Information retrieval in Current Research Information Systems.

Andrei S. Lopatenko
Vienna University of Technology
Gusshausstrasse 28 / E015
A-1040 Vienna, Austria
+43(1)58801-41573
andrei@derpi.tuwien.ac.at

ABSTRACT

In this paper we describe the functional requirements for research information systems and problems which arise in the development of such a system. Here is shown which problems could be solved by using knowledge markup technologies. In this article one DAML + OIL ontology for Research Information System is offered. The already developed ontologies for research analyzed and compared. The architecture based on knowledge markup for collecting research data and providing access to it is described. It is shown how RDF Query Facilities can be used for information retrieval about research data.

Keywords

Current Research Information System, Ontology, Information Retrieval, DAML, RDF, Knowledge Markup, scientific publishing

INTRODUCTION

Information about research results, projects, publications, organizations, researchers and so on published on the web play a more and more pervasive role in modern research. The increasing dependence of modern research on already achieved research results requires to have ability to retrieve research information in a more efficient way.

Information overload by the exponential rise of amount of information makes it difficult for researchers to find relevant information. To solve these problems a number of Current Research Information Systems (CRIS) is being developed.

But in most cases such systems do not solve their task of providing complete and actual information with a minimum of information noise. This is one reason that researchers are not prone to publish results about their research via information systems. Publishing usually is limited to researcher's or project's web pages.

To provide actual and complete information for interested persons, information from research web pages also should be included into information retrieval operations.

Usually researchers' or policy-makers' demands for research information is not limited to information from one single system. Research information in any science or technology area is scattered among a number of heterogeneous information systems. There is a strong need to gather information or to point researchers to systems where information can be found. It is very important to know if the gathered research information is actual and complete.

We are developing the AURIS-MM information system (Austrian Research Information System - MultiMedia enhanced) to provide research information to interested consumers in a more attractive way. The system is being developed coming from the existing AURIS (Austrian Research Information System) and FoDok-Online (Research Documentation of Vienna University of Technology).

Our experience and newest web technologies showed us that centralized database systems are very efficient but not the best solution to provide access to research data due to a widespread distribution of the research data over the web.

The new version of AURIS-MM is based on Semantic Web technologies

RDF – Resource Description Framework www.w3.org/rdf

RDFS – Resource Description Framework Schema www.w3.org/rdf

DAML + OIL (DARPA Agent Markup Language + Ontology Inference Layer) www.daml.org

ONTOLOGY DEVELOPMENT FOR SCIENCE

Some efforts already were done to provide to researchers, industry, policy-makers efficient information access to research data from some sectors of science and access to research limited to organization (university research information systems), or limited to geographical boundaries (national networks, ERGO[ERGO] – European Research Gateways Online) .

The development and use of such systems has shown that it is very hard to collect complete and up-to-date data about research in a sector or in an organization like a university in a central system due to the huge effort of periodically copying or keying in the data by the providers.

Due to the fact that already huge amount of data is provided on internet web pages of projects, researchers, universities, it is hard to get researchers provide their data once more into a centralized system.

Full text search engines like Google (http://www.google.com) index among others also pages with research information. But they can not limit search to trusted data, understand context of the page and provide search based on meaning of the data.

One of the possible ways to collect data about research is the page annotation. Knowledge can be annotated on the page in such a way that automatic tools can collect and understand it [BL-2001, Hend-2001, Erd-2001]

Ontologies make possible that software agents can understand knowledge which is marked up [Staab-2001, SWA] . The benefits of ontologies and Semantic Web use for scientific publishing were described at [Lee-2001]

Some effort is already done to develop markups for scientific data.

SHOE[Hefl-99, SHOE] is a small extension to HTML which allows to annotate some knowledge about web page content. SHOE is a very simple language for declaring ontology, defining classification, relationship, inference rules, categories, etc. SHOE was developed in the Department of Computer Science, University of Maryland. SHOE specification, tools, SHOE ontology in plain text and DAML, examples are accessible at the SHOE home page

Several ontologies for university and research data were developed for SHOE. There are the University ontology and the Computer Science Department ontology (http://www.cs.umd.edu/projects/plus/SHOE/onts/index.html).

