



Deliverable D2.2

Costruzione e Popolazione di Ontologie di Dominio

Responsabile:	Di Martino Beniamino
Afferenza	Seconda Università di Napoli
Autori	Prof. Rocco Aversa, Ing. Carlo Baia, Prof. Ida Caracciolo, Prof. Beniamino DI Martino, Prof. Pasquale Femia, Ing. Angelo Martone, Ing. Francesco Moscato, Prof. Francesco Palmieri, Ing. Massimiliano Rak, Ing. Gianmarco Romano, Ing. Salvatore Venticinque, Prof. Rosanna Verde
Afferenza	Seconda Università di Napoli
Autori	Prof. Paola Velardi
Afferenza	Università di Roma "Sapienza"
Autori	Gaetano Guerriero, Barbara Fiaschetti
Afferenza	SPACE
Autori	Dr.ssa Federica Pesce, Dr. Fabrizio Melorio
Afferenza:	STURZO

PROGETTO LC3	Revisione n*	0	Del	----
--------------	--------------	---	-----	------



INDICE

- [TR2.2.1 Stato dell'Arte dei modelli e dei linguaggi standard per la rappresentazione di ontologie nel Semantic Web](#)
- [TR2.2.2 Rassegna ed Analisi delle tecnologie e degli strumenti per la costruzione Computer Assisted di Ontologie](#)
- [TR2.2.3 Rassegna ed Analisi delle metodologie, delle tecniche e degli strumenti per la derivazione automatica di Ontologie](#)
- [TR2.2.4 Definizione di una tecnica per la derivazione automatica di Ontologie da corpora documentali](#)
- [TR2.2.5 Realizzazione di uno Strumento Prototipale per la derivazione automatica di Ontologie da corpora documentali](#)
- [TR2.2.6 Definizione di una tecnica per la derivazione di Ontologie da strutture di documenti gerarchicamente o relazionalmente organizzate](#)
- [TR2.2.7 Realizzazione di uno strumento Prototipale per la derivazione di Ontologie da strutture di documenti gerarchicamente o relazionalmente organizzate](#)
- [TR2.2.8 Ontology Manager: architettura e funzionalità](#)
- [TR 2.2.9 Costruzione automatica di una ontologia in formato OWL sul dominio "Il rapimento di Aldo Moro"](#)
[D 2.2.9 Elenco dei concetti specifici del dominio "Aldo Moro"](#)
- [TR2.2.10 Esperienze di Costruzione di Ontologie di Dominio e relative Annotazioni Semantiche](#)
- [TR2.2.11 Applicazione di metodologie e tecniche Statistiche per la costruzione di ontologie di dominio ad un corpus relativo al dominio "Rapimento Aldo Moro"](#)
- [TR2.2.12 Definizione di un MAG Repository per lo storage sicuro dei contenuti multimediali granulari, in grado di gestire contenuti semanticamente annotati](#)
- [TR2.2.13 Supporto alla definizione architetture del Sistema MAG Teca](#)

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------



Lo scopo di questo task è la definizione e realizzazione di tecniche per la creazione, gestione e popolazione di ontologie di dominio, cioè di una rete semantica di concetti e relazioni mediante le quali annotare documenti testuali, cercando di privilegiare tecniche automatiche o semiautomatiche.

Nel seguito si descrivono nel dettaglio le attività eseguite ed i risultati conseguiti dalle Unità di ricerca della Seconda Università di Napoli (DII e JM), coadiuvata su alcuni punti dal gruppo di ricerca dell'Università la Sapienza, unità operativa del Consulente UNIMED.

Sono stati analizzati i principali modelli, linguaggi standard e tecnologie disponibili nell'ambito del semantic web, con particolare riferimento alle ontologie. Inoltre, si stanno individuando le principali ontologie di livello superiore (upper ontology) ed gli attuali linguaggi e framework di modellazione dei *Semantic Web Services*.

Sono state analizzate e confrontate le principali tecnologie per la rappresentazione strutturata di informazioni quali XML, XML Schema, RDF, RDF Schema e OWL. Inoltre, si sono analizzate e confrontate le principali ontologie di livello superiore (upper ontology) quali DOLCE, OpenCYC, Ontologia di Russel e Norvig, Sumo e Wordnet. Vedi **TR2.2.1 Stato dell'Arte dei modelli e dei linguaggi standard per la rappresentazione di ontologie nel Semantic Web.**

Sono stati individuati ed analizzati i principali strumenti e piattaforme tecnologiche, commerciali ed open source, per la costruzione computer assisted di ontologie. Tali strumenti e tecnologie sono stati classificati in base ai modelli ed alle tecniche adottate, alle funzionalità offerte, alle piattaforme di sviluppo e di esecuzione.

