Semantic technologies and linked data, with a case study at the Consiglio Nazionale delle Ricerche (CNR)

Aldo Gangemi

Web evolution and resource dereferencing

Web is evolving from a web of documents (often called Web 1.0) to a web of entities (called, with subtle differences in meaning, semantic web, web of data, Web 3.0). This evolution is passing also through the availability to users to edit its contents and generate complex social networks through simple interaction paradigms (known as social web or Web 2.0). This is happening primarily thanks to a deeper exploitation of the Web architecture designed since the nineties,[1] which enables the dereferencing and linking of web resources (identified by means of a Web address), through simple communication protocols (e.g. HTTP). For example, when one writes the address (URI) http://www.cnr.it (the web address of the portal of Consiglio Nazionale delle Ricerche, CNR) in a browser, the HTTP client of the browser dereferences that address by communicating with a server at CNR, which returns a HTML page, visualized on its turn

[1] http://www.w3.org/TR/webarch.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5457

by the browser. Other web pages can be present in the visualized page, so creating a network of hypertextual links, which enables the browsing experience. This is basically the web of documents. Sometimes dereferencing is indirect, as in the case when an address represents a call to a database, e.g. when looking for one's tax data in the Agenzia delle Entrate (the Italian fiscal authority) web site: this is still the web of documents, but the documents are generated out of a query to a database, whose answer is rendered in HTML by using XML stylesheets. The case of Web 2.0 is a more sophisticated indirect dereferencing, which also enables direct changes to a database performed by users: applications such as voice protocols, email, tagging, automatic log analysis, opinion mining etc. converge in rich, customizable and dynamic HTML pages, as in the case of Facebook.

Two difficult problems: identity and semantic interoperability

Web 1.0 and 2.0 have two limitations, which actually exist in information systems since centuries ago: identity and semantic interoperability. The identity issue arises e.g. in the following example. Aldo Gangemi has different homepages (one on his institute site, ISTC-CNR, one from the wiki of his lab, STLab, one on the semanticweb.org site etc. He is also registered on many other portals providing services to the citizen, to members of associations, conference committees, commercial services etc. Moreover, he has several accounts of social web applications (e.g. Facebook, Gmail, Flickr, iTunes etc.). Even more, Aldo Gangemi is a datum within public or personal databases, like Google Scholar, DBLP etc.; that datum has identifiers that are owned specifically by those databases, gathering then a sort of ”positional” identity within one of their tables). Finally, Aldo Gangemi is cited in other web pages: articles, bibliographic references, event reports. Now the issue is: how can we know that (the physical or social person) Aldo Gangemi is the entity denoted by his homepages, registrations, accounts, database IDs, citations? Intuitively, the issue is not limited to persons, but it impacts on everything that has an identity: places, organizations, products, services, events, laws, ideas, concepts, fictional things etc. The semantic interoperability issue, besides purely system-oriented problems (e.g. different computational platforms), arises from the identity issue: if we cannot resolve the identity of something across the different sources and systems that refer to it, it gets really difficult to aggregate (i.e. assemble) and integrate (i.e. appropriately connect) the information about it. This is quite limiting when considering that the relations between something and something else can be similar within different systems: the relation between Aldo Gangemi and the email messages addressed to him, or between him and his recipients, are similar in any emailing system, but those systems assign different identities to the same persons, if any. In addition, each system works on a proprietary infrastructure: different languages, formats, protocols etc. All this makes data integration between different systems partial in the best cases.

Some traditional solutions

In the last years, a sort of cartel has emerged between commercial service providers such as Facebook, Google etc., in order to make social network data interoperable: this however concerns only data exchange that are commercially interesting for those systems, and third party applications that count on them. Database management systems use complex procedures to integrate their data when it is required: schema integration, identity resolution, data warehousing etc. Each process is typically made ad hoc on a pair of databases. Partial solutions for data integration also come from data mining or natural language processing techniques. For example, there are effective statistical approaches for named entity recognition and resolution, as well as for discovering similarity and indirect relations in data. Document annotation is an approach that comes back at least to the beginning of 20th century: a document, or part of it (paragraphs, terms) are annotated with a category or tag taken from some knowledge organization system: thesauri, classification schemes, nomenclatures, controlled vocabularies, which have developed in most scientific, library, and commercial disciplines. Exemplar cases of similar large efforts include SnoMed, ICD, MeSH (medicine), Getty thesaurus (cultural heritage), Agrovoc (agriculture) etc. Recently, annotation procedures are assisted either by computational support for manual annotation, or by automatic annotation algorithms (e.g. text classification), with variable precision.