OIL (Ontology Inference Layer) [OIL, Fens-2000] - "is a proposal for a web-based representation and inference layer for ontologies, which combines the widely used modeling primitives from frame-based languages with the formal semantics and reasoning services provided by description logics. It is compatible with RDF Schema (RDFS), and includes a precise semantics for describing term meanings (and thus also for describing implied information)." OIL was sponsored by the European Community via the IST projects Ibrow and On-To-Knowledge.

In the OIL for research data there were developed SWRC (Semantic Web Research Community Ontology) (http://ontobroker.semanticweb.org/ontologies/swrc-onto-2000-09-10.oil) and KA2 (Ontology of Knowledge Acquisition community) .

DAML (DARPA Agent Markup Language)[DAML] - ontology markup language, was developed as an extension to RDF and RDFS. DAML allows to specify ontologies and markup pages for automatic knowledge extraction. The last version of DAML is named DAML + OIL. DAML specifications, examples, tools, ontologies are published at DAML home page.

Several ontologies for research information are developed in DAML. Among them: DAML version of SHOE University ontology (http://www.cs.umd.edu/projects/plus/DAML/onts/univ1.0.daml), SWRC (Semantic Web Research Community) ontology (http://www.semanticweb.org/ontologies/swrc-onto-2000-09-10.daml), homework assignment ontology (http://www.ksl.stanford.edu/projects/DAML/ksl-daml-desc.daml).

A more complete list of ontologies for research data as well as for metadata standards, thesauri and system architectures please find at the European Current Research Information Systems Platform home page (http://www.eurocris.org) and at Andrei Lopatenko's Resourse Guide to Metadata for Science, Research and Technology (http://derpi.tuwien.ac.at/~andrei/Metadata_Science.htm)

ONTOLOGY

So, the main goal of our ontology development was to develop an ontology which will help users of research information to retrieve relevant information.

The Primary use cases of information retrieval for CRIS are [Jeff-98, CERIF-2000, Lind-2000, Aks-2000]

- Retrieving information about research results by researchers or students for results reuse. The estimation of research results.
- Seeking collaborators which can take part in research projects as partners, sell their expertise, results and intellectual rights
- Finding facilities and equipment which can be used for research
- Assess and access to Research and Development capabilities by policymakers
- Finding ongoing research and technology activities and results of projects by users in commerce and industry
- Finding the sponsors for a new research project

The ontology should contain terms already known to developers of Current Research Information system to make it more easy to integrate new infrastructure with the old ones.

There are not a lot of metadata standard for science. The review of them have been done at [Grot-98,Lop-01].

Math-Net developed a metadata format based on Dublin Core and RDF Schema for mark up of knowledge about content of researchers and institutes pages[MathNet]. Math-

Net metadata set allows describe Researchers/Research groups/organizations, projects, results, events, publications.

In our ontology development we decided to use CERIF-2000 metadata standard (Common European Research Information Format)[CERIF-2000]

According to CERIF documents [CERIF] “CERIF 2000 is a set of guidelines meant for everyone dealing with research information systems. The CERIF 2000 guidelines are developed by a group of experts from the EU Member States and Associated Member states, under the co-ordination of the European Commission.”

Now CERIF 2000 is used by several groups of developers and researcher in different EU member states, it is proved and stable. Also different group of developers are well-acquainted with CERIF-2000 what will let make a process of ontology more easy

Despite excellence of CERIF as metadata format for research, there are certain lacks in CERIF in description some types of research information resources. In development of our ontology we decided to enrich it with terms, slots from some other ontologies, to make it more suitable for research information retrieval.

In the next table is provided comparison of enriched CERIF ontology with a few already developed ontologies (they were described earlier)

Table 1. Comparison of selected ontologies for science

CERIF 2000	Math-Net ontology	SWRC Semantic Web Research Community	University Ontology
Person			
Yes. Not classified in CERIF	Yes.	Advanced hierarchy suitable for research and education	Advanced Hierarchy suitable for research and education
Project			
Not classified in CERIF	Yes	Yes. Classified.	No
Organization			
Yes. Classified	Yes	Close to CERIF classification	Only educational
Publication			
Advanced classification which can server to research and educational IS	Yes	Close to CERIF classification of publications. Grey literature is not included	Close to CERIF classification of publications. Grey literature is not included
Event			
Yes. Vary basic classification	Conferences	Yes. Very close to CERIF	Conference
Equipment			
Yes.	No	No	No
Patent			
Patent	No	No	No
Product/Research result			
Product	Only software and software libraries	Yes	Only software product
Expertise skill/Research topic			
Expertise skill	Yes Subject Value	Research Topic	No
Multimedia elements			
Multimedia elements No	No	No	No
Sites/pages			
No	Yes	No	No

