LREC 2006 Pre Conference Workshop

Towards a Research Infrastructure for Language Resources

Workshop: 22. May 2006
Magazzini del Cotone Conference Center, Genoa, Italy
Main Conference: 24-26. May 2006

Second Call for Papers

http://www.mpi.nl/lrec/2006
Changed deadline for Abstract Submissions: February 17th 2006 !!!
Submission per email to: lrec.workshop2006@mpi.nl

Background

Many teams are working hard on establishing a sound framework for eHumanities where language resources play a fundamental and enabling role both with language as object of research and language as carrier of meaning. The future researcher wants to interact with an integrated and interoperable domain of language resources that is persistent, accessible and extendable. Here, language resources are meant in the more general sense, i.e. they cover data resources (texts of different sorts, annotated multimedia recordings, lexica, grammars, geographical databases etc), tools (aligners, annotators, parsers, taggers, meaning extractors etc) and knowledge sources (metadata, data category registries, relation registries and ontologies). Only a solid and sustainable research infrastructure breaking national boundaries will help us to realize the researcher's dream. Persistence will be of crucial importance, since researchers will only invest time if they see potential benefits.

Many projects have been carried out at national, European and international levels that have helped us to test frameworks, to build up basic technologies, to improve standardization, to create language resource archives and to test new forms of interaction and collaboration. To just mention a few initiatives from the domain of language resources that were carried out (not meant to be comprehensive):

- for standardization work: TEI, EAGLES, ISLE, MILE, ISO TC37/SC4
- for metadata frameworks: DC, IMDI, OLAC, MPEG7, METS
- for schemas: LMF, TIPSTER, EAF, MAF
- for knowledge representation: ISO DCR, GOLD
- for registration, integration and services: INTERA, ECHO, DAM-LR, LIRICS

All are built upon strong international backbone network infrastructures, emerging Grid middleware and common standards and frameworks such as XML, RDF and web services. In addition we can refer to national formation processes that will form the pillars for a sustainable international research infrastructure. In Europe for example we can refer to AHDS (UK), DANS (NL) and CNRS-eScience (FR) as examples for national centers for the humanities.

Based on the experience we have over the years in the language resource community and the national backbones we can conclude that the Language Research Community is ready to establish solid research infrastructures. This is the reason why the EARL initiative for a European distributed Archive for Language Resources was placed recently on the roadmap for European research infrastructures (ESFRI).

Language Resource Centers

Language resource centers are the key pillars for such research infrastructures. They can be digital archives that, by their nature, should be based on principles and technologies that enable digital sustainability, such as: (1) Web-accessible metadata standards for resource management and cataloguing (2) Separation of the mutable physical structure from the logical one relevant for researchers; (3) Preservation of bit-stream representations by regular migration to new technology and by distributing them; (4) Facilities to allow interested and qualified researchers to add new data or upload new versions of existing data; (5) Easy and flexible user access to the resources; and (6) Utilization frameworks that take into account the heterogeneity of the resources in terms of linguistic

data types, structural differences and differences in linguistic terminology. But there can be other centers that maintain registries of useful components, schemas and tools.

All centers share the same basic characteristics: (1) they have to be embedded in national research strategies for the humanities; (2) they have to commit themselves to offer stable services and (3) they must be willing and able to act as partners in international scenarios. The latter includes the need to define the organizational, legal and ethical basics of federations. Already now we have a network of relevant international collaborations such as DELAMAN[1], OntoLex[2] and ISO TC37/SC4[3].

Goals

As well as addressing questions as to what the organizational pillars of research infrastructures and the exact identity of federations of language resource centers and archives might be, the workshop will discuss and share information about technologies that can help in setting up and managing large research infrastructures for language resources. All technologies that are important and currently being tested out in European or international projects should be critically discussed to understand their potential and state of maturity. Some time will be devoted to discussing roadmap issues.

Papers

We would like to invite the submission of papers that have the potential to contribute to building a stable and sustainable research infrastructure for language resources, with digital archives as one of their key pillars. A number of keywords may indicate the scope of the workshop:

- organizational, legal and ethical aspects of federations and RI
- aspects of sustainability
- standards improving interoperability
- Grid middleware technologies
- frameworks facilitating integration
- registries for all type of language resources
- archiving strategies for digital resource

We would like to motivate contributors to give summarizing papers that compare existing technologies and organizational frameworks based on project experience and that evaluate them with respect to their suitability for stable and sustainable research infrastructures and archive federations in the language resource domain as part of future eHumanities.

We expect extended abstracts of about 1000 words to be submitted as WORD, PDF or ASCII documents per email to the following address: “lrec.workshop2006@mpi.nl”. The final papers should not have more than 4 pages.

Dates

Deadline for Abstracts:	February 17th 2006
Notification of acceptance:	February 24th 2006
Final Papers:	March 31st 2006
Workshop CDROM, Website with contributions and program, etc:	May 14th 2006
Workshop:	May 22nd 2006

Location

Magazzini del Cotone Conference Center, Genoa, Italy

Organizers

Peter Wittenburg	Max-Planck-Institute for Psycholinguistics, Nijmegen, Netherlands
Remco van Veenendaal	Dutch Institute for Lexicology (INL), Leiden, Netherlands
Heidi Johnson	AILLA, Texas University, Austin, USA
Linda Barwick	PARADISEC, University of Sydney, Australia

Questions

Questions about the workshop can be addressed to: lrec.workshop2006@mpi.nl

[1] DELAMAN = Digital Endangered Languages and Music Archives Network; www.delaman.org
[2] OntoLex = www.ilc.cnr.it/ontolex2005/
[3] ISO TC37/SC4 Management of Language Resources; www.tc37sc4.org

Program Committee

Victoria Arranz	ELDA, Paris
Linda Barwick	Paradisec, U Sydney
Jeannine Beeken	TST Center – INL, Leiden
Hans Bennis	Meertens Institute, Amsterdam
Steven Bird	U Melbourne and U Pennsylvania
Daan Broeder	MPI for Psycholinguistics, Nijmegen
Lou Burnard	Oxford University Computing Services
Nicoletta Calzolari	ILC, Pisa
Khalid Choukri	ELDA, Paris
Helen Dry	E-Meld, LinguistList, Michigan
Maria Gavrilidou	ISLP, Athens
Gary Holton	U Alaska, Fairbanks
Michel Jacobson	LACITO, Paris
Heidi Johnson	AILLA, Austin
Peter van der Kamp	Institute for Dutch Lexicology, Leiden
Boyd Michailovsky	LACITO, Paris
Richard Moyle	AMPM, Auckland
David Nash	AIATSIS, Canberra
David Nathan	ELAR Archive, SOAS, U London
Nelleke Oostdijk	CLS, Nijmegen
Stelios Piperidis	ILSP, Athens
Laurent Romary	LORIA, Nancy
Florian Schiel	BAS, Munich
Gary Simons	SIL International, Dallas
Sven Strömqvist	Linguistic Department, U Lund
Nicholas Thieberger	PARADISEC, Melbourne
Remco van Veenendaal	TST Center – INL, Leiden
Peter Wittenburg	MPI for Psycholinguistics, Nijmegen
Martin Wynne	Oxford Text Archive, UK