmente usate per i documenti destinati agli esseri umani, ma estese nei Linked Data per condividere informazioni leggibili automaticamente dai computer.

Il paradigma dei Linked Data si basa sui seguenti principi¹⁴:

1. Usare gli URI per denominare (identificare) le entità.
2. Usare HTTP in modo che queste entità possano essere trovate (o interpretate, o "de-referenziate").
3. Fornire informazioni utili sulla risorsa quando viene richiesta, usando standard come RDF, SPARQL, ecc.
4. Nella descrizione di un'entità, riferirsi sempre alle altre entità usando i loro URI.

¹⁴ Testo di riferimento: <http://linkeddatabook.com/editions/1.0/>

In altre parole, non solo la risorsa deve essere *dereferenziata* (cioè all'URI della risorsa si trova effettivamente la risorsa stessa!), ma insieme alla risorsa deve essere fornita una descrizione della risorsa stessa. La descrizione può essere fornita ad esempio via RDF (o SPARQL) e dovrebbe includere relazioni con altre risorse, anch'esse riferite attraverso i loro URI, ed essere espressa secondo un vocabolario di termini anch'essi dereferenziabili. Tipicamente, la modalità di pubblicazione differisce tra i dati e i vocabolari che si usano per descriverli. I vocabolari normalmente vengono pubblicati tal quali, come file rdf: nel caso in cui un singolo termine del vocabolario debba essere riferito, esso verrà concatenato all'URI del vocabolario attraverso il cosiddetto *fragment identifier* che segue il simbolo # nell'URI, come avviene nel comunissimo URI del concetto di classe in RDF:

<http://www.w3.org/2000/01/rdf-schema#Class>; questo metodo fa affidamento sulla capacità del client di per conto proprio identificare i singoli termini nel file RDF che corrisponde all'URI del vocabolario. Nel caso dei dati, essi vengono invece normalmente pubblicati tramite un *endpoint* che supporti un linguaggio di interrogazione specifico per RDF come SPARQL, oppure tramite API di tipo REST (benché nulla impedisca a un fornitore di dati di renderli disponibili come file RDF online, soprattutto nel caso di *dataset* di piccole dimensioni).

Nei Linked Data, la descrizione di una risorsa è un misto di termini dovendosi provenire da vocabolari RDF diversi (sia proprietari sia standard). Le linee guida suggeriscono l'utilizzo di vocabolari standard, ma quando ciò non è possibile, impongono di fare in modo che i propri termini siano comunque de-referenziabili. Nel caso in cui sia necessario stabilire una relazione con un termine equivalente o simile tra vocabolari diversi, è possibile stabilire la relazione tra i due mediante le proprietà `owl:equivalentClass` e `owl:equivalentProperty` (oppure usando le proprietà `rdfs:subClassOf` e `rdfs:subPropertyOf` di RDF o quelle `skos:broadMatch` e `skos:narrowMatch` di SKOS per esprimere una relazione più lasca).

Lo strumento deputato a pubblicare online i dati nel paradigma Linked Data è una Linked Data Platform (LDP), il cui schema astratto è illustrato in figura 7. La funzione di tale piattaforma è duplice: da un lato, mantenere i dati, dall'altro, renderli accessibili via http. Si noti che i dati possono essere disponibili in formato RDF e quindi immagazzinati in un software apposito (*triple store*), oppure essere disponibili in formato diverso (per esempio, perché originati da altre attività) e convertiti in formato RDF per la pubblicazione. Dato che i LDP supportano la creazione e il mantenimento di grafi diversi per rappresentare i dati, spesso le triple vengono rappresentate come quadruple, aggiungendovi un elemento che rappresenta il grafo di appartenenza.

L'accesso ai dati, quindi, può essere realizzato sia tramite un *endpoint* SPARQL (spesso disponibile, nel caso dei triple store pubblici, anche come interfaccia utente di interrogazione), oppure tramite API programmate utilizzando librerie (normalmente integrate nella LDP) che permettono l'interazione con il triple store e/o l'inoltro di interrogazioni SPARQL per via programmatica.

LINKED DATA PLATFORM

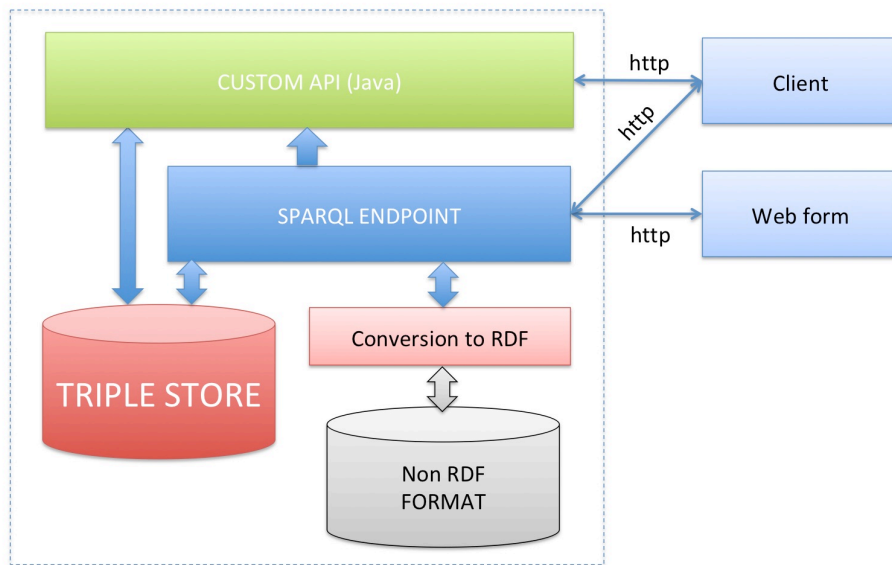


Figura 7: Schema astratto di una Linked Data Platform

Un esempio di accesso duplice è fornito dal portale Europeana, meta-portale dei beni culturali in Europa (www.europeana.eu) la cui piattaforma LOD, implementata tramite il server Virtuoso, mette a disposizione degli sviluppatori una duplice modalità di accesso, sia tramite endpoint SPARQL, sia tramite un'API sviluppata appositamente per replicare le singole funzioni di ricerca nel catalogo disponibili dall'interfaccia utente. Esempio di accesso via SPARQL è anche l'endpoint pubblico del progetto Wikidata: <https://query.wikidata.org>.

Riferimenti bibliografici

- Arndt, R., Troncy, R., Staab, S., Hardman, L., & Vacura, M. (2007). COMM: designing a well-founded multimedia ontology for the web. In *The semantic web* (pp. 30-43). Springer, Berlin, Heidelberg.
- Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1* (pp. 86-90). Association for Computational Linguistics.
- Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3), 75.
- Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening ontologies with DOLCE. *Knowledge engineering and knowledge management: Ontologies and the semantic Web*, 223-233.
- Gangemi, A. (2005, November). Ontology design patterns for semantic web content. In *International semantic web conference* (pp. 262-276). Springer, Berlin, Heidelberg
- Guarino, N., & Welty, C. A. (2009). An overview of OntoClean. In *Handbook on ontologies* (pp. 201-220). Springer Berlin Heidelberg.
- T. Heath and C. Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, 1-136. Morgan & Claypool.
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.
- Niles, I., & Pease, A. (2001, October). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001* (pp. 2-9). ACM.
- Niles, I., & Pease, A. (2003). Mapping WordNet to the SUMO ontology. In *Proceedings of the ieee international knowledge engineering conference* (pp. 23-26).
- Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., & Gangemi, A. (Eds.). (2012). *Ontology engineering in a networked world*. Springer Science & Business Media.