After the comparative analysis of the CERIF ontology, selected ontologies and some research information systems, it was recognized that CERIF ontology could be a base technology due to richness of base terms and relevance to RIS. But in some areas there are certain lacks in CERIF. Enriching CERIF ontology with terms from other ontologies can be useful for research information systems

The primitive units of the CERIF ontology are *Person*, *Project*, *Organization Unit*, *Publication*, *Event*, *Site* (Internet service/page), *Equipment*, *Result*, *Multimedia element*, *Research topic* (Expertise skill).

Research results which can be reused might be described in *publications* (articles, thesis, technical reports, etc.). *Research results* might be described precisely (*Research result* or *Product*). They can be presented by advanced presentation techniques - *Multimedia element*, which maybe video, images, drawing, diagrams, MS PowerPoint presentations.

Research results are results of research *projects*, invented by *persons(researchers, students),* in *organization units* (universities, labs, institutes, departments). Information about *expertise skills* of persons, organizations can be also significant for estimation of research results.

Some research results are patented and valuable information about them can be stored in *patents.*

To make search of research results more easy information about any entity can be classified by *research topics.*

To find a partner. Partner might be an *organization unit* or *person,* which has relevant for partner seeker *research results* and *experience.* Information about results and experience of partner can be extracted from its *publications,* description of the *projects.*

Information about organization units, publications, results, projects, persons can be stored on the *sites.* No research information system store all relevant information. So users need to know about other information system, which can help in search research results, partners.

To help user find information, data about other research data relevant *sites* and internet services should be provided to user.

Research may need *equipment* or *facilities.* Information about those entities also should be retrievable and searchable.

Table 2. Research Information Ontology terms

- Organization unit
  - Enterprise
  - Higher Education Establishment
    - University
    - Faculty
    - Institute
  - International organization
  - Joint Research Center
  - Non-research private non-profit
  - Non-research public sector
  - Private research center
  - Private non-profit research center
  - Public research center
  - Laboratory
  - Research Group
- Project
  - European project
  - Fundamental research project
  - Applied research project
  - Financed by official bodies project
- Person
  - Researcher
  - Student
- Product/Research result
  - Fundamental
  - Applied
  - Software
    - Software library
    - Information system
  - Compound
  - Process
  - Technology
  - Algorithm
  - Documentation
    - Proposal
- Event
  - Conference
  - Cultural event
  - Exhibition
  - Political event
  - Sport event
  - Trade fair
  - Workshop
- Publication
  - Abstract
  - Book
  - Conference paper
  - Conference proceedings
  - Dissertation
  - Guideline
  - Index
  - Journal article
  - Lecture
  - Multimedia
  - Patent
  - Report
  - Review

Equipment

Multimedia element

Audio

AudioVisual

DataForMultimedia(data for scientific software modules, such as GIS)

ExecutableFile(which visualize information, process, etc)

Flash

Image

RealMedia

ShockWave

Slide presentation

Video

Site

Organization's site

Project's site

Personal home page

Publication on the web

List of the publications

Reference page

Information system

Library (access to articles)

Research Information System (access to research data- projects, persons, organizations)

The complete ontology and set of terms are presented at http://derpi.tuwien.ac.at/~andrei/Metadata_Science.htm.

For ontology development CERIF-2000 Guidelines and Subject Index recommendations were used, as well Multimedia Ontology [Hunt-2001] and science and university ontologies mentioned early.

As a guidelines for ontology development we used [Noy-2001, Noy-G]

INFORMATIONAL RETRIEVAL ARCHITECTURE

The research data for retrieval should be collected, analyzed. To make possible analysis and understanding of meaning of data by software, they should be published in format understandable by software agent or annotated. Then annotations should be collected, analyzed, if it is considered necessary, they should also be transformed into one model/format. During search operation queries and data should be processed by search engines and response should be send to information consumers

So the process of information retrieval consists of

1. knowledge markup (by researcher)
2. harvesting marked-up knowledge by crawlers or software agents
3. transforming harvested data into formats appropriate for metadata repository/search engines
4. loaded into repository
5. retrieved by search engines according to users request

WEB PAGE ANNOTATION

So the ontology can serve for understanding meaning of data. But to make data understandable by software agents, they should be provided in a format, which agent can parse

A number of annotation tools are described in [Staab-2001].

