The Workshop/Tutorial Programme

9.00	P. Wittenburg (MPI for Psycholinguistics, Nijmegen)	Welcome and Introduction to Research Infrastructures
9.10	T. Varadi (Hungarian Academy of Sciences, Budapest)	Putting Language Resources Infrastructure to the Test: the ESFRI Challenge
9.35	M. Theofilatou (EC, DG Research Brussels)	Research Infrastructures and FP7
10.10	S. Furui (Tokyo Institute of Technology, Japan)	Research Infrastructures for Systematization and Application of Large-scale Knowledge Resources
10.35	L. Barwick (Sydney University and DELAMAN)	Research Infrastructures – the Australian Perspective
11.10	Coffie break	
11.30	M. Wynne, Sheila Anderson (AHDS, London)	The Arts and Humanities Data Service: research infrastructure in the UK
11.50	N. Calzolari (ILC-CNR, Pisa)	Community Culture in Linguistics – an international perspective
12.10	B. Maegaard (CST, Uiversity of Copenhagen)	Organization Models for RI and existing Infrastructures
12.30	G. Francopoulo et al (ISO TC37/SC4)	The relevance of Standards for RI
13.30	Lunch Break	
14.30	D. Nathan, et al (DAM-LR Project, SOAS London)	Foundation of a Federation of Archives
14.50	D. Broeder et al (DAM-LR Project, MPI Nijmegen)	Integrated Services for the Language Resource Domain
15.10	A. Yli-Jyrä (Helsinki University)	Common Infrastructure for Finite-State Methods and Linguistics Descriptions
15.30	A. Itai (Israel Institute of Technology, Haifa)	Knowledge Center for Processing Hebrew
15.50	O. Streiter et al (?)	Design Features for the Collection and Distribution of basic NLP Resources for the World's Writing Systems
16.30	Coffie Break	
17.00	S. Krauwer, T. Varadi, P. Wittenburg	CLARIN – the next steps
17.30	Panel Discussion	Next Steps Towards a Research Infrastructure for Language Resources
18.30	Closed CLARIN Meeting (only for delegates of member institutions)	
19.30	End	

Workshop/Tutorial Organiser(s)

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

Workshop/Tutorial Programme 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

Table of Contents

Author Index

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

http://www.mpi.nl/lrec/2006

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, the term "language resources" is used 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 that transcends national boundaries will help us to realize the researcher's dream. Sustainability is of crucial importance, since researchers will only invest time if they see potential benefits that last.

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 of those initiatives from the domain of language resources (not meant to be exhaustive):

- 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, TELRI, ECHO, DAM-LR, LIRICS

These are all built on 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), CNRS-eScience (FR) and Max-Planck-Digital-Library (D) as examples for national centers for the humanities.

ESFRI Process

In Europe the issue of pan-European infrastructures to support future eScience scenarios received increasing attention during the last year. This is mainly inspired by the goals of the European Strategy Forum on Research Infrastructures (ESFRI) to establish a priority roadmap for infrastructures. Many disciplines are currently in the process of designing and organizing for research infrastructures that are seen as mature enough to be funded. Based on the experience we

have gained over many years with the language resource community and based on the current existing national infrastructure situation we concluded that the Language Resource and Technology Community is ready to establish such a solid research infrastructure. This is the reason why the CLARIN initiative (Common Language Resources and Technology Infrastructure, http://www.mpi.nl/clarin) was formed, covering institutions from almost all countries in Europe, CLARIN intends to apply for funds in the 7th Framework Program of the EC. It is obvious that Language Resources and Technology have to offer services to the humanities disciplines as well and perhaps even beyond. This is the reason that CLARIN has to synchronize with other initiatives with a broader scope such as EROHS (http://www.portedeurope.org/IMG/pdf/Projet_EROHS-ESFRI.pdf) and DARIAH. Also the European Science Foundation started an initiative focusing on establishing research infrastructures called HERA (http://www.esf.org/esf_genericpage.php).

