

DS5 – Intelligent Resource Discovery

Ontologies and Semantic Web

Outline

- What technologies ?
 - Semantic Web
 - Ontologies
- How they connect to the VO
 - Registry, UCD, ADQL
- What components can be built?
- Suggested implementations

What technologies?

- Ontologies and Semantic Web are very active research domains
- XML-based
- Strong support by W3C – definition of open standards (approved after years of discussions!)
- Interest of many commercial companies in these techniques

Ontologies

- An ontology is a formal description of a set of concepts and their relationships to each other
- Why develop one ? (Ontology 101)
 - to share common understanding of the structure of information among people or software agents
 - to enable reuse of domain knowledge
 - to make domain assumptions explicit
 - to analyze domain knowledge

Ontologies

- Ontologies rely on Description Logics
- Reasoners can make inferences
 - **Ontology:** 'Seyfert2' isA 'galaxy'
 - **Query:** astronomer wants to select 'galaxy'
 - **Reasoner:** 'Seyfert2' also matches astronomer's query
- Building an ontology is an iterative, collaborative process:
 - domain ontologies
 - task ontologies

Ontologies

- Ontologies can be edited/stored with different tools/formats:
 - DAML + OIL (Ontology Inference Layer)
 - **OWL** (Web Ontology Language)
- Quite mature editors:
 - OILed
 - Protégé

Ontologies



OWL Web Ontology Language Guide

W3C Recommendation 10 February 2004

This version:

<http://www.w3.org/TR/2004/REC-owl-guide-20040210/>

Latest version:

<http://www.w3.org/TR/owl-guide/>

Previous version:

<http://www.w3.org/TR/2003/PR-owl-guide-20031215/>

Editors:

Michael K. Smith, Electronic Data Systems, michael.smith@eds.com

Chris Welty, IBM Research, chris.welty@us.ibm.com

Deborah L. McGuinness, Stanford University, dlm@ksl.stanford.edu

Please refer to the [errata](#) for this document, which may include some normative corrections.

See also [translations](#).

Copyright © 2004 W3C® ([MIT](#), [ERCIM](#), [Keio](#)), All Rights Reserved. W3C [liability](#), [trademark](#), [document use](#) and [software licensing](#) rules apply.

Abstract

The World Wide Web as it is currently constituted resembles a poorly mapped geography. Our insight into the documents and capabilities available are based on keyword searches, abetted by clever use of document connectivity and usage patterns. The sheer mass of this data is unmanageable without powerful tool support. In order to map this terrain more precisely, computational agents require machine-readable descriptions of the content and capabilities of Web accessible resources. These descriptions

Semantic Web



Technology and Society
domain

Semantic Web
Activity

Semantic Web

*The **Semantic Web** provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework ([RDF](#)), which integrates a variety of applications using XML for syntax and URIs for naming.*

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." -- Tim Berners-Lee, James Hendler, Ora Lassila, [The Semantic Web](#), Scientific American, May 2001

On this page: [Activity Statement](#) | [Specifications](#) | [Publications](#) | [Presentations](#) | [Groups](#)

Nearby: [Advanced Development](#) | [SWAD-Europe](#) | [Simile](#) | [Semantic Web Coordination](#) | [RDF](#) | [RDF Core](#) | [RDF Data Access](#) | [Web Ontology](#) | [Best Practices and Deployment](#) | [Interest Group](#) | [Developer Tools](#)

Semantic Web

- Data shared and reused among application and community boundaries
- RDF: Resource Description Framework (XML-based, with URIs and namespaces)
- Goal:
 - communication between software agents, users
 - discover valuable applications (resource discovery) and exploitation paths (workflows)

Application in biology

- BioHaystack (IBM Watson Research)

- Integration of data from heterogeneous databases

- access protocols
- data formats
- softwares
- RDF as underlying model
- Link to myGRID

The screenshot displays the BioHaystack application interface within an Eclipse Platform. The main window shows a Genbank entry for NM_001240. The interface is divided into several sections:

- Commands:** Search using NotQuiteAblast, View in QMol.
- Active contexts:** E-mail, Media center, Organize.
- Suggested contexts (check to accept):** No suggestions.
- Starting Points:** Configure Haystack, Create collection, Development tools, E-mail, Find something, Get information into Haystack, Jot down information, Learn about Haystack and the Semantic Web, Media center, My documents, files and projects, News.
- Go to:** Input field.
- Search for:** Input field.
- Starting Points:** Navigator.

The main content area displays the Genbank entry for NM_001240, including a DNA double helix icon and the URL: <http://www.ncbi.nlm.nih.gov/locustlink/90>. The entry details include:

- Sequence Summary:** Name: None specified; click here to add. Genbank Locus: NM_001240. Length: 2568. Strandedness: single-stranded. Topology: linear. Division: PRI. Date Last Modified: 04-OCT-2003. Date Created: 19-MAR-1999. Source: Homo sapiens (human). Organism: Homo sapiens. Taxonomy: Eukaryota; Metazoa; Chordata.
- External Reference:** CDD:COG5333, GI:17978466, GeneID:904, dbSNP:2291728, dbSNP:3013, dbSNP:3741632, taxon:9606, urn:lsid:ncbi.nlm.nih.gov:lsid:i3c.org:locustlink:90, urn:lsid:ncbi.nlm.nih.gov:lsid:i3c.org:omim:6025.
- Pubmed:** 75K small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. A model of repression: CTD analogs and PIE-1 inhibit transcriptional elongation by P-TEFb. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. CDK9 has the intrinsic property to shuttle between nucleus and cytoplasm, and enhanced expression of cyclin T1 promotes its nuclear localization. Cooperative interaction between HIV-1 regulatory proteins Tat and Vpr modulates transcription of the viral genome. Cyclin K functions as a CDK9 regulatory subunit and participates in RNA polymerase II transcription. Identification of multiple cyclin subunits of human P-TEFb. Interactions between human cyclin T, Tat, and the transcription response element (TAR) are disrupted by a cysteine to tyrosine substitution found in mouse cyclin T. Isolation and characterization of the human cyclin T1 promoter. MAQ1 and 75K RNA interact with CDK9/cyclin T complexes in a transcription-dependent manner. Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter. Optimized chimeras between kinase-inactive mutant Cdk9 and truncated cyclin T1 proteins efficiently inhibit Tat transactivation and human immunodeficiency virus gene expression. P-TEFb containing cyclin K and Cdk9 can activate transcription via RNA.

The bottom of the interface shows a sequence visualization with a rainbow-colored arc representing the sequence, with markers at 856 bp, 642 bp, and 428 bp.

Application to the VO

- Astronomy is a user community which is familiar with many issues motivating the Semantic Web research:
 - Resource description in the VORegistry, with unique identifiers and XML format: mapping to RDF?
 - VOTable, with metadata and data grouped in the same document: metadata sharing is at the core of the semantic web
 - UCD and the IAU thesaurus for the semantic description will be very valuable resources for building an astronomical ontology

UCD and ontologies

- Collaboration between CDS and french IT labs (LORIA, IRIT)
- Syntactic rules for building UCDs from words
- UCD validation
- Classification

UCD and ontologies

The screenshot displays the Protégé 2.1.1 interface for editing an ontology. The title bar indicates the file path: `file:/home/seb/Onto/ucdvalidator/Ontologie/ucd.pprj, OWL Files`. The menu bar includes Project, Edit, Window, OWL, Wizards, Code, and Help. The toolbar contains various icons for file operations and editing. The main workspace is divided into several panes:

- Subclass Relationship / Asserted Hierarchy:** A tree view showing the class hierarchy. The class `a:phot.mag` is selected, and its subclasses are listed, including `a:phot.mag.bc`, `a:phot.mag.bol`, `a:phot.mag.distmod`, `a:phot.mag.reddfreesb`, `a:phot.mag.sb`, `a:phys.absorption`, `a:phys.abund`, `a:phys.acceleration`, `a:phys.albedo`, `a:phys.angmomentum`, `a:phys.at`, and `a:phys.columndensity`.
- Class Properties:** The `a:phot.mag` class is selected, showing its name and the `rdfs:comment` property.
- Annotations:** A table for class annotations with columns for Property and Value.
- Asserted Conditions:** A list of conditions for the class, categorized into Necessary & Sufficient, Necessary, and Inherited. The conditions are:
 - `a:phot` (Necessary & Sufficient)
 - `a:word` (Necessary)
 - `∃ a:need a:filtre` (Necessary)
 - `∃ a:allow a:primary-option` (Inherited, from `a:pri...`)
 - `∃ a:allow a:quantity-option` (Inherited, from `a:qua...`)
- Properties:** A list of properties for the class, including `a:allow` and `a:need`.

The interface also includes a status bar at the bottom with the text "Logic View" and "Properties View".

New components

- Build a 'core ontology', shared by different applications
 - IAU thesaurus for research domains, objects types
 - UCD for measured quantities
- Then, specific 'task ontologies'

Registry exploration

- Goal: retrieve relevant datasets from the registry, using the ontology
 - search for "late type *" would be interpreted by a reasoner, and recursively matched to all sub-categories of "late type *"
 - in case of null result, suggest request broadening (TTauri star -> young star)
- Could also be used for registry population:
 - interactive keywords refinement

Generic queries

- Goal: send queries to remote server without knowing its structure
 - use UCD-like syntax in some VOQL, possibly interpreted using contents of the registry

Computing quantities

- Goal: compute measurements using mathematical combinations of other data present in some catalogues
 - e.g. compute redshift from radial velocities
 - ontology could impose restriction like "object must be extragalactic"
- Ultimately, query against "virtual columns"
 - select $z < 1$ on catalogue not containing z