For page annotation we use two tools: OntoMat and AURIS-MM metadata generating facilities.

OntoMat [OntoMat] is a user-friendly interactive webpage annotation tool. It includes web browser and ontology browser. Ontology browser supports DAML + OIL ontology exploration. Web browser supports web browsing, highlighting parts of the web pages and creating annotations based on highlighted part of the pages. To annotate the web page researcher needs to open web page in the browser, then open ontology from provided by project URL. Then the researcher can crate annotation highlighting regions of the page and describing them in ontology browser according to the ontology terms, relation and attributes. OntoMat automatically creates RDF annotation and new web page with included RDF annotation. The annotated web pages can be published on the web instead of annotated.

AURIS-MM metadata generating facilities generated RDF description of the data from AURIS-MM Relational database.

To create annotated web page, researcher needs input data about his research (projects, publications, etc) into AURIS-MM, and the use metadata generating facility just by pressing buttons. Generated RDF file then can be published on the web directly, or can be embedded into the web page.

The generated RDF file for the object has a persistent location in the AURIS-MM, which can be used as an identifier for that object. This is very important because information about the one object can be asserted on different pages. OntoMat supports only annotation and does not generate persistent URLs, because it is annotation tool.

Currently AURIS-MM does not support any ontology for semantic annotation as OntoMat does. But it supports vocabularies and thesaurus for advanced annotations, also it supports workflows and allows to re-use already inputted data.

Fig. Annotation of the page

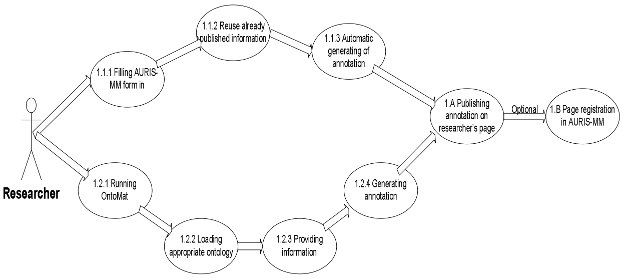
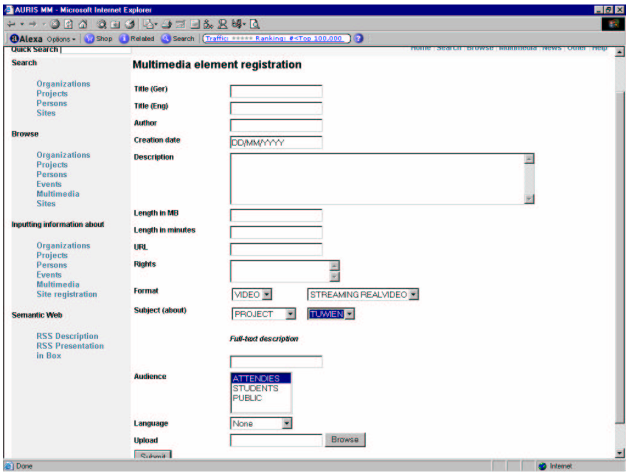

1.1.2 Reuse already published information
1.1.3 Automatic generating of annotation
1.1.1 Filling AURIS-MM form in
1.A Publishing annotation on researcher's page
Optional
1.B Page registration in AURIS-MM
Researcher
1.2.1 Running OntoMat
1.2.4 Generating annotation
1.2.2 Loading appropriate ontology
1.2.3 Providing information

Fig. The registration of multimedia element.



Multimedia element registration

Title (Ger)

Title (Eng)