Language Resource Centers

Language resource centers are the key building blocks for such research infrastructures. They can be digital archives that, by their nature, should be based on principles and technologies that enable accessibility and 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 that can play a role here should also share some basic organizational 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. Recently, the partners of the DAM-LR (Distributed Access Management for Language Resources, http://www.mpi.nl/dam-lr) project which is building a federation of archives based on typical Grid components took the initiative to create the Live Archives document (http://www.mpi.nl/dam-lr). It summarizes the principles that should guide the work of Language Resource Archives and received already broad support.

International Networking

The Language Resource and Technology community can also refer to several networks of relevant international collaborations such as TEI, ACL, COCOSDA, DELAMAN, OntoLex, ISO TC37/SC4 and many others guaranteeing that the development of standards and technology is broadly discussed.

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 also be devoted to discussing roadmap issues.

Programme

The workshop offers an interesting programme with a mix of invited and submitted papers. There are contributions from European and international colleagues concentrating on political/organizational and there are more technological oriented papers. The programme will end with an open discussion about the next steps for the CLARIN initiative where also all sorts of related aspects can be discussed. S. Krauwer, T. Varadi and P. Wittenburg who mainly pushed the CLARIN work in collaboration with M. Everaert will be open for all kinds of comments and questions.

After the workshop there will be a closed meeting of all registered CLARIN members. Those who are not yet registered could either talk with one of the three CLARIN coordinators or one of the already registered members about the terms of becoming a member.

Putting Language Resources Infrastructure to the Test: the ESFRI Challenge

Tamás Váradi

Linguistics Institute, Hungarian Academy of Sciences
varadi@nytud.hu

The Language Resources and Language Technology community is one of the most dynamically growing vibrant communities of recent years. This is well attested by the history of LREC itself. It did not take ten years for it to become a massive event drawing several hundred contributors. Language resources are clearly seen as a cornerstone of research activities that provide impetus to a number of related fields ranging from hard core ICT projects to general-interest language preservation and querying. Still, it is fair to say that the main driving force behind language resources has been the language technology industry ever craving for more and more data.

The European Strategy Forum on Research Infrastructures (ESFRI) was launched in 2002 with the aim of working out a common platform on research infrastructures in Europe, and "to act as an incubator for international negotiations about concrete initiatives". The current ESFRI activities are focussed on creating a Roadmap of new research infrastructures of pan-European interest, which is due to be published by the autumn of 2006.

ESFRI members are one of two persons delegated by each member country who are typically high ranking officials in charge of cultural/scientific policy. Preparatory work has been going on in three Steering Groups devoted to Physical Sciences and Engineering, Biological and Medical Sciences, and Social Sciences and Humanities respectively. Their work was helped by Expert Groups consisting of 8 – 10 members. The Social Sciences and Humanities Working Group has two Expert Groups, one to cover Social Sciences the other devoted to Cultural Heritage. ESFRI started out reviewing existing Roadmaps with a view to integrating them in the ESFRI Roadmap but the main source of information for identifying potential infrastructure initiatives for the Roadmap was a questionnaire circulated through ESFRI members, inviting applications for projects to be identified for inclusion in the Roadmap.

There were three major initiatives from the domain of language resources and language technology submitted independently of each other, which in the end were consolidated into a single proposal named CLARIN. The CLARIN proposal preserved much of the broad community forming objectives of the EARL initiative but focussed its aims on serving the ESFRI SSH community.

This is challenge number one. The ESFRI call is to propose research infrastructure for the social sciences and humanities. Linguistics is clearly within this domain but important as it is, language resources and language technology has a much broader relevance than serving the needs of linguists. In fact, we need to drive this point home with most people outside our field because otherwise the popular view that language resources/technology is about linguistics prevails. But as far as the current ESFRI proposal is concerned, it is not enough to get our aims and scope of relevance clearly established. We need to constantly remind ourselves that CLARIN is not an infrastructure for our own community. It is an infrastructure meant to serve the needs of the social science and humanities researchers. This is a completely new role in that the traditional user base of our community was keen and, most importantly, able to make use of what our field had to offer. In contrast, we can expect no such readiness with humanities scholars. They may not even be aware of the benefits of using language resources and the relevant language technology in their own research. Hence the challenge to build an infrastructure that provide services that are capable of not just making resources and technology available but readily usable as well by the target audience. It requires actively promoting the use of language resources and technology and also providing them in as much tailored to the perceived needs of the target audience as possible. I am sure the discussion in the present workshop will result in many useful ideas for the strategies to follow to achieve this objective.

