Chapter 5

# Development and Creation Software


### *Abstract*

*Chapter 5 describes software tools for use in the development of linked data. The majority of tools are aimed at software developers and often require programming and server management skills. Tools described facilitate data element creation, validation of Semantic Web structures, link creation, and linked data searching.*


The Semantic Web and its use of linked data are a work in progress, and so are the tools and programs that one would use with them. At this time, tools are aimed at developers and often require programming and server management skills. Some tools are cross-platform, but many assume a UNIX environment.

Tools can be considered in development and tend to have been created by someone in order to get a specific job done or in the course of an experimental project. They may not be maintained; therefore, it is always a good idea to look at the documentation pages for evidence of current updates.

I haven't included here all types of tools. In particular, there are database tools specifically designed to manage databases of triples. There is more information on these tools on the triplestore Wikipedia page.

*Wikipedia: Triplestore*
http://en.wikipedia.org/wiki/Triplestore.

There exist a number of lists of tools. Here are a few:

**Programming with RDF**
http://www.rdfabout.com/programming.xpd

A short list of programming environments by languages

**Developers Guide to Semantic Web Toolkits**
http://www4.wiwiss.fu-berlin.de/bizer/toolkits/index.htm

An extensive list with comparison tables that indicate additional technical details (e.g., supported databases, query languages). Most importantly, this site includes information on the number of developers working on the toolkit, the number of downloads (if known), and whether there is an active mailing list. This information gives an idea of the size of the project and whether it has an active user community for support and sharing.

**Semantic Web Development Tools**
http://www.w3.org/2001/sw/wiki/Tools

The W3C's Semantic Web wiki provides information on about 275 tools divided into about two dozen categories (e.g., RDF generator, SPARQL endpoints, validator). Because it is a wiki, you can easily see the history of the entries and therefore get a quick idea of the freshness of the information.

**Sweet Tools**
http://www.mkbergman.com/sweet-tools

Mike Bergman has compiled a list of about one thousand tools in a small number of categories. The list is searchable by category or keyword in the name or description. This is a handy list if you are looking for

a tool in a particular programming language because you can filter retrievals by language. Be sure to check the webpages of the tool to get an idea of its level of use and community support.

**Library Linked Data Tool for Developers**
http://www.w3.org/2001/sw/wiki/LLDtools

This list of tools was developed during the work of the W3C Library Linked Data Incubator Group. If kept up-to-date, it could become the source of library-specific linked data tools.

Most of these lists are limited to tools that are open source and free. There are some commercial tools, but they can be expensive and are designed for enterprise-level development.

## Metadata Definition and Development

All data elements and controlled lists that you wish to employ in linked data must be defined using Semantic Web standards. Defining these is not unlike creating an XML schema; in particular it is as complex, tedious, and error-prone as any schema production. For this reason good tools can be very welcome. Below are just a few; please check the W3C Semantic Web website for other news and for new ones as they are developed.

### *Protégé*

- **Name:** Protégé
- **Creator:** Stanford Center for Biomedical Informatics Research
- **URL:** http://protege.stanford.edu
- **Created:** 2004
- **Updated:** October 4, 2011

Protégé (figure 5.1) is the software most frequently cited for the creation and editing of metadata definition files using OWL. It was developed at the Stanford Center for Biomedical Informatics Research and was supported by grants by a number of research and scientific agencies including National Institutes of Health and the National Library of Medicine. Protégé provides an interface that displays the relationships between classes and subclasses and between elements. It allows for editing of metadata elements without the barrier of code. It is not unlike some commonly used XML editors in its functionality, but instead of XML the underlying data is in OWL and RDF.

### *Open Metadata Registry—Elements*

- **Name:** Open Metadata Registry
- **Creators:** Diane Hillmann, Jon Phipps
- **URL:** http://metadataregistry.org
- **Created:** 2005
- **Updated:** 2011

The Open Metadata Registry (OMR) is a website that allows you to create element sets simply by filling in a template (figure 5.2). The underlying software then produces valid RDF that can be downloaded or used directly from the site. Some of the capabilities of OWL are not available in the OMR, but its user-friendliness makes it suitable for beginners. There is a sandbox where you can experiment with linked data element creation without needing to obtain a domain name beforehand. OMR has been used to define and publish metadata elements from the Resource Description and Access standard, as well as FRBR, FRAD, and ISBD.