Sono stati analizzati e confrontati i principali strumenti Open Source per la costruzione computer assisted di ontologie, tra i quali Apollo, OIEd, OntoEdit e OntoStudio, OntoSaurus, Protégé, WebODE, WebOnto, OntoMaker. Sono stati inoltre analizzati e confrontati i principali tool e framework che supportano la pubblicazione, la scoperta e la composizione dei Semantic Web Services, tra i quali OWL-S (precedentemente DAML-S), WSMO, SWSF e WSDL-S. Vedi **TR2.2.2 Rassegna ed Analisi delle tecnologie e degli strumenti per la costruzione Computer Assisted di Ontologie**

E' stata effettuata l'analisi delle principali metodologie e tecniche nonché i principali strumenti e piattaforme tecnologiche, commerciali ed open source, per la derivazione automatica di ontologie. Tali strumenti e tecnologie sono stati classificati in base ai modelli ed alle tecniche adottate, alle funzionalità offerte, alle piattaforme di sviluppo e di esecuzione.

Sono stati analizzati e confrontati i principali strumenti Open Source per la derivazione automatica di ontologie, tra i quali Text2Onto, OntoLearn, OntoBuilder, OntoLT, JAKTE, OntoGen2.0, ASIUM, SEISD. È stata inoltre effettuata un'analisi delle principali tecniche di derivazione automatica di ontologie, classificando le tecniche e gli strumenti in base al tipo

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------



di input. Vedi **TR2.2.3 Rassegna ed Analisi delle metodologie, delle tecniche e degli strumenti per la derivazione automatica di Ontologie**

E' stato definito un metodo per la derivazione semiautomatica di ontologie da corpora documentali, che prevede l'applicazione di tecniche di *machine learning*, basate su classificazione automatica di tipo *supervised* (basate su *training sets*) ed *unsupervised* gerarchico. La procedura che si sta definendo si applica ad insiemi di testi costituenti un corpus documentale di dominio. Tale procedura produce una struttura gerarchica (albero) ad ogni nodo della quale è associato un insieme di termini, derivati mediante procedura di clustering dal corpus documentale in input, che potranno rappresentare concetti candidati a costituire l'ontologia di dominio rappresentata dalla struttura gerarchica derivata, e che andranno selezionati e validati manualmente. Gli stessi documenti o porzioni di testo, associati alle categorie derivate, potranno poi essere annotati "*coarse grained*" con i concetti associati (vedi task 2.3).

Si è definita quindi una procedura di ontology learning da corpora documentali che si articola nelle seguenti fasi elaborative: analisi testuale di documenti, indicizzazione con inverted index dei concetti e delle parole presenti nel testo, classificazione dei documenti testuali attraverso algoritmo di clustering gerarchico, utilizzo della tassonomia di documenti creati nel passo precedente per la rilevazione di concetti e relazioni tra quest'ultimi, validazione e modifica delle relazioni trovate attraverso WordNet, creazione di un'ontologia in linguaggio RDF/OWL. Vedi **TR2.2.4 (DRAFT) Definizione di una tecnica per la derivazione automatica di Ontologie da corpora documentali**

Si sta progettando e realizzando uno strumento prototipale (*OntoClust*) che implementa la tecnica di ontology learning da corpora documentali precedentemente definita. Vedi **TR2.2.5 (DRAFT) Realizzazione di uno Strumento Prototipale per la derivazione automatica di Ontologie da corpora documentali**

Si sta definendo una tecnica che permette l'estrazione semi-automatica di una ontologia dalla struttura implicita o esplicita di relazioni associate ad un insieme di documenti. Esempi di tali strutture possono essere quella gerarchica presente in una struttura di directories in cui documenti siano organizzati; o una tassonomia a cui siano associati documenti classificati in accordo ad essa; oppure la struttura relazionale rappresentata dai links degli ipertesti presenti in un sito web o la loro strutturazione logica. La struttura relazionale viene estratta e rappresentata mediante grafi (in Java) ed anche mediante insiemi di relazioni predicative (in Prolog), insieme alle caratteristiche sintattiche e lessicali estratte dai nodi informativi (documenti, o pagine web) associati a tale struttura. Complementare alla determinazione di relazioni semantiche ottenuta dalla rappresentazione e l'analisi strutturale, vi è la determinazione di relazioni semantiche tra i

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------



nodi a partire dall' analisi del testo associato ai nodi stessi, ed utilizzando tesauri quali Wordnet e Multiwordnet come base di conoscenza. Si sta inoltre definendo la procedura di validazione da parte dell' esperto di dominio dopo ogni passo di costruzione automatica. La procedura di ontology learning da strutture di documenti gerarchicamente o relazionalmente organizzate provvisoriamente definita si articola nelle seguenti fasi elaborative: parsing della struttura (sito web, directory, tassonomia descritta in RDFS, Prolog), individuazione di concetti facendo riferimento ad altre ontologie di tipo Upper Ontology (Dolce, Sumo,...), definizione ed individuazione di pattern di concetti analoghi presenti all'interno della struttura, identificazione di occorrenze diverse di uno stesso concetto facendo ricorso all'individuazione di sinonimie mediante WordNet (MultiWordNet), individuazione di sottoconcetti (subclass) e suoi attributi. (Ad esempio, con WordNet (MultiWordNet) è possibile utilizzare la relazione di Meronimia (isPartOf – potrebbe individuare un attributo); visualizzazione della rappresentazione a grafo dell' ontologia che passo dopo passo si sta costruendo, consentendo anche la validazione da parte dell'utente. Infine traduzione della rappresentazione interna in una ontologia in linguaggio OWL. Vedi **TR2.2.6 (DRAFT) Definizione di una tecnica per la derivazione di Ontologie da strutture di documenti gerarchicamente o relazionalmente organizzate.**

Si sta progettando e realizzando uno strumento prototipale (*OntoExtract*) che implementi la tecnica di ontology learning da strutture di documenti gerarchicamente o relazionalmente organizzate precedentemente definita. Vedi **TR2.2.7 (Draft) Realizzazione di uno strumento Prototipale per la derivazione di Ontologie da strutture di documenti gerarchicamente o relazionalmente organizzate**

E' stata definita l' architettura e le funzionalità dell' Ontology Manager, previsto dal progetto come modulo indipendente rispetto al MAG repository (MAG Teca), che immagazzina e gestisce le ontologie prodotte (automaticamente o manualmente) e/o utilizzate. L' ontology manager gestisce la produzione manuale delle ontologie (mediante strumenti di editing), la memorizzazione delle ontologie, la gestione delle differenti versioni delle ontologie stesse, l' accesso ed il browsing condivisi delle stesse, la ricerca di concetti e relazioni nelle ontologie stesse. Vedi **TR2.2.8 (DRAFT) Ontology Manager: architettura e funzionalità**

Si sta realizzando, a cura del gruppo di ricerca guidato dalla Prof. Paola Velardi dell' Università la Sapienza (Unità operativa del Consulente UNIMED), uno strumento prototipale web-based per la validazione collaborativa di una ontologia semi-automaticamente appresa. Tale strumento, denominato TAV (TAXonomy Validator), è reso disponibile liberamente sul sito: <http://lcl.uniroma1.it/tav>. Rispetto a strumenti disponibili per la costruzione e manutenzione di ontologie (vedi Protégé) TAV consente come

