LT World: Ontology and Reference Information Portal

Brigitte Jörg*, Hans Uszkoreit°, Alastair Burt°
German Research Center for Artificial Intelligence (DFKI), Language Technology Lab
*Alt-Moabit 91c, 10559 Berlin, °Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
E-mail: {brigitte.joerg; hans.uszkoreit; alastair.burt}@dfki.de

Abstract

LT World (www.lt-world.org) is an ontology-driven web portal aimed at serving the global language technology community. Ontology-driven means, that the system is driven by an ontological schema to manage the research information and knowledge life-cycles: identify relevant concepts of information, structure and formalize them, assign relationships, functions and views, add states and rules, modify them. For modelling such a complex structure, we employ (i) concepts from the research domain, such as person, organisation, project, tool, data, patent, news, event (ii) concepts from the LT domain, such as technology and resource (iii) concepts from closely related domains, such as language, linguistics, and mathematics. Whereas the research entities represent the general context, that is, a research environment as such, the LT entities define the information and knowledge space of the field, enhanced by entities from closely related areas. By managing information holistically – that is, within a research context – its inherent semantics becomes much more transparent. This paper introduces LT World as a reference information portal through ontological eyes: its content, its system, its method for maintaining knowledge-rich items, its ontology as an asset.

1. Introduction

LT World started its activities in 2001 in the German Competence Center for Language Technology, funded by the German Federal Ministry of Education and Research (Capstick et. al. 2003). Since its second release in 2004, all LT World information and knowledge has been ontologically structured (Uszkoreit and Jörg 2003). After the successful completion of the initial platform, DFKI continued to develop the system through own resources. Since 2009, LT World has been receiving support from the BMBF-funded project TAKE[1], and, to a smaller degree, from CLARIN[2], a pan-European effort to create, coordinate and make language resources and technology available and usable for scholars in the Humanities and Social Sciences. As a result, the LT World portal has recently been set up with Zope[3], an open source application framework and Plone[4], its content management system, enabling user registration and thus, community participation. Zope and Plone offer rich sets of continuously improved features, and are supported by a lively developer community. The technology behind the LT World portal (Burt and Jörg 2008) is being prepared to move towards its next generation with Plone3, where the user interfaces underwent major changes. In parallel, the LT World ontology has been extended towards a richer coverage of natural language resources and tools. Further development of the LT World portal and ontology are a central component within the new Network of Excellence META-NET, forging META[5] – the Technology Alliance for a Multilingual Europe. LT World is serving as the knowledge portal of this initiative.

The LT World approach to ontology modelling has been top-down and pragmatic. The information portal grew from a need to support human access to the field by structuring its knowledge, and to enable machine access by formalizing this knowledge, while, at the same time, facilitating maintenance and extension of the knowledge base.

2. Content

LT World records have been collected from the initial phase, and organised within four major classes: Information and Knowledge (I&K), Players and Teams (P&T), Research and Development Systems (R&D), Communication and IPR (C&I). Every instance aligns with the underlying ontology by metadata; attributes and relationships inherited from multiple classes or super classes. The number of LT World records as of March 2010 is the following:

- **I&K**: technologies (112); information sources (269)
- **P&T**: organisations (2563); projects (891); people (3087)
- **R&D**: systems (326); products (607); repositories (16)
- **C&I:** news (2487); events (1827); patents (1106)

The presented structure is also applied to the public portal, where the maintenance of records requires an organisation with workflows, and a distinction is necessary between the record types, for their maintenance. The I&K-type records are field specific and require a deep knowledge of the area. The P&T, R&D and C&I-type records are less discipline-specific and require two kinds of maintenance actions. Essential attributes like name, title, or acronym, and relationships like person-participate-project or person-affiliate-organisation are intuitively clear. However, much richer and specialised knowledge is required to relate research domain records to I&K-type records as for project-apply-resource, or project-develop-technology; locating a record within the knowledge space.