The web of data

In 2006, Tim Berners-Lee introduced linked data, a simple and elegant method[2] to realize some practical data identity integration and interoperability on the Web. Linked data are aimed at realizing a web of data (or Entities, or Things, depending on the interest to data management, to entity linking, or to sensors and things in the physical world). Linked data is one of the technologies for the semantic web (discussed in the next section), and consists of four principles and many good practices. The principles include:

1. use web addresses (URI) as names for entities/things;

[2] http://www.w3.org/DesignIssues/LinkedData.html.

2. use HTTP URIs so that people can look up and dereference those names;
3. when someone looks up a URI, provide useful information, using the standards (RDF, RDFS, SPARQL, OWL, RIF);
4. include links to other URIs, in order to be able to discover more things and data.

Among good practices, it's useful to mention those that have best supported the Linking Open Data (LOD) bootstrap, whose state of play is visualized periodically as a cloud[3]:

- use open licenses to obtain highly reusable data;
- use non-proprietary formats (e.g. CSV instead of Excel);
- use W3C open standards (typically RDF,[4] SPARQL,[5] OWL[6]) to identify things, so that people can point at your stuff, new links can be created, better queries and more extended reasoning can be performed.

These practices also fit recommendations from the Open Data movement, and are currently adopted in many different fields, including Public Administration data[7] and are used in the integration and enrichment of data, for example for the expert finding task.[8]
The LOD Cloud contains linked data from many different domains, in particular biomedicine, cultural, multimedia, bibliographic, geographic etc. An example of the potential of linked data is shown in

[3] http://linkeddata.org.

[4] http://www.w3.org/RDF/.

[5] http://www.w3.org/TR/rdf-sparql-query/.

[6] http://www.w3.org/2004/OWL/.

[7] http://data.gov; http://data.gov.uk; http://dati.gov.it.

[8] http://data.cnr.it.

figure 1, a graph built automatically by an application (RelFinder[9]), which incrementally visualizes the relations between any two entities, provided that they have an identity on the web of data. In the figure 1, graph building starts from the entities:

```
<http://dbpedia.org/wiki/Neo-positivism>
<http://dbpedia.org/wiki/Francis_Bacon>
```

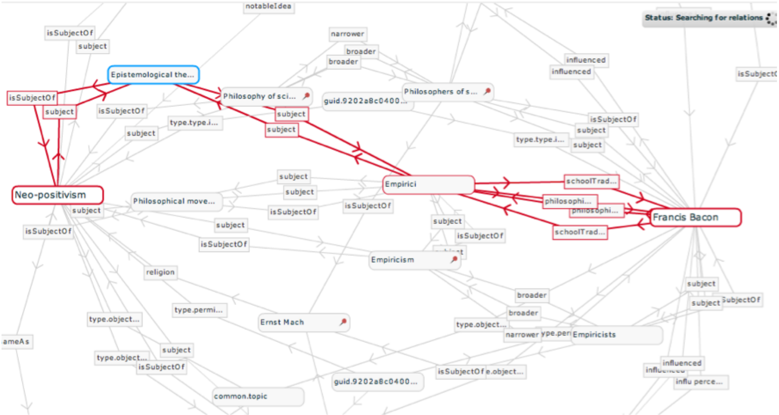


Figure 1: The emerging relations between two entities across the Linking Open Data graph.

Semantic web standards

W3C open standards, primarily RDF,[10] SPARQL[11] and OWL,[12] enable elegant and homogeneous representation of, as well as querying

[9] http://www.visualdataweb.org/relfinder.php.
[10] http://www.w3.org/RDF/.
[11] http://www.w3.org/TR/rdf-sparql-query/.
[12] http://www.w3.org/2004/OWL/.

and reasoning on, the data from most traditional data structures and data models.

RDF is based on a recursive data structure, called triple, made of a Subject, a Predicate, and an Object, analogously to the most abstract grammatical structure of Western languages, SVO (Subject-Verb-Object).