The OMR treats all updates to elements as new versions and, like a wiki, all previous versions of an element or its descriptive text can be viewed or even used in programs. Each entry can have a status that indicates whether the term is new and provisional or has been given production status. This makes it possible to add terms so that they are visible to the relevant user community for discussion before they become part of the standard element set.

The OMR supports RDF, some OWL properties, and also SKOS (in its Vocabularies section). This includes support for multiple languages for the display of elements and vocabulary terms.

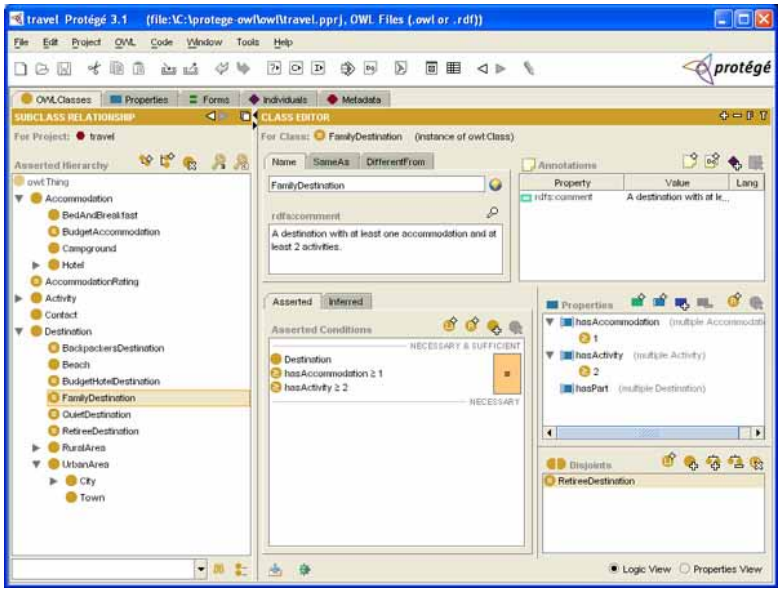


**Figure 5.1**
Sample Protégé page

### *OpenVocab*

- **Name:** OpenVocab
- **Creator:** Talis, Ian Davis
- **URL:** http://open.vocab.org
- **Updated:** 2011

OpenVocab is a simple, online, form-based input tool for the publication of Semantic Web metadata elements (called `properties`). To make good use of this site one must understand the basic Semantic Web concepts around element definition, such as domains and ranges. OpenVocab registers all terms under its own domain name, which makes this site an easy way to add a few needed elements without having to manage a domain name and an online presence for the resolution of the URIs.

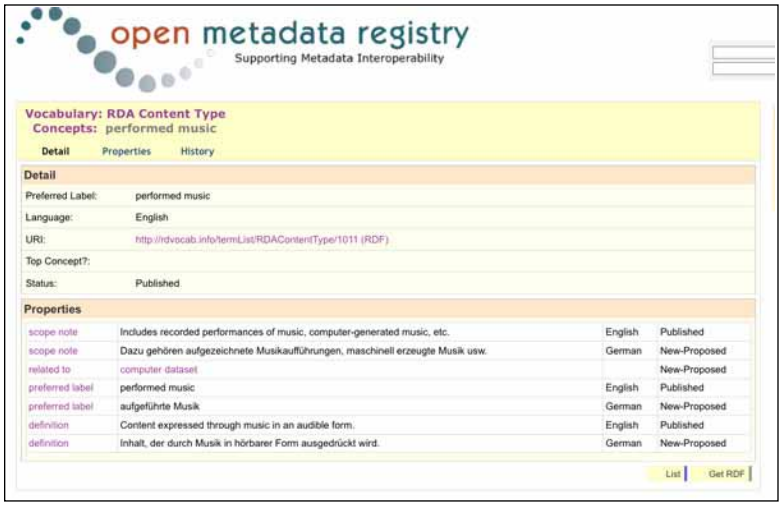


**Figure 5.2**
Sample OMR page

## List and Thesaurus Development

### *PoolParty*

- **Name:** PoolParty
- **Creator:** Semantic Web Company
- **URL:** http://poolparty.biz
- **Updated:** 2012

PoolParty (figure 5.3) is a commercial application that helps users create and manage thesauri. It includes automatic indexing of full text documents using controlled vocabularies and a database that is created from automatic and human indexers. PoolParty has educational licensing, and the product includes a plug-in to WordPress, the commonly used blogging software.