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------



caratteristica aggiuntiva di “spostare” nodi e sottoalberi dell’ontologia, mantenendo traccia degli autori di una modifica e “colorando” le modifiche via via apportate dal team di validazione, e ricostruendo la rappresentazione OWL dell’ontologia. E’ stata inoltre realizzata, con l’aiuto di alcuni strumenti dell’Università Sapienza (TermExtractor, TAV) una ontologia del dominio del progetto LC3, “Attentato ad Aldo Moro”, popolando l’ontologia “general purpose” WordNet con 1291 concetti aggiuntivi caratterizzanti il dominio. L’ontologia è stata costruita come segue: 1) Partendo da un insieme di documenti (articoli di giornale, trascrizioni di telegiornali ecc.) forniti dal partner Unicity, sono stati dapprima identificati, utilizzando il sistema TermExtractor, i termini specifici del dominio (ad es. *terrorismo, ultrasinistra, demoproletario.*) nonchè nomi propri (es. *Brigate Rosse, Palazzo Chigi, via Fani*) e le espressioni polirematiche (es. *stella a cinque punte, comitato di fabbrica*). 2) Poichè molti altri termini del dominio in questione sono di tipo generale, si è utilizzata come base di partenza EuroWordNet per la lingua italiana, che è stata trasformata in formato OWL e caricata sul sistema TAV. 3) Infine, l’ontologia EuroWordNet è stata arricchita con gli (oltre 1200) concetti di dominio estratti nella fase 1. L’elenco dei termini inseriti è contenuto nell’ **Allegato A2.2.9 Elenco dei concetti specifici del dominio “Aldo Moro”**, mentre i dettagli sono riportati nel **TR2.2.9 Costruzione di una ontologia in formato OWL sul dominio “Il rapimento di Aldo Moro”**.

Sono state costruite (in modalità manuale, non semiautomatica) ontologie di dominio, e relative annotazioni semantiche, per un numero di domini; in particolare per il dominio definito dal progetto MICHAEL (Multilingual Inventory of Cultural Heritage in Europe), che si pone l’obiettivo di censire le collezioni digitali del patrimonio culturale europeo, per il dominio giuridico (basata su una tassonomia realizzata da esperti di dominio a partire dalla classificazione di circa 50.000 testi di ambito giuridico), e per il dominio “caso Aldo Moro”.

In particolare, si sta costruendo una ontologia in formato OWL sul dominio “Il Progetto Michael”. Il progetto MICHAEL (Multilingual Inventory of Cultural Heritage in Europe), coordinato in Italia dal Ministero per i Beni e le Attività Culturali, si pone l’obiettivo di censire le collezioni digitali del patrimonio culturale europeo. Per la descrizione delle collezioni MICHAEL ha adottato un proprio data-model.

La costruzione dell’ontologia in formato OWL a partire dal modello dei dati delle collezioni MICHAEL è stata effettuata seguendo le seguenti fasi progettuali:

- Trasformazione del modello dati MICHAEL, di tipo XML, in un modello dati di tipo ontologico, attraverso l’utilizzo di un software open source per la creazione di ontologie (Protégé);

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------



- Testing e valutazione dello schema dati ontologico, attraverso il popolamento dello stesso con una serie di dati descrittivi campione delle collezioni digitali censite;
- Applicazione di tecniche di annotazione semantica sul modello dati ontologico, attraverso l'utilizzo di un tool progettato e sviluppato nell'ambito del progetto LC3 (OverFa).

Ulteriori dettagli sono reperibili nel **TR2.2.10: Esperienze di Costruzione di Ontologie di Dominio e relative Annotazioni Semantiche.**

Si sta costruendo una ontologia in formato OWL sul dominio giuridico ed effettuando la relativa Annotazione di documenti giuridici. Questa attività si incentra sulla realizzazione di un'ontologia di dominio giurisprudenziale e sulla successiva annotazione semantica di documenti in base all'ontologia realizzata. La definizione dell'ontologia del dominio giuridico si è basata su una tassonomia realizzata da esperti di dominio a partire dalla classificazione di circa 50.000 testi di ambito giuridico. L'attività ha portato all'individuazione di 18 macrocategorie, tra cui una relativa alla filosofia e una alla storia del diritto. Il numero di categorie alla base di questa attività ammonta a circa 9.500. Le seguenti fasi di progetto hanno permesso la costruzione dell'ontologia obiettivo dell'attività:

- Clusterizzazione delle categorie
- Individuazione delle Classi di Concetti ontologici
- Individuazione di regole di inferenza per permettere un processo di individuazione semi-automatico di concetti ontologici.
- Applicazione del sistema così costruito alla derivazione automatica di ontologie.