Listing 1: Sample RDF triples.

```
<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://www.w3.org/2000/01/rdf-schema#label>
  ''Aldo Gangemi''

<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  <http://www.cnr.it/ontology/cnr/personale.owl#
      UnitaDiPersonaleInterno>

<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://purl.org/dc/terms/subject>
  <http://dbpedia.org/resource/Category:Semantic_Web>
```

RDF triples can be queried via the SPARQL language.

Listing 2: Query on triples in 1

```
SELECT ?l
WHERE {
?x <http://purl.org/dc/terms/subject>
<http://dbpedia.org/resource/Category:Semantic_Web>.
?x <http://www.w3.org/2000/01/rdf-schema#label> ?l}
```

The query in listing 2 on the preceding page gets the answer:

```
l
''Aldo Gangemi''
```

Each triple contains Subjects and Objects that have a type, which is on its turn a Class, e.g.

```
<http://www.cnr.it/ontology/cnr/personale.owl#
    UnitaDiPersonaleInterno>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <http://www.w3.org/2002/07/owl#Class>
```

Each triple contains a Predicate (or Property), which together with classes forms the vocabulary (also called schema or ontology) used by a dataset. In cases where logical validation and reasoning is required, a vocabulary is defined in the OWL (Ontology Web Language) standard,[13] a language that allows the use of automated reasoners to derive logical inferences out of data structures. For example, an automated reasoner infers the inverses of existing triples, the symmetric triples, the triples holding transitively (when appropriate rules have been defined for the vocabulary) etc.
With the expressive power of OWL and SPARQL on the web of data, one can make complex questions to heterogeneous knowledge sources, e.g. in the Romal Law domain, the following natural language questions can be formalized as queries, but terms need to be mapped to appropriate entity types in RDF and OWL. In this case, underlined terms are supposed to be mapped as classes, bold-faced **terms** as properties, and *terms* in *italics* as specific entities or values:

-which *Roman Law* sources **contain** maxims **concerning** *stipulation*, cite *Ulpian*, and **include** commentaries **published in** *the last 10 years*?

[13] http://www.w3.org/2004/OWL/.

-which cases **appeared in** *Common Law* systems **contain** interpretations **relative to** contracts **analogous to** *stipulatio*?

In order to improve vocabulary quality and inference capabilities, additional axioms need to be defined (e.g. what type of entities can be analogous to what, what can be cited in what etc.). Therefore, vocabulary design requires a certain accuracy and quality control, which can be obtained by means of approaches oriented to user requirements, and with the reuse of standard vocabularies and ontology design patterns,[14] known to describe the domain of interest, and/or solving the modeling problems emerging from user requirements.

Semantic applications

Availability of large open data can provide a good motivation to develop next generation applications, which build on both existing and novel solutions, focused on the semantic paradigm: using meaning of data as a widespread organizational schema.
The life cycle of a semantic application is typically the following:

1. reengineering existing data, by producing datasets in RDF triples (data) and OWL (vocabularies);

2. linking between entities in multiple datasets, and production of new triples ;

3. extraction of new entities and triples by means of data mining and natural language processing techniques, and production of new triples;

[14] http://www.ontologydesignpatterns.org.

4. reasoning on the logical structures obtained from previous steps, and possible production of new triples (materialization);

5. publishing of datasets on appropriate platforms, for SPARQL querying;

6. presentation of enriched data to be used by web users: textual, graphic, rich snippets, explorative etc.

The life cycle reflects a multiple interpretation of the term semantics. In steps 2 and 3, we refer primarily to the linguistic semantics that is implicit in the analyzed texts; related technologies are those of text and data analysis, and aims at recognizing entities, names, terms, relations, facts, topics etc. Only once we have extracted them, we can produce new formal triples. In steps 1, 4, 5, we refer to logical (or formal) semantics of data and schemas; related technology is basically what we have mentioned in previous sections as "semantic web" (which is a mix of web science and knowledge representation). In step 6., we refer to the semantics of user interaction.