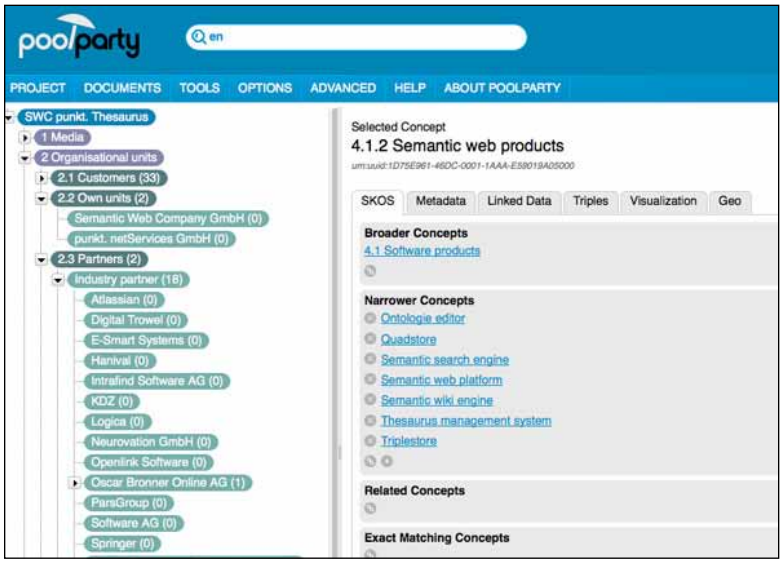


**Figure 5.3**
Sample PoolParty page

### *Open Metadata Registry—Vocabularies*

- **Name:** Open Metadata Registry
- **Creators:** Diane Hillmann, Jon Phipps
- **URL:** http://metadataregistry.org
- **Created:** 2005
- **Updated:** 2011

The Open Metadata Registry can be used to create controlled lists that are then made available by the underlying software as SKOS vocabularies. The OMR hosts the nearly seventy controlled lists from the RDA standard.

## General Programming

Jena and Pellet are commonly used programming tools that are under continuing development, although there are others to be found in the lists above. Both are Java-based. Redland is also included here as an example of a tool that uses the C libraries and is currently supported.

### *Jena*

- **Name:** Jena Platform
- **Creator:** HP Labs, Talis
- **URL:** http://incubator.apache.org/jena

- **Created:** 2004
- **Updated:** 2011

Jena is an open source programming toolkit for developing Semantic Web applications in Java. Jena includes modules for reading and writing RDF graphs, managing files of linked data, and navigating the links in the graphs. It assumes that your data has been described in OWL and/or RDFs and that your instance data follows that description. Jena functions primarily through an RDF API and makes use of SPARQL querying of RDF data. Jena was developed at HP Labs and is now an Apache project.

### *Pellet*

- **Name:** Pellet
- **Creator:** Clark and Parsia
- **URL:** http://clarkparsia.com/pellet
- **Created:** 2008
- **Updated:** 2011

Pellet is the most commonly used "reasoner" for RDF. A reasoner is software that acts on the relationships and constraints that have been defined for the metadata in OWL. Examples of constraints are property/subproperty and class/subclass relationships, datatype declarations, declarations that classes or properties are disjoint, and vocabulary relationships like OWL's "same as." Constraints in RDF are different from what one may be accustomed to from database definitions or XML, but they are powerful because they can operate over chains of relationships. Pellet can be integrated with programming environments including Jena.

### *Redland RDF Libraries*

- **Name:** Redland RDF Libraries
- **Creator:** Dave Beckett
- **URL:** http://librdf.org
- **Created:** 2000
- **Updated:** 2011

Similar to Jena in functionality, Redland is a set of C libraries.

## Viewers

There are those among us who happily read code, but it can be difficult for most of us to form a mental picture of the contents of a complex schema unless it is given a clear and logical display. You will find that some authors of term lists and schemas already provide them in an organized display. For those that do not, there are some options. Note that the programming toolkits, such as Protégé, can also be used as viewers.

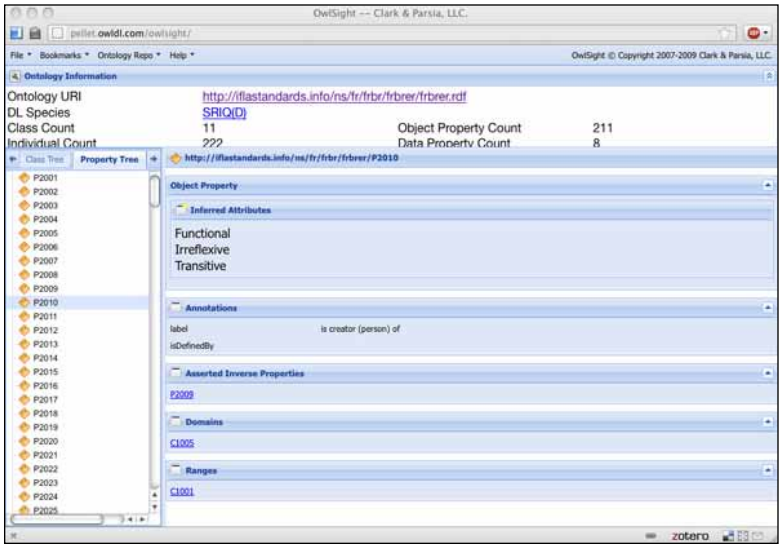


**Figure 5.4**
Sample OwlSight page

### *Manchester Ontology Browser*

- **Name:** Manchester Ontology Browser
- **Creator:** University of Manchester
- **URL:** http://owl.cs.manchester.ac.uk/browser

Created at the University of Manchester as part of the CO-ODE project, the Manchester Ontology Browser is widely used. The code is available if you wish to install it at your own site. Alternately, there is an HTML interface that can display an OWL document from a URI (presuming that the URI links to the OWL code) or as text pasted into a form for quick viewing.

### *OwlSight*

- **Name:** OwlSight
- **Creator:** Clark and Parsia
- **URL:** http://pellet.owldl.com/ontology-browser/
- **Updated:** 2008

OwlSight (figure 5.4) is a program that displays OWL ontologies inside your browser. It places a menu at the top of the browser page that allows you to input a URI or to input the OWL code into a form. It can expand or close any hierarchies and uses Pellet to display information about constraints that are defined in the ontology.

## Converters and Validators

There is a huge amount of data of all types that already exists in nonlinked data formats. In many cases the data can be converted to linked data, even though the converted data may be less than ideal since it was not designed originally for the Semantic Web. Many of the datasets in the Web of linked data today have been converted in this way. Two data converters highlighted here convert from a standard database to RDF and from natural language to RDF. The W3C maintains a list of converters. These include converters from BibTex, e-mail, Excel, Flickr data, iCalendar, MARC, OAI-PMH, and vCard. Note that the MARC conversion is actually a conversion from MARC to MODS and then to RDF.

*W3C: Converter to RDF*
http://www.w3.org/wiki/ConverterToRdf

#### *Stuff 2 RDF*

- **Name:** Stuff 2 RDF
- **Creator:** Christopher Gutteridge, University of Southampton
- **URL:** http://graphite.ecs.soton.ac.uk/stuff2rdf
- **Created:** 2010
- **Updated:** 2010

Stuff 2 RDF is software that provides conversions between the various serializations of RDF (RDF/XML, Turtle, N3, etc.). These conversions are interesting as learning and testing tools, and I find that they also are useful for formatting downloaded data that is hard to read. Most of the validator services also provide conversion between serializations. Stuff 2 RDF can also convert from CSV or Excel to RDF and is available for download if you wish to make heavy use of it. It is a PHP program.

#### *W3C RDF Validator*

- **Name:** W3C RDF Validator
- **Creator:** W3C
- **URL:** http://www.w3.org/RDF/Validator
- **Updated:** 2007

The W3C RDF Validator webpage allows you to validate RDF/XML either by pasting it into a form or by providing the URI of the RDF. It will validate the RDF and optionally provide a graph (in a variety of formats) and an expression of the RDF as triples. For simple RDF this is an easy way to get a downloadable graph, although graphs quickly become unwieldy as data becomes more complex.

#### *D2R Server*

- **Name:** D2R Server
- **Creators:** Chris Bizer, Richard Cyganiak
- **URL:** http://www4.wiwiss.fu-berlin.de/bizer/d2r-server
- **Created:** 2006
- **Updated:** 2010

D2R Server is a tool for publishing relational databases on the Semantic Web by converting the relational database to linked data. Definitely a developer tool, D2R Server sits between a database in a standard database management system that is compatible with a common version of SQL and the Semantic Web. D2R Server provides an interface into the non-RDF data, thus turning nearly any database into linked data.

## Discovery of Metadata and Services

One of the great difficulties in working in linked data today is that the linked data space is not well indexed, so discovery of available metadata and services is rather hit-or-miss. It reminds me of the early days of the Internet before search services, when you operated with a list of well-known sites that were sources of files of software.

#### *Swoogle*

- **Name:** Swoogle
- **Creator:** eBiquity (University of Maryland, Baltimore County)
- **URL:** http://swoogle.umbc.edu
- **Created:** 2007
- **Updated:** daily

Swoogle is a search engine primarily for Semantic Web ontologies, that is, for documents that define elements and vocabularies. It also searches for documents that include RDF statements, although this is a much larger search space and may not be fully covered. Swoogle can search for elements from a particular ontology, for RDF found in documents, and for individual terms (useful for linking) and can search across ontologies. As of December 2011, Swoogle was indexing almost twelve million URLs and over one billion triples.

## Link Creation and Term Mapping

As more data is created for use on the Web, there is a need to make useful connections between data from different sources. In cases where two communities use the same data element, like `dc:title`, linking can be done directly. However, different communities have different metadata needs, and they define metadata specific to

their purpose. In the pre-RDF world, one brings together data that has been defined separately using cross-walks or data element maps. With RDF you accomplish the same thing (actually something better) by creating meaningful links between elements. For example, in library data there are many types of title: title proper, uniform title, parallel title. To allow library data to be used along with simpler bibliographic data based on Dublin Core, you would create a link from the library data elements for title to the Dublin Core title. This link could be of a type that indicates that the library titles are more specific than the Dublin Core title element, or it could simply say that each of the library elements for title is similar to the Dublin Core element. This link then allows an application gathering titles to treat the Dublin Core titles and the library titles as equivalent for the purposes of that application without altering the data elements themselves. The links form connections that are not unlike "broader term" references in library authority data. The original data remains intact, but it is possible to operate on the broader concept.

Note that links are made between the defined data elements, not in the data itself, and therefore are valid for all data created with that element set.

Some links between data elements can be created algorithmically, but to make the Web of linked data intellectually rich, much linking will need to be done by humans who understand the concepts in the element sets. Therefore we need programs that will help us make those links, and some already exist.

### *Snoggle*

- **Name:** Snoggle
- **Creators:** Dave Kolas and others
- **URL:** http://snoggle.semwebcentral.org
- **Created:** 2007
- **Updated:** 2007

Snoggle is a graphical tool that allows you to make mappings between terms, including terms from different vocabularies. The example given on the website is mapping an element for "Person" in one vocabulary to an element for "Employee" in another. Snoggle makes the connections in an editor that looks like the graphs that are usually employed to illustrate linked data, with boxes as "things" and arrows as relationships. New relationships between things are created by connecting boxes with arrows. Snoggle then renders this graph as code. Snoggle is a Java application, but at the time of this writing the authors were not making any guarantees that it would indeed run on all platforms.

### *Silk*

- **Name:** Silk Link Discovery Framework
- **Creators:** Robert Isele and others, Freie Universität Berlin
- **URL:** http://www4.wiwiss.fu-berlin.de/bizer/silk
- **Created:** February 1, 2009
- **Updated:** November 17, 2011

The Silk site is subtitled, "A link discovery framework for the Web of data." It is a developer tool built of programs and a linking language that allow data providers to programmatically discover linking vocabularies on the Web. Silk can also be used in a local application to generate links and has features like caching and data distribution to gain greater efficiency in link usage.

## Searching

Searches against linked data are done using SPARQL, the query language that operates on triples. Most sites with linked data will provide a basic form connected to their *SPARQL endpoint,* the API that accepts queries in SPARQL and returns results. There are some sites that provide a SPARQL endpoint for more than one linked dataset. These sites generally accept a fully formed SPARQL query, so you must be versed in that language to make use of them.