Ulteriori dettagli sono reperibili nel **TR2.2.10: Esperienze di Costruzione di Ontologie di Dominio e relative Annotazioni Semantiche.**

Si sta costruendo una ontologia in formato OWL sul dominio "Il rapimento di Aldo Moro". Il lavoro riguarda la realizzazione di un'ontologia che descriva i concetti fondamentali riguardanti il "Caso Moro", e l'annotazione semantica delle fonti documentali più rilevanti ad esso inerenti. Per "Caso Moro", si è voluto intendere non solo i 55 giorni del sequestro ma anche il contesto storico politico e sociale in cui si è svolto il sequestro, il dopo sequestro, i processi, e tutto ciò che ha girato intorno al "caso Moro". Ulteriori dettagli sono reperibili nel Technical Report **TR2.2.10: Esperienze di Costruzione di Ontologie di Dominio e relative Annotazioni Semantiche** ed inoltre nell' articolo: Beniamino Di Martino, Pasquale Femia, Laura Manelli, Sergio Muzzupappa, "Metodologie e strumenti informatici per una ontologia sul caso Aldo Moro", Scritture di Storia, n. 5, pp. 359-380, luglio 2008, Edizioni Scientifiche Italiane.

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------



Sono state applicate metodologie di analisi statistica testuale per la riduzione dimensionale dello spazio di rappresentazione dei testi (Analisi delle corrispondenze binarie), e per la classificazione documentale automatizzata, su un corpus costituito da 90 articoli estratti da quotidiani che trattarono gli avvenimenti che si susseguirono in seguito al sequestro di Aldo Moro. Ulteriori dettagli sono reperibili nel Technical Report **TR2.2.11 Applicazione di metodologie e tecniche Statistiche per la costruzione di ontologie di dominio ad un corpus relativo al dominio "Rapimento Aldo Moro"**

Nel seguito si descrivono i risultati conseguiti dall' Unità di ricerca SPACE, coadiuvata dal Consulente Istituto Luigi Sturzo.

E' stata definita l' architettura del MAG repository (MAG Teca) e della sua evoluzione per arrivare a gestire gli oggetti granulari (*Contlet*), al fine di consentire una gestione delle annotazioni semantiche, la loro archiviazione sicura e il relativo recupero delle stesse. In particolare si sono conseguiti i seguenti risultati:

- strutturazione della piattaforma base di digital repository, basata sulla Teca Digitale studiata da SPACE, allo scopo di gestire MAG estesi;
- produzione di una specifica dettagliata delle estensioni a questa digital library in grado di gestire in modo indipendente i singoli oggetti CONTLET semanticamente e geograficamente annotati, consentendone la loro individuazione puntuale e il loro successivo recupero;
- realizzazione di alcuni dimostratori software in ambiente open source in grado di consentire la validazione delle funzionalità della digital library, e la messa a punto delle funzionalità base di cooperazione tra applicazioni.

Ulteriori dettagli sono reperibili nel Technical Report **TR2.2.12 (Draft) Definizione di un MAG Repository per lo storage sicuro dei contenuti multimediali granulari, in grado di gestire contenuti semanticamente annotati.**

Il supporto fornito dal Consulente ILS, sulla base dell'esperienza maturata come consulente del Ministero per i Beni e le Attività Culturali sul progetto della Biblioteca Digitale Italiana e Network Turistico Culturale, nonché come produttori di specifiche outlines narrative, è descritto nel report **TR2.2.13 (Draft) Supporto alla definizione architeturale del Sistema MAG Teca**

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------



PUBBLICAZIONI PRODOTTE

F.Moscato, B.Di Martino: “Semantic Web and Semantic Information Management”, Int. J. Web and Grid Services, Vol. 4, No. 3, 2008.

F. Moscato, B. Di Martino, S. Venticinque, A. Martone, “OverFA: A collaborative Framework for Semantic Annotation of Documents and Web Sites”, to be published in: International Journal of Web and Grid Services (IJWGS), Inderscience Press.

B. Di Martino, “Semantic Web Services Discovery based on Structural Ontology Matching”, to be published in: International Journal of Web and Grid Services (IJWGS), Inderscience Press.

Beniamino Di Martino, Angelo Martone, Francesco Moscato, Salvatore Venticinque, “A versioning based framework for semantic annotation of Web documents: OVerFA” , Proc. of Int. Conf. on Methods, Moels and Information Technologies for Decision Support Systems (MTISD 2008), Lecce, Italy, 18-20 sept. 2008.

Beniamino Di Martino, Pasquale Femia, Laura Manelli, Sergio Muzzupappa, “Metodologie e strumenti informatici per una ontologia sul caso Aldo Moro”, Scritture di Storia, n. 5, pp. 359-380, luglio 2008, Edizioni Scientifiche Italiane.

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------