[1] TAKE: http://www.dfki.de/lt/project.php?id=Project_539&l=en
[2] CLARIN: http://www.clarin.eu/
[3] Zope: http://www.zope.org/
[4] Plone: http://www.plone.org/
[5] META: http://www.meta-net.eu/

As LT World knowledge is ontologically structured, all the records are highly interlinked: contextually, field-specific, and field related. Most records are related to I&K-type records, as indicated with the arrows in figure 1, where the relationships from the "Language Technology" concept are propagated to sub concepts and their contained records.

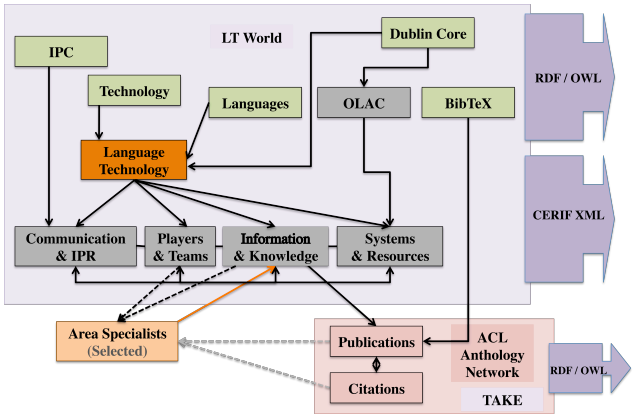

Figure 1: LT World ontology, portal structure and maintenance concept

3. Knowledge Organisation

At physical level – within the database – each record is identified via ID, and represented by its essential attributes (i.e. for organisation: homepage, name, variants), and relationships (i.e. developed) as indicated in figure 2.

```
id="obj_60215"
homepage="http://www.dfki.de/"
name="German Research Center for Artificial
  Intelligence"
nameVariants="DFKI GmbH; Deutsches
  Forschungszentrum für Künstliche Intelligenz"
abbreviation="DFKI"
contact="info@dfki.de"
location="67663 Kaiserslautern"
country="Germany"
developed="obj_80987", "obj_77625", "obj_77931"
investigated=" obj_88216", "obj_61059"
```

Figure 2: Example record for organisation DFKI represented by attributes and relationships

At the public portal, record IDs resolve to URLs, revealing their ontological commitment. For the DFKI record in figure 2: http://www.lt-world.org/kb/players_and_teams/organisations/academic_institutions/obj_60215, it is obvious that an organisation is a player in a team, and, DFKI is an academic institution. Physically, LT World relationships are established by IDs within the database, i.e. obj_60215 referring to a product obj_80887 through a relationship "developed" or to a person obj_88216 through a relationship "investigated". For portal users, the referred objects are presented as hyperlinks labeled with the referred object's name.

4. Ontology

The LT World ontology has been implicitly presented while introducing the content, and the record type structure of the LT World portal. Knowledge-rich LT World concepts belong to the technology and resource parts of the ontology. The LT World ontology has recently been extended to allow for a richer coverage of LT Resources. The need grew from the TAKE, project, to build a quality, human-annotated seed set of papers for machine learning. For TAKE, papers have been annotated with (a) technology types, with (b) resource types, and with (c) supported language. For (a) and (b) at least one, and at a maximum three annotation relationships were allowed, as presented in figure 4.

For the technological relationships (a), the LT World technology records were readily linkable. For the resource relationships (b), the LT World ontology has been extended as indicated in figure 3:

- LT Resources
  - Data Resources
    - By Annotation
    - By Form
    - By Language
  - Knowledge Resources
    - Linguistic
    - Non-Linguistic
  - Systems
    - Status Type
    - Application Type
    - Basic Type
  - Theories and Models
    - Algorithmic
    - Linguistic
    - Other

Figure 3: LT World Resource ontology (two-level) view

The ontological resource extensions within the LT World ontology will be added with the forthcoming portal release, consolidating existing standards and catalogues.

5. Knowledge-rich Records

It has become clear, that the wealth of collected information in LT World is of heterogeneous nature, and the maintenance of the knowledge-rich records requires highly specialized experts from the community. Knowledge changes constantly and especially in a research environment – frequently. Such changes need to be reflected and incorporated into the records regularly, to keep the knowledge of the field up-to-date, and, as a reference, available for the community, and the next generation. A Wikipedia-like approach may seem a temptation, however the complexity of the field requires more authority.

We consider the LT World technology records as knowledge-rich: http://www.lt-world.org/kb/ information_and_knowlege/technologies/ (such as "Grammars", "Machine Translation", "Information Extraction").

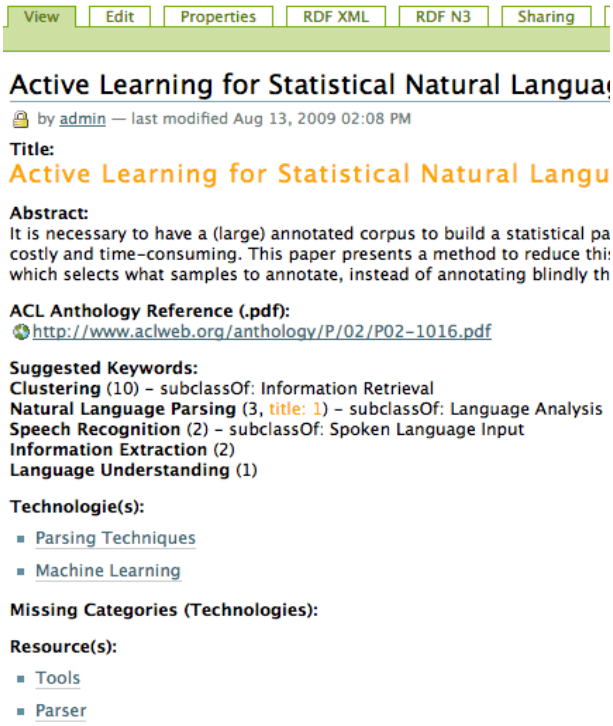

Figure 4: TAKE publication extract, with technology and resource relationships, and LT World technology term matches in "Suggested Keywords"

To maintain the knowledge-rich records, we started to systematically involve carefully selected area specialists from the global LT community. In order to select these area editors on objective criteria from respective groups of currently active recognized intellectual leaders within each area we designed a pre-selection process that gives us for an area the scientists most active and most cited in the particular field. As there is no place where such information can be found, we calculate it from the most highly recognized body of computational linguistics literature in a few steps:

(1) Collection of ACL Anthology papers from ACL + COLING 2002 - 2008
(2) Transformation of pdf to text
(3) Extraction of technology terms
(4) Identification of authors for fields
(5) Incoming citations for authors
(6) Author activity (ranking) in the ACL Anthology Network

The selection process for LT World area specialists started from the collection of ACL Anthology papers, namely the ACL and COLING conference series in the years from 2002 to 2008, which cover an international range of experts. In the second step, pdf papers were transformed into text files with Apache PDFBox, as presented in (Schaefer et al. 2007). From these text files, the extraction of LT World technology terms (i.e. "Information Extraction", "Machine Translation", "Grammar") started.

Files preserving the highest number of occurrences of terms (including synonyms and variants) have been ranked highest. For these highest ranked files, the author names have been identified (ACL BibTeX files according to paper IDs), inversely revealing authors' activities in technological fields. Additionally to the ranking of authors based on term matches in the text files, the number of incoming citations based on the authors' last name matches, has been taken into account. For this step, the text files were computed with ParsCit (Councill, et. al. 2008), resulting in normalizations of authors' names from references in the paper collection. Finally, the authors' ranking in the ACL Anthology Network was taken into account, to validate the determination process.