Technologies oriented to linguistic semantics allow e.g. to recognize entities in texts, and to resolve their identity with respect to known datasets. Once identity has been resolved, it is possible to enrich the dataset with known relations between that entity and other entities. For example, given the following text from the proceedings of European Union Parliament:

> The sensitivities of Northern Ireland are too important for any ill-informed bandwagoning on the International Fund for Ireland. Raytheon has been welcomed to Derry by no less than Nobel Peace Prize winners, John Hume – one of our own colleagues, and David Trimble. Raytheon will be funded by the Industrial Development Board in Northern Ireland. Not one euro nor one Irish pound from the International Fund for Ireland is going to Raytheon.

it's possible to use a "named entity recognizer" like http://www.alchemyapi.com/api/demo.html in order to recognize several names (e.g. Northern Ireland, International Fund for Ireland, Derry, John Hume etc.), whose identity can be automatically resolved by an "entity resolver" like http://wit.istc.cnr.it/stlab-tools/wikifier as e.g. the entities:

```
<http://dbpedia.org/resource/Northern_Ireland>
<http://dbpedia.org/resource/John_Hume>
<http://dbpedia.org/resource/Derry>
<http://dbpedia.org/resource/David_Trimble>
```

Once identified, we can query LOD to find out other linked entities, e.g.

```
<http://dbpedia.org/wiki/Mark_Durkan>
<http://dbpedia.org/wiki/David_Cameron>
```

A complex graph emerging from LOD triples when the four entities above are searched together for their links can be then retrieved (and e.g. visualized in the RelFinder tool[15]). For example, the following triples are found:

```
<http://dbpedia.org/wiki/Mark_Durkan> <http://dbpedia.org/
    ontology/placeOfBirth> <http://dbpedia.org/resource/Derry>

<http://dbpedia.org/resource/Derry> <http://dbpedia.org/
    ontology/country>
<http://dbpedia.org/resource/Northern_Ireland>
<http://dbpedia.org/wiki/Mark_Durkan> <http://dbpedia.org/
    ontology/predecessor> <http://dbpedia.org/resource/
    John_Hume>
```

[15] http://www.visualdataweb.org/relfinder.php.

```
<http://dbpedia.org/resource/Northern_Ireland> <http://dbpedia.
    org/ontology/leaderName>
<http://dbpedia.org/resource/Derry>
```

Figure 2 summarizes this kind of simple process: linguistic knowledge can be used to enrich (and give access to) formal knowledge, while the latter, besides generating the implicit knowledge that is implicit in triples (deductive inferences), can also be used as background knowledge by the algorithms that extract new linguistic knowledge (as applied in the entity resolution task).

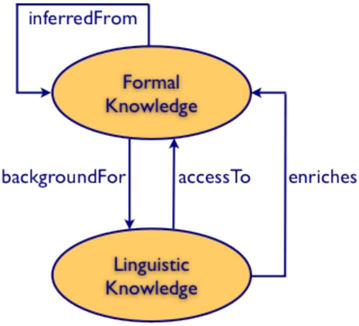


Figure 2: The hybridization cycle of linguistic and logical techniques.

Linguistic and formal interpretation, as well as the integration of related technologies, enable a hybrid methodology that empowers knowledge structuring and querying. A recent spin to that hybridization can be seen in the deep knowledge extraction task, as implemented in the FRED tool.[16] FRED deeply analyzes sentences, produces formally correct structures in RDF and OWL, and enriches the results with entity resolution. Figure 3 shows a fragment of the RDF graph produced by FRED from the EU Parliament sample

[16] http://wit.istc.cnr.it/stlab-tools/fred.

sentence.

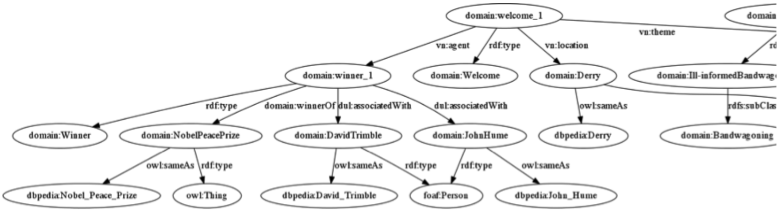


Figure 3: An excerpt from the RDF graph produced by FRED on the EU sample sentence.