The other challenge concerns financial viability. The ESFRI Roadmap is supposed to be a select choice of initiatives that have the seal of approval by ESFRI as meeting the criteria of the scientific soundness and viability, pan-European relevance and maturity. Being on the Roadmap means no guarantee for being funded by the EU. In fact, ESFRI depends on the willingness of member countries to buy into the projects identified on the Roadmap. The biggest challenge for CLARIN, then, is to round up enough national support to make the project economically

feasible with minimal EU contribution. This is, again, an issue that calls for detailed discussion in our workshop.

Support to Research Infrastructures (RI) in the 7th Framework Programme (FP) for Research and Technological Development (2007 – 2013)

Maria Theofilatou

Research Infrastructures unit, European Commission, DG Research

1. What are Research Infrastructures?

The term "Research infrastructures" refers to facilities, resources and services that are needed by the research community (ies) to carry out their research in all scientific and technological fields). Examples include major equipment or a set of instruments; knowledge based resources such as collections, scientific archives and/or structured information; enabling Information and Communication Technology-based infrastructures; and any other entity of a unique nature that is used for scientific research. Research infrastructures may be "single-sited", "distributed", or "virtual" (the services being provided electronically). Last but not least, "One size does not fit all" and it is recognised that the needs for research infrastructures may vary considerably from one scientific field to another, from physical sciences and engineering to environmental, biological, biomedical and social sciences and humanities

2. Evolution of the Community support to Research Infrastructures

Community support to research infrastructures has been developed with success over consecutive Framework Programmes. The budget has increased from around € 30 million in FP2 to € 735 million in the current FP6 (2002-2006), which includes € 220 million for the further development of the communication network development for all researchers in Europe, Geant and Grids. Under this total budget for RIs in FP6, 142 projects have been funded and support has been given to 259 RIs serving directly more than 20000 users. Through the so-called Integrating Activities, Community support in FP6 concentrates primarily on activities which are related to the coherent use and development of existing research infrastructures. These activities cover access of researchers to research facilities, the networking of research infrastructures, and joint research projects to improve their performance. Through design studies and limited contribution to construction studies, the Community actions have started to support the development of new (or the major upgrading of existing) research infrastructures of pan-European significance. The DAM-LR project is an example of an FP6 construction study, which is aiming at creating a single, virtual linguistics resource out of four linguistics research institutions from the Netherlands, Sweden and the United Kingdom.

3. The Research Infrastructures' action in the FP7 proposal

There can be no doubt that state–of-the-art research infrastructures are essential for Europe's researchers to stay at the forefront of research development, and thus, an important task for the EU's research policy and FP7. In the Commission's proposal for FP7, this action forms part of the Specific Programme on "Capacities" and its budget is expected to increase considerably in FP7.

- How to reduce fragmentation of efforts and structure better, on a European scale, the way existing research infrastructures operate in a given field.? How to enhance existing research infrastructure capacity in the European Research Area (ERA)? One of the main objectives of the Research Infrastructure programme will be to continue to optimise the use and development of the best research infrastructures existing in Europe through the following activities:

  - *Reinforcement of Integrating Activities* to structure better, on a European scale, the way research infrastructures operate, in a given field, and promote their coherent and cross-disciplinary use. They will be implemented through both a "bottom up approach" to calls for proposals, open to all fields of science and technology as well as a "targeted approach" to calls for proposals and in close cooperation with the other FP7 thematic priorities
  - *ICT based e-infrastructures*: to foster the development and evolution of high-capacity and high-performance communication and grid infrastructures.