Author

Creation date: DD/MM/YYYY

Description

Length in MB

Length in minutes

URL

Rights

Format: VIDEO | STREAMING REALVIDEO

Subject (about): PROJECT | TUWIEN

Full-text description

Audience: ATTENDIES | STUDENTS | PUBLIC

Language: None

Upload

COLLECTING METADATA

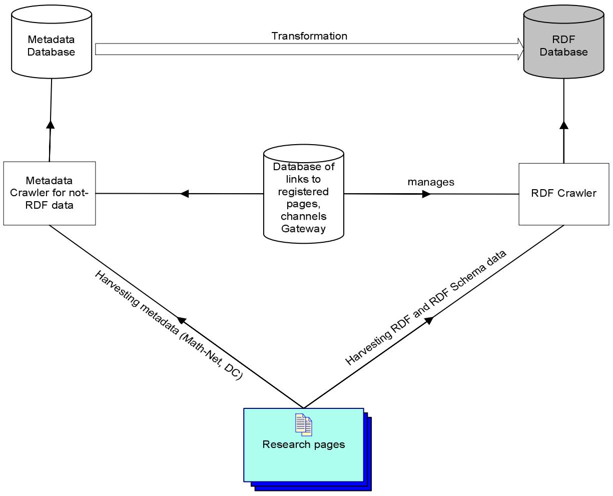
To make knowledge annotated on the web pages accessible for retrieval, it should be collected, analyzed, stored and made accessible for query engine.

Harvesting (collecting) RDF metadata possible by using RDF Crawler (http://ontobroker.semanticweb.org/rdfcrawl/index.html) – java application, which can crawl web pages and collect RDF data. After crawling RDF Crawler produces one file which store all RDF data and declaration of all used RDF Schemas.

The data about research now provided in different markup formats. Austrian research information system, Math-Net(http://www.math-net.org) and other societies use different markups to annotate date.

In our approach all data should be converted to RDF to be accessible for search and analysis through one search engine.

Fig. Metadata collecting into RDF database


Metadata Database
Transformation
RDF Database
Metadata Crawler for not-RDF data
Database of links to registered pages, channels Gateway
manages
RDF Crawler
Harvesting metadata (Math-Net, DC)
Harvesting RDF and RDF Schema data
Research pages

QUERYING COLLECTED METADATA, GETTING KNOWLEDGE FROM ANNOTATIONS

Once the annotated metadata were collected, how to use them?

There are several tools which can be used to search annotated pages.

SHOE Search Engine – Semantic Search (http://www.cs.umd.edu/projects/plus/SHOE/search/) search registered annotated pages. User of search engine can choose ontology, then choose type of resource he searches, create very simple filter conditions and search SHOE metadata database.

Our approach assumes that data would be described in RDF or can be translated into RDF by transformation procedure. Also to provide search services for researcher query facilities should be able to search data by its meaning (type of resource or property), values of attributes (properties) and relation between resources.

There are several query engines for RDF[Karv-2000], Squish, Ontobroker, Redland RDF Application Framework, MetaLog, RDF Data Query Language.

In our project to query RDF database Sesame RDF Query Repository and Querying Facility is used.

Sesame supports RQL (RDF Query Language) [Vass] which is being developed by ICS-FORTH Institute. Sesame supports storing both RDF and RDF Schema information. Querying Facilities of Sesame supports Schema information about subclasses and subproperties, searching by attributes values, resource relations.

Table. Examples of SESAME queries to retrieve research information

http://derpi.tuwien.ac.at/~andrei/cerif.rdfs#Person All persons in database (and any subtype of a person, -researchers and student)
http://derpi.tuwien.ac.at/~andrei/cerif.rdfs#Researcher All persons who are researchers (or any subtype of researchers)
^***http://derpi.tuwien.ac.at/~andrei/cerif.rdfs#Researcher*** All persons, who are researchers and not any subtype of researcher
select X,Y ***from #Project {X}. #project_persons{Y}, {Z} #expertise_skill {E}*** ***where X = Z and N = "Semantic Web"*** All projects in Semantic Web with description of persons participation in them If the organization or person, or Research Information System asserts new type of project – software project and in RDF Schema provides that it is a subtype of AURIS-MM, then it will also searched.
select X,Y ***from ^#Project {X}. #project_persons{Y}, {Z} #expertise_skill {E}*** ***where X = Z and N = "Semantic Web"*** Only projects in Semantic Web asserted as exactly CERIF projects and participants of those projects

Sesame provides application interface through HTTP protocol, so application can query and update network RDF databases.

CONCLUSIONS

Use of Semantic Web technologies might be very fruitful for development of Research Information Systems.

The annotation of knowledge make it more easy to researchers and research organization to assert information about their research for dissemination. No need to register it in a number of information systems. Software agents can collect information and understand its meaning

Not only research data but also new domain knowledge can be also asserted and shared for use.

Query engines for Semantic Web due to that inference abilities and schema exploration can make development of Research Information System more easy then conventional technologies like Relational Database management systems because exploration of domain knowledge is very crucial for CRIS systems .

ACKNOWLEDGMENTS

I thank Walter Niedermayer and all AURIS-MM project staff, Vienna University of Technology for support and helpful comments on previous versions of this article.

REFERENCES

ERGO European Research Gateways Online http://www.cordis.lu/ergo

BL-2001 Berners-Lee T., Hendler J., Lassila O., The Semantic Web, Scientific American, May 2001

Hend-2001 Hendler J., Agent and the Semantic Web, IEEE Intelligent Systems Journal, March/April 2001

Erd-2001 M. Erdmann, A. Maedche, H-P. Schmurr, and S. Staab, From Manual to Semi-automatic Semantic Annotation, Linkцping Electronic Articles in Computer and Information Science, Vol 6 (2001)

SWA Semantic Web Activity http://www.w3.org/2001/sw

Lee-2001 Berners-Lee T., Hendler J., Scientific publishing on the 'semantic web', 12 April, The Nature, http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm

Hefl-99 Jeff Heflin, James Hendler, and Sean Luke, SHOE: A Knowledge Representation Language for Internet Applications, Technical Report CS-TR-4078 (UMIACS TR-99-71). 1999. http://www.cs.umd.edu/projects/plus/SHOE/pubs/#tr99

SHOE SHOE home page. http://www.cs.umd.edu/projects/plus/SHOE/

OIL Ontology Inference Layer web site http://www.ontoknowledge.org/oil

Fens-2000 D. Fensel et al.: OIL in a nutshell In: Knowledge Acquisition, Modeling, and Management, Proceedings of the European Knowledge Acquisition Conference (EKAW-2000), R. Dieng et al. (eds.), Lecture Notes in Artificial Intelligence, LNAI, Springer-Verlag, October 2000.

DAML DARPA Agent Markup Language. http://www.daml.org

Jeff-98 K. G. Jeffery, “ERGO: European Research Gateways Online and CERIF: Computerized Exchange of Research Information Format”, ERCIM News N. 35, 1998, http://www.ercim.org/publication/Ercim_News/enw35/jeffery.html

CERIF-2000 Common European Research Information Format 2000 Guidelines. ftp://ftp.cordis.lu/pub/cerif/docs/cerif2000.htm

Lind-2000 Niclas Lindgren, Anita Rautamдk, Managing Strategic Aspects of Research, CRIS-2000, (ftp://ftp.cordis.lu/pub/cris2000/docs/rautamdki_fulltext.pdf)

Aks-2000 Dag W Aksnes, Johanne-Berit Revheim, The Application of CRIS for Analyzing Research Output - Problems and Prospects, CRIS-2000 (ftp://ftp.cordis.lu/pub/cris2000/docs/aksnes_fulltext.pdf)

Grot-98. M. Grotschel, L. Lugger, "Scientific Information systems and Metadata", Classification in the Information Age. Proc. of the 22nd Annual GfKl Conference, Dresden, March 4-6, 1998.

Lop-01. Lopatenko A. S., Kulagin M. V. "Current Research Information Systems and Digital Libraries. Needs for integration", to appears in proceedings of "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Sep. 2001

MathNet. Math-Net Application Profile http://www.iwi-iuk.org/material/RDF/1.1/profile/MNPage/

CERIF CERIF Homepage. http://www.cordis.lu/cerif

Hunt-2001 J. Hunter, "Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology", SWWS, Stanford, July 2001

Noy-2001 Noy N. F., Ontology Engineering, Semantic Web Working Symposium, 2001, Stanford

Noy-G Noy N. F., McGuinees D. L., Ontology Development 101: A Guide to Creating Your First Ontology, http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

Staab-2001 Staab, S., Maedche, A., and Handschuh, S.: Creating Metadata for the Semantic Web: An Annotation Framework and the Human Factor. Technical Report, 2001 http://www.aifb.uni-karlsruhe.de/WBS/sha/papers/semantic-annotation.pdf

OntoMat Webpage annotation tool. http://ontobroker.semanticweb.org/annotation/ontomat/index.html

Karv-2000 Karvounarakis G., Querying RDF Metadata and Schemas Technical Report, Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH), Crete, Greece, http://www.ics.forth.gr/proj/isst/RDF/rdfquerying.pdf

Vass G. K. Vassilis, C. D. Plexousakis, S. Alexaki, “Querying Community Web Portals”, http://139.91.183.30:9090/RDF/publications/sigmod2000.html