### *Virtuoso SPARQL Query Editor*

- **Name:** Virtuoso SPARQL Query Editor
- **Creator:** OpenLink Software
- **URL:** http://virtuoso.openlinksw.com

Virtuoso is a multiformat data storage and retrieval system. The Virtuoso SPARQL Query Editor allows one to run queries against a linked dataset, as well as queries that cross sets of linked data. It natively understands a large number of namespace prefixes, so it isn't necessary to define those namespaces in the query. The service is available as a Web format at http://demo.openlinksw.com/sparql.

### *Semantic Web Search Engine (SWSE)*

- **Name:** Semantic Web Search Engine
- **Creator:** Deri
- **URL:** http://swse.deri.org
- **Updated:** 2010

The Digital Enterprise Research Institute, National University of Ireland, Galway has developed a Semantic Web search engine that provides a single search box for keyword searching. SWSE is experimental in nature and cannot deliver all results, but the results that it does deliver can give you can idea of what data exists on the Semantic Web. For example, a search on Ireland retrieves results from sources GeoNames and

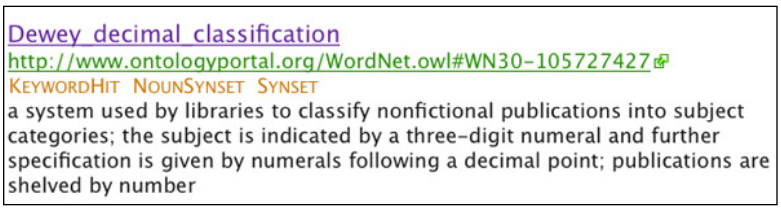


**Figure 5.5**
A single search result from SWSE

DBpedia as well as databases of airport information and a music database. The search results look very much like the results from a Google search, with titles, links, and sometimes snippets (see figure 5.5).

## Data Creation

As I mentioned earlier in this report, most linked data today is being created programmatically by converting existing data to linked data on a large scale. There are efforts to make linked data creation available to the rest of us. Some of these operate on full text; others are form-based and accept data entry from humans.

### *Open Calais*

- **Name:** Open Calais
- **Creator:** Thomson Reuters
- **URL:** http://www.opencalais.com
- **Updated:** 2008

Open Calais is software from Thomson Reuters that operates in the original spirit of the Semantic Web by identifying data in natural language documents and coding them as linked data. The software is impressive in its ability to identify people, places, events, and a variety of facts like position in a company. The document viewer allows you to enter text into a form and see the Open Calais analysis. Although the resulting RDF is specific to the Calais vocabulary (which therefore limits linking), it is an interesting proof of concept for the semantic analysis of plain text. The software can plug into Drupal or WordPress to add Semantic Web functionality to blogs. For content providers with a large document archive, Open Calais can tag entire archives to create searchable metadata.

Open Calais has an active user community with discussion forums where users can share information.

### *DBpedia Spotlight*

- **Name:** DBpedia Spotlight
- **Creator:** DBpedia
- **URL:** http://dbpedia.org/spotlight
- **Updated:** 2011

DBpedia Spotlight is an application that helps you annotate texts with DBpedia elements and term lists. You can download the software or use an online form to provide the text you wish Spotlight to annotate. For each term that matches a concept or term in DBpedia, the tool gives a list of possible matches. For example, the term *library* in a text gives a number of proper names including the term, but also offers a choice between *library* and *library (computing)*. You can narrow down your choices by selecting one or more categories from DBpedia, Freebase, or Schema.org.

### *FOAF-a-Matic*

- **Name:** FOAF-a-Matic
- **Creator:** Leigh Dodds
- **URL:** http://www.ldodds.com/foaf/foaf-a-matic
- **Created:** 2003

FOAF-a-Matic is an online tool that anyone can use to create a FOAF profile. It is a simple HTML form that prompts for the basic FOAF information, like name, e-mail address, website, and so on. It returns a FOAF profile in RDF that can be placed on the Web. You can see the profile I created at `http://kcoyle.net/foaf.rdf`. While limited to this one function, FOAF-a-Matic proves that the creation of linked data is something that anyone can do with the right software.