- How to develop a European long-term approach to the creation of new research infrastructures? What mechanisms to put in place to ensure a coherent and strategy-led approach to policy making on research infrastructures of pan European interest? How to facilitate multilateral initiatives leading to the better development, construction and use of research infrastructures in Europe? These are some of the questions, which have been continuously in the focus of high level discussions and reflections in recent years.

  The other major objective of FP7 is, therefore, to support the construction of new research infrastructures. Community action will be primarily based on the work of ESFRI[1] and the development of the first Roadmap for such new infrastructures.

  The Community action will include the following activities

  - *Design studies:* to promote the creation of new research infrastructures by funding feasibility studies*;*
  - *Support to Construction of new infrastructures* (or major upgrades of existing ones)

  In FP7, the construction of new infrastructures (incl. major upgrades) will follow a two-stage approach with a first stage supporting the *preparatory phase* and a second stage supporting the *implementation phase.* The preparatory phase will involve the finalisation of the detailed construction plans, of the legal organisation, of the management and multi-annual planning. The European Commission will act as a "facilitator" during this preparatory phase where selected projects will be funded with a maximum Community financial contribution of up to 50% of the total eligible costs. Only projects which will have completed successfully the preparatory phase will proceed to the construction phase.

  FP7 financial support for the construction phase may vary from 0% to a substantial financial contribution where there is critical need for such support. The possible EC funding to the construction of new RI could take different forms: a direct grant from FP7 to the construction costs, a guarantee of EIB loans through the Risk Sharing Finance Facility and ad-hoc decisions based on Article 171 of the Treaty.

***More information on EU Research Infrastructures :*http://www.cordis.lu/infrastructures/**

[1] Following a Commission initiative, the European Strategy Forum on Research Infrastructures (ESFRI) was set up in 2002. ESFRI brings together representatives of research ministers, and representatives of the European Commission to help developing a European policy on Research infrastructures, based on a medium to long term vision of the scientific needs in Europe. (http://www.cordis.lu/esfri)

Research Infrastructures for Systematization and Application of Large-scale Knowledge Resources

Sadaoki Furui
Department of Computer Science
Tokyo Institute of Technology
furui@cs.titech.ac.jp

To function at optimal levels of efficiency, the 21st century, which is being described as the century of knowledge resources, will require the construction of sophisticated, accessible, large-scale knowledge resources in every domain of research, education and daily life. Knowledge in this context refers to the structured representation of observed content, and the application of the rules that underlie this representation to information interpretation, problem solving, and information creation. Another way of describing this concept is as the comprehensive, integrated form of information which has been verified as valid, for a specific topic, and which is therefore more significant than a mere collection of data or observations. Thus a knowledge resource is the large-scale accumulation of usable knowledge, combined with meta-knowledge, and it represents a much more sophisticated object than mere content. While various individual knowledge bases exist today, inconsistent development approaches, lack of communication among participating research organizations, and the high level of complexity inherent to the project mean that these knowledge bases are usually dfficult to manage, extend or utilize.

In order to resolve these growing problems, a five-year COE (Center of Excellence) sponsored program, the 'Framework for the Systematization and Application of Large-scale Knowledge Resources' was launched at the Tokyo Institute of Technology in 2003. Since then, those involved in the project have been conducting a wide range of interdisciplinary research, combining information and knowledge from the humanities with technology from the natural and information science fields, in order to establish a large scale framework for the systematization and application of large-scale knowledge resources in electronic mediums. Figure 1 illustrates the strategy of the COE program, with integrated hierarchies for building infrastructure for research and education, investigating systematization technology, building fundamental knowledge resources, and constructing knowledge resource applications. Large-scale systems for computation, information storage and retrieval have been installed as infrastructure to support research and education. For the systematization of large-scale knowledge resources, statistical theories, graph theory, logic, ontology, and metadata techniques, as well as approaches to refining various traditional methods are being investigated. Various fundamental knowledge resources such as spoken language-resources, written-language resources, audio-visual resources and Web-based resources are being constructed in accordance with these technologies. On top of these fundamental knowledge resources, application oriented resources, including educational (e-learning) resources, classic literature, documents on historical sites, broad-casing resources and Web-knowledge resources are also being constructed.