The selection method of area specialists as presented in its first version, worked out very well for some LT World fields, whereas for other fields the results were not good enough due to disturbances from substring matches, or due to a limited number of terms or synonyms available in the ontology. An area specialist will finally be responsible for field updates during a certain period of time. A "field" is represented by an LT World technology record, and indicated with figure 5.

```
url: http://www.lt-world.org/kb/
information_and_knowledge/technologies/…
name: technology name
abbreviation: abbreviated technology name
name variants: name variant;
description: Technology (XY) is […]
related projects: record ID1, record ID2
related organisations: record ID3
related persons: record ID4, record ID5
related systems: record ID6, record ID7
related resources: record ID8, record ID9
related publications: record ID10, record ID11
```

Figure 5: LT World technology dummy record from a database perspective

The area expert will define the range of his field or subfield by giving it a name, an abbreviation, name variants or synonyms. S/he will maintain a textual definition or description of the field and provide the relevant relationships with projects, organisations, people, systems, resources, and publications. From the records, users should be able to get in contact with the technology and understand its essence. Current relationships with projects, organisations or people, systems or products as well as resources and publications should represent the state of the art in the area. Each link is a physical reference to a record in the LT World database, itself represented by its own attributes and relationships. Figure 6 shows one technology record, namely the Machine Translation record from a logged-in user's view.

Links referring to not yet existing records require the creation of new records for linkage, according to the ontology. Each new record will thus become part of the ontology and extend the LT World knowledge base. The maintenance of LT World records is fully web-based by user authentication, and is guided through its underlying, ontology-driven system.

Figure 6: LT World Machine Translation technology record from an editor's perspective

6. Exchange and Integration

The entire LT World portal is driven by its underlying ontology; each record is organised according to the schema. An integration of the ontology with the entire portal ensures that the portal information always commits to the ontology, and in the same way, the ontology is always up-to-date with the portal content. The ontology is available in RDF and OWL format upon request, and will become available with META for registered users. The ontological schema has been modeled with Protégé[6], a free modeling tool.

In addition to the ontologically integrated, and thus, highly interrelated and complex representation of information and knowledge, a rather modularized representation of records and references seems useful, in particular, with respect to information exchange. Therefore, we started to provide record-type information with CERIF[7] XML, a format, recommended by the European Commission to Member States to manage current research information. A first snapshot of CERIF representations has been prepared for CLARIN with organisation and system records, including some relationships: http://www.lt-world.org/clarin/

For an integration of record lists from other systems, a semi-automated merging tool has been set up. In this way, LT World has incorporated system and product information from the ACL NLSR Registry[8] hosted at DFKI. The merging environment works for every record type. It enables furthermore, a semi-manual cleaning of duplicate records.

7. System Architecture

The architecture behind LT World is unique in that it is able to derive the schema for the underlying content management system[9] from an OWL[10] ontology. The schema generation software (Burt and Jörg 2008) provides a basic web site for viewing, searching and editing data, importing and exporting the data, with an OWL ontology as its only input.

The advantages of a tight integration of the ontology and the web site are fivefold:

1. The data published on the web site have a clear, formal semantics.

2. The information architecture behind the site can be specified without programming knowledge through ontology editors such as Protégé.

3. The ontology defines a natural means to import and export data as serializations of a well-defined RDF[11] graph, in any of the formats (XML, Turtle, Manchester Syntax) commonly used for this task. Moreover, the underlying system provides a natural REST-based protocol for such import and export.

4. Through its reliance on URIs and RDF, the system allows information from LT World to be immediately integrated with other information on the net in accordance with the principles of Linked Data[12].

5. The data can be processed by the growing list of tools for RDF-based knowledge management. In particular, in addition to the native querying facilities of the host content management system, the data can be queried through SPARQL[13] or related query languages (Frank et al. 2005).

8. Acknowledgements

LT World is supported by the BMBF-funded project TAKE with contract no. 01IW08003, to a smaller degree by the EU-funded project CLARIN via grant agreement no. INFRA-2007-2.2.1.2. Current development is being co-funded by the European Commission through META-NET, grant agreement no. 249119.