A case study at Consiglio Nazionale delle Ricerche (CNR)

A practical application of data designed by following the presented semantic application lifecycle is data.cnr.it, a group of datasets that contain research data from the Italian National Research Council (CNR), enriched by automatic extraction of linguistic knowledge, automatic categorization of person and project profiles, and automated materialization of logical inferences.

The vocabularies for the datasets have been designed as modular ontologies (figure 8 on page 269), which are aligned to reference vocabularies from the LOD Cloud. Part of the taxonomy from the CNR vocabularies is shown in figure 9 on page 269. Where possible, data are linked to public entities, e.g. from DBpedia[17] or GeoNames.[18] Data can be accessed in different ways, depending on who is going to consume them. Information systems will use the data.cnr.it

[17] http://dbpedia.org.

[18] http://www.geonames.org.

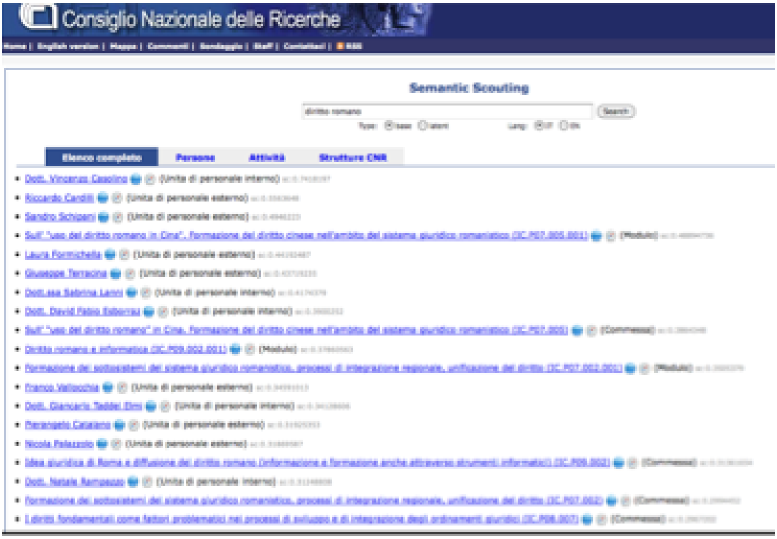
SPARQL endpoint to execute queries. Human users have also other ways to search, query, or explore data. We have designed the Semantic Scout,[19] an exploratory browser for human consumption of data: a web application supports the expert finding task on scientific competences existing at CNR, based on how data are represented in data.cnr.it datasets.

Figure 4 on the facing page shows the Semantic Scout keyword search interface for finding who works on Roman Law at CNR. Figure 5 on the next page shows hierarchical browsing in the semantic social network built from knowledge related to people found with the keyword search. Figure 6 on page 268 shows the exploratory search interface to data in a spherical space. Figure 7 on page 268 shows the exporting functionality of results obtained from the choices of the user during the exploratory browsing.

Conclusions

Semantic technologies provide a simple solution to the identity and interoperability issues, exploiting direct access to data, web standards, formal precision of schemas, and easy hybridization between techniques oriented to the extraction of linguistic semantics, and those oriented to the management of formal semantics.

[19] http://bit.ly/semanticscout.


Consiglio Nazionale delle Ricerche
Home | English version | Mappa | Commenti | Sondaggio | Staff | Contattaci | RSS

Semantic Scouting

diritto romano [Search]

Type: ● base ○ latent   Lang: ● IT ○ EN

Elenco completo | Persone | Attività | Strutture CNR

- Dott. Vincenzo Casolino (Unita di personale interno)
- Riccardo Cardilli (Unita di personale esterno)
- Sandro Schipani (Unita di personale esterno)
- Sull' "uso del diritto romano in Cina". Formazione del diritto cinese nell'ambito del sistema giuridico romanistico (IC.P07.005.001) (Modulo)
- Laura Formichella (Unita di personale esterno)
- Giuseppe Terracina (Unita di personale esterno)
- Dott.ssa Sabrina Lanni (Unita di personale interno)
- Dott. David Fabio Esborraz (Unita di personale interno)
- Sull' "uso del diritto romano" in Cina. Formazione del diritto cinese nell'ambito del sistema giuridico romanistico (IC.P07.005) (Commessa)
- Diritto romano e informatica (IC.P09.002.001) (Modulo)
- Formazione dei sottosistemi del sistema giuridico romanistico, processi di integrazione regionale, unificazione del diritto (IC.P07.002.001) (Modulo)
- Franco Vallocchia (Unita di personale esterno)
- Dott. Giancarlo Taddei Elmi (Unita di personale interno)
- Pierangelo Catalano (Unita di personale esterno)
- Nicola Palazzolo (Unita di personale esterno)
- Idea giuridica di Roma e diffusione del diritto romano (informazione e formazione anche attraverso strumenti informatici) (IC.P09.002) (Commessa)
- Dott. Natale Rampazzo (Unita di personale interno)
- Formazione dei sottosistemi del sistema giuridico romanistico, processi di integrazione regionale, unificazione del diritto (IC.P07.002) (Commessa)
- I diritti fondamentali come fattori problematici nei processi di sviluppo e di integrazione degli ordinamenti giuridici (IC.P08.007) (Commessa)

Figure 4: Search with the Semantic Scout.

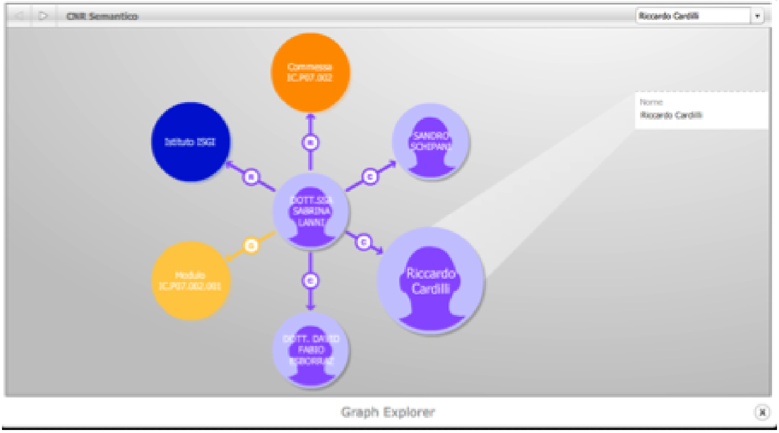
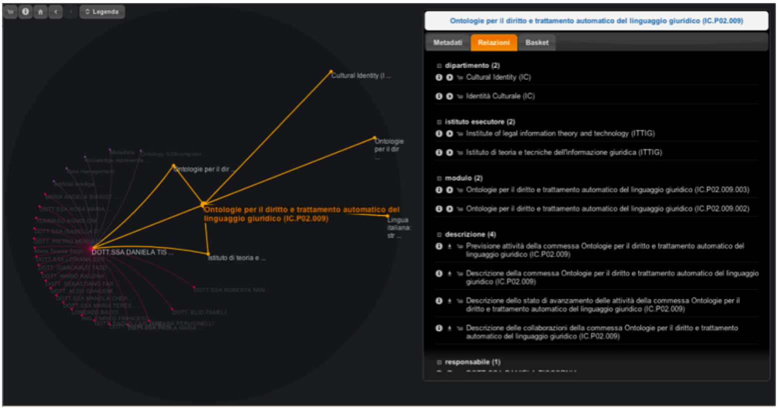


Figure 5: Semantic browsing with the Semantic Scout.



Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)

Metadati | Relazioni | Basket

⊟ dipartimento (2)

Cultural Identity (IC)

Identità Culturale (IC)

⊟ istituto esecutore (2)

Institute of legal information theory and technology (ITTIG)

Istituto di teoria e tecniche dell'informazione giuridica (ITTIG)

⊟ modulo (2)

Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009.003)

Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009.002)

⊟ descrizione (4)

Previsione attività della commessa Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)

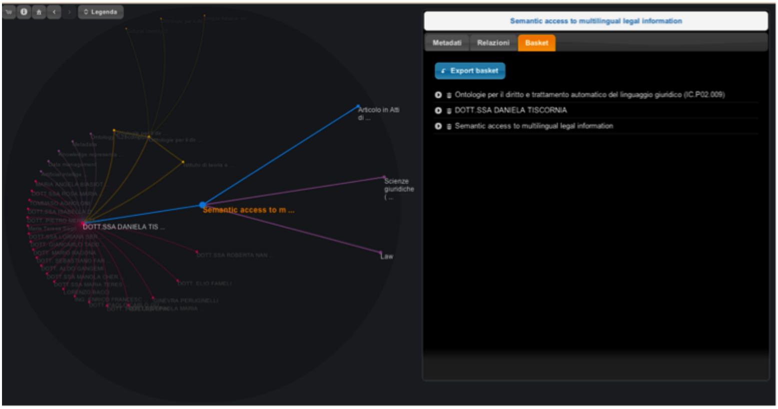
Descrizione della commessa Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)

Descrizione dello stato di avanzamento delle attività della commessa Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)

Descrizione delle collaborazioni della commessa Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)

⊟ responsabile (1)

Figure 6: Exploratory search interface of data.cnr.it data in a spherical space.



Semantic access to multilingual legal information

Metadati | Relazioni | Basket

Export basket

Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)

DOTT.SSA DANIELA TISCORNIA

Semantic access to multilingual legal information

Figure 7: Exporting results of semantic exploratory search based on user choices.

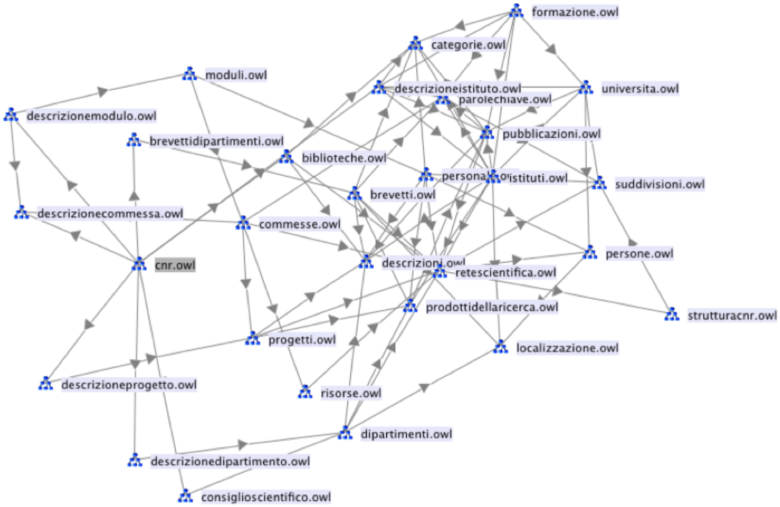


Figure 8: The network of CNR modular ontologies.

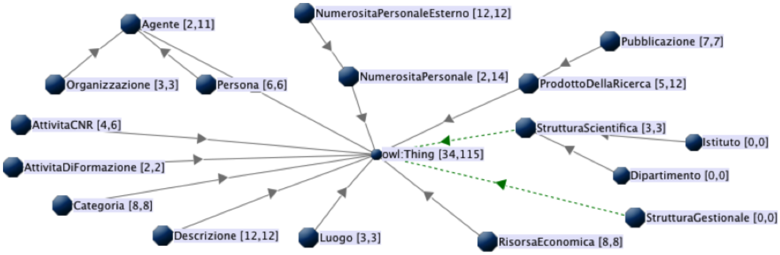


Figure 9: Taxonomy from the core ontology of CNR.

ALDO GANGEMI, Semantic Technology Lab, Institute of Cognitive Sciences and Technologies (CNR-ISTC), Roma, and Université Paris 13 – CNRS –Sorbonne Cité (France).
gangemi@loa-cnr.it

Gangemi, A. ”Semantic technologies and linked data, with a case study at the Consiglio Nazionale delle Ricerche (CNR)”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5457. DOI: 10.4403/jlis.it-5457. Web.

ABSTRACT: Governmental data are being published in many countries, providing an unprecedented opportunity to create innovative services and to increase societal awareness about administration dynamics. In particular, semantic technologies for linked data production and exploitation prove to be ideal for managing identity and interoperability of administrative entities and data. This paper presents the current state of art, and evolution scenarios of these technologies, with reference to several case studies, including two of them from the Italian context: CNR's Semantic Scout, and DigitPA's Linked Open IPA.

KEYWORDS: Library linked data; Semantic web; Governmental data; DigitPA's Linked Open IPA; CNR's Semantic Scout

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15