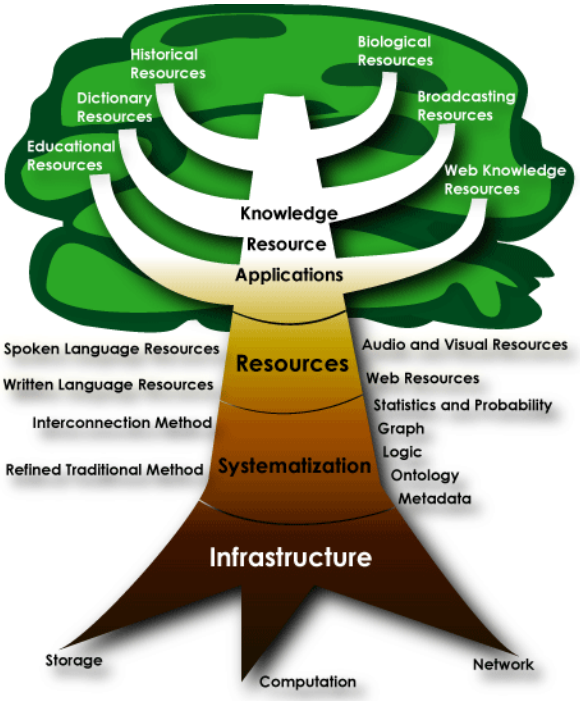


Fig. 1 - The strategy of the COE program.

Reference

[1] "Proceedings of the International Symposium on Large-scale Knowledge Resources (LKR 2006)", Tokyo, Japan, March 1-3, 2006.

Research Infrastructures for Language Resources – the Australian Perspective

Linda Barwick

University of Sydney

In this talk I will discuss the prospective place of research infrastructure for the humanities in general, and for language resources in particular, in the context of Australian demography and current Australian government initiatives for supporting research infrastructure through the National Collaborative Research Infrastructure Strategy (NCRIS) and the E-Research Initiative. Examples will be provided from PARADISEC (the Pacific and Regional Archive for Digital Sources in Endangered Cultures), an Australian collaborative research facility for providing services of digital preservation, curation and online access to Australian researchers' field recordings of endangered languages and musics of the Asia-Pacific region.

Since Australia is a large country with a relatively small geographically dispersed population, government investment in communication and transport infrastructure has been essential to constitution and governance of the national community. In 2004 the Australian Government announced the National Collaborative Research Infrastructure Strategy (NCRIS), allocating funding of A$542 million until 2010/11, with the expectation of further funding from the states and other funding agencies such as the Australian Research Council (ARC). NCRIS aims to support research areas in which Australia has the potential to do excellent world-class research, and which fit in with government-approved national research priorities (economic development, social and health related outcomes, environmental sustainability and national security).

Although Australian researchers in regional languages and cultures continue to undertake world-class research (hence meeting criterion 1), the humanities in general and language studies in particular are not favoured under the national research priorities as they are currently defined. The NCRIS strategy therefore at present only supports generic infrastructure requirements of humanities research (such as libraries and repository storage). Various humanities bodies in Australia, including the Australian Academy of the Humanities, are currently lobbying the government to improve recognition of the role of the humanities in Australia's international research standing, and hence improve funding for language resource infrastructure. One good feature of NCRIS is its recognition of the changing nature of national and international research, with its emerging emphasis on multi-disciplinary, networked, collaborative research teams.

Another Australian government initiative that is focussed in this direction is the current development of an e-Research infrastructure strategy. Although at an early stage, the e-Research coordination committee does recognise explicitly the needs and interests of the humanities and social sciences. This committee bases Australia's e-Research infrastructure strategy on existing and planned robust high-bandwidth advanced communications networks; distributed high-performance computing and data storage capacities; accessible data and information repositories and research instruments and facilities; and agreed standards and coordinated middleware development. In 2005 the ARC introduced a pilot funding scheme in e-Research, in which several of the funded projects have a language orientation (including the EthnoER project in which PARADISEC is currently participating).