6 Protégé: http://protégé.stanford.edu/

7 CERIF: http://www.eurocris.org/cerif/cerif-releases/

8 ACL Natural Language Software Registry: http://registry.dfki.de/

9 LT World is driven by the Plone content management system and the Zope web application sever.

10 OWL: Web Ontology Language

11 RDF: Resource Description Framework

12 Linked Data: Tim Berners-Lee, Linked Data - Design Issues, 2006, http://www.w3.org/DesignIssues/LinkedData.html

13 SPARQL: W3C Recommendation 2008

9. Outlook

The extension of services and further system developments will be a central component in the Network of Excellence META-NET, forging META – Technology Alliance for Multilingual Europe. In this role, LT World will serve as the knowledge portal of META, an interface between initiatives and communities of researchers, technology providers, and other stakeholders, also strengthening the collaboration with and among other initiatives such as ELRA, FlaReNet, CLARIN, LDC, and the Language Grid.

10. References

Burt, A., Jörg, B. 2008. Automatic Generation of a Content Management System from an OWL ontology and RDF import and export. *Linked Open Data Triplification Challenge, I-Semantics 2008*, Graz, Austria.

Capstick, J.; Declerck, T.; Erbach, G.; Jameson, A.; Jörg, B.; Karger, R., Uszkoreit, H.; Wahlster, W.; Wegst, T. (2002) COLLATE: Competence Center in Speech and Language Technology. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Canary Islands, Spain, May 2002.

Bechhofer, S.; van Harmelen, F.; Hendler J.; Horrocks, I.; McGuinness, D. L.; Patel-Schneider, P.-F.; Stein, L. A. (2004), OWL Web Ontology Language. W3C Recommendation 2004.

Bird, S.; Dale, R.; Dorr, B. J.; Gibson, B.; Joseph, M. T.; Kan M-Y.; Lee, D.; Powley, B.; Radev, D.R. Tan, Y.F.: The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. *InProceedings: LREC 08,* Marrakesh, Morocco, May 2008.

Councill, Isaac G.; Giles, C. Lee; Kan, Min-Yen (2008): ParsCit: an open-source CRF reference string parsing package. *In Proceedings of LREC'08*, Marrakesh, Morrocco, May 2008.

Principled Design of the Modern Web Architecture. *In Proceedings of ICSE2000*, Limerick, Ireland, June 2000, pp. 407-416.

Frank, A.; Krieger, H.-U.; Xu, F.; Uszkoreit, H.; Crysmann, B.; Jörg, B.; Schäfer, U. (2005). Question Answering from Structured Knowledge Sources. In: *Journal of Applied Logic, Special Issue on Questions and Answers: Theoretical and Applied Perspectives*, Amsterdam, 42 pages.

Hayashi, Y.; Declerck, T. ; Buitelaar, P. ; Monachini, M. (2008): Ontologies for a Global Language Infrastructure. In : Jonathan Webster, Nancy Ide, Alex Chengyu Fang (eds.) : In *Proceedings of the 1st International Conference on Global Interoperability for Language Resources*, January 9-11, Hong Kong, China, Pages 105-112.

Lassila, O. and Swick R. R. (1999): Resource Description Framework (RDF) Model and Syntax Specification, W3C Proposed Recommindation.

Schäfer, U., Uszkoreit, H., Federmann, C., Marek, T., Zhang, Y. (2007): Extracting and Querying Relations in Scientific Papers. In Proceedings of the *31st Annual German Conference on Artificial Intelligence, KI 2008*, Springer LNCS 5243, pages 127-134, September 2008, Kaiserslautern, Germany.

Prud'hommeaux, E. and Seaborne, A. (2008): *SPARQL Query Language for RDF*. W3C Recommendation 2008

Uszkoreit H. and Jörg, B. (2003): A Virtual Information Center for Language Technology: Ontology, Datastructure, Realization. In: *Nordisk Sprogteknologi 2002*. Museum Tusculanums Forlag, Kobenhavns Universiteit, 2003.