The Arts and Humanities Data Service: the experience of building a research support infrastructure in the UK

Sheila Anderson and Martin Wynne

Arts and Humanities Data Service, UK

1. The AHDS in the UK research support infrastructure

The Arts and Humanities Data Service (AHDS) is a UK national service aiding the discovery, creation and preservation of digital resources in and for research, teaching and learning in the arts and humanities. Currently, the AHDS covers five subject areas:

- Archaeology,
- History,
- Literature, Languages & Linguistics
- Performing Arts,
- Visual Arts.

The AHDS is organised via an Executive at King's College London and five AHDS Centres, covering the five subject areas above, and hosted by various Higher Education Institutions. The AHDS is funded by the Joint Information Systems Committee (JISC) and the Arts and Humanities Research Council (AHRC).

The key functions of the AHDS are advice to funding bodies, advice to resource creators, and the archiving, distribution and preservation of resources.

1.1. Services to funders

One of the key functions of the AHDS is to provide an advisory service to funding bodies such as the AHRC, JISC and the British Academy. The AHDS gives advice to applicants on the technical aspects of their funding proposals, carries out assessments of the proposals for the funding body, offers advice to grant holders, offers advice on strategic and policy issues relating to electronic resources, and offers an archiving service for project outputs. These services are particularly well integrated into the processes of the funding schemes of the AHRC. Recipients of AHRC grants who are creating electronic resources are normally expected to deposit the resources with the AHDS within three months of the end of the project.

The innovative work and the organisational structure of the AHDS have already been taken as a model for the creation of an Economic and Social Data Service (ESDS), and have informed the work of many other new initiatives and structures in the UK.

1.2. Advice to resource creators

AHDS staff offer a free initial advisory service to UK academics who are creating electronic resources. In addition to dealing with directly with enquiries, various published materials cover data creation, access and delivery, and data deposit and preservation.

The AHDS has created a series of Guides to Good Practice focusing on the practical steps necessary to make a successful digital resource. Some of these are subject-based, while others cover cross-disciplinary topics. The latest title is *Developing Linguistic Corpora: a guide to good practice*. Case Studies review various projects which are creating or completing digital resources in the arts and humanities. Various issues are covered, such as funding, preservation, and using resources in teaching. Information Papers are shorter publications relating to specific technical questions. They are not focussed on any particular subject area but rather deal with the various aspects of a digitisation project (e.g. project management, metadata, XML editors).

The AHDS offers various workshops and events which give further advice to those interested in creating, maintaining and using digital archives.

2. The AHDS and language resources

The centre responsible for language resources is AHDS Literature, Languages & Linguistics, which is hosted by the Oxford Text Archive (OTA) in the University of Oxford. The OTA has been in operation for 30 years and has been the centre responsible for literary and linguistic subject areas since the foundation of the AHDS in 1996. AHDS Literature, Languages & Linguistics benefits from its location in Oxford University Computing Services in close proximity to projects and services including the Text Encoding Initiative, the British National Corpus, the Oxford e-Science Centre and the JISC Open Source Advisory Service (OSSWatch).

The language resource holdings of the AHDS include many historical and literary texts, language corpora, lexical data, and other types of literary and linguistic dataset. As with all AHDS resources, they are catalogued using the AHDS Common Metadata Format, and can be discovered via subject-specific portals and via the AHDS cross-search catalogue. AHDS Literature, Languages & Linguistics also shares discovery metadata with the Open Language Archives Community (OLAC).

AHDS Literature, Languages & Linguistics plays a central role in the UK in promoting good practice in the creation of language resources, in promoting the use of language resources in research and

learning and teaching, and archiving language resources and in developing new and improved ways to deliver them to the user.

3. AHDS and the European Infrastructure

Just as astronomers require a virtual observatory to study the stars and other distant objects in the galaxy, researchers in the humanities need a digital infrastructure to get access to and to study the sources that are until now hidden and often locked away in cultural heritage institutions. Only a fraction of the analogue sources in archives, libraries and museums is as yet available in a digital form. Of course, more and more sources are being digitized, but the permanent and open access to the information they contain is only yet beginning. It is therefore not surprising that the recent US Cyber Infrastructure for the Arts and Humanities report proposed as its grand vision 'access to all surviving humanities and cultural heritage information across all of time and space'.

Since the publication of that report, the National Science Foundation in the US has produced a draft strategy aimed at implementing many of the recommendations contained in the report. The challenge for Europe is to ensure that the development of a research infrastructure for cultural heritage and the humanities that can match, or supersede, that of the US. The infrastructure that is needed for the humanities is indeed very much comparable to infrastructures for the natural sciences such as the virtual observatory, and the kind of organisation and grid-based techniques that are required also show a surprising degree of similarity.

The AHDS is working with partners across Europe to identify and develop the key elements and activities for a Research Infrastructure (RI) that take steps towards achieving this grand vision for European humanities and cultural heritage information, and would provide an infrastructure that eventually could support access to all surviving humanities and cultural heritage information for Europe. Such a Research Infrastructure would:

1. Provide a coordinated infrastructure across Europe that would act as a catalyst to bring together the best efforts of national initiatives, organisations and individuals in order to provide upgraded and enhanced European wide actions, initiatives and services that could not be provided at local or national level.

2. Provide a coordinated infrastructure that would act as both a catalyst and support for the development of national services and digitisation programmes aimed particularly at those European countries without such services and programmes.

3. Provide a coordinated infrastructure that would act as a catalyst to bring together the different sectors involved in cultural heritage and humanities information management and access – education, memory and cultural heritage institutions and organisations, and the commercial sector – in order that they might work together for the benefit of both themselves and the research communities across Europe.

4. Provide a coordinated infrastructure that would act as a catalyst for the enhancement and promotion of digital scholarship in the humanities and arts across Europe, including facilitating cross-disciplinary research and the sharing of content, tools and methods across communities of practice and discipline domains. This would also ensure that the arts and humanities did not work in isolation but took note of developments across the social, physical and medical sciences.

Community Culture in Language Resources – An International Perspective

Nicoletta Calzolari

Istituto di Linguistica Computazionale – CNR
Via Giuseppe Moruzzi N° 1 – I-56124 Pisa – ITALY
glottolo@ilc.cnr.it

Abstract

I highlight a few issues which I consider of relevance with respect to the infrastructural role of Language Resources. I underline some of the circumstances and attitudes which are specific of the European approach, and sketch how I see the current situation in the LR field and what I think is of highest priority with respect to implementing an open Language Infrastructure. My objective is to show that it is imperative that there is an underlying global strategy behind the set of initiatives which are/can be launched in Europe and world-wide, and that a global vision and cooperation among different communities is necessary to achieve more coherent and useful results.

1. The growth of a Language Resources community culture

1.1. Setting the scene

Since the '80s it has become clear that Language Resources (LR) have progressively acquired a larger role in Human Language Technology (HLT), also in view of developing innovative and robust technologies or to integrate existing ones to achieve more advanced applications. This process achieved a crucial step through the acknowledgment of the infrastructural role of LRs, first recognized by A. Zampolli to whom we also owe the term itself 'Language Resources' [1]. This trend was very influential in the formation of the strategy of the European Commission (EC) in the '90s and in the launching of many European LR related projects and initiatives, the conditions and time being ripe for the speeding up of a major effort in LR development. LRs started to be considered as the necessary common platform on which to base new technologies and applications, a recognition which is nowadays widely accepted for the development and takeoff of our field.

Also the concept of reusability – directly related to the importance of "large scale" LRs within the dominant data-driven approach – has contributed significantly to the structure of many R&D efforts [2]. Many large international projects in this area, on both sides of the Atlantic and in Japan, were motivated by this idea. After the first pioneering EC projects on LRs already in the '80s - ESPRIT BRA ACQUILEX and EUROTRA-7 – there was a flourishing of international projects and activities (see also [3] for an overview) that contributed to substantially advance knowledge and capability of how to represent, create, acquire, access, tune, maintain, standardize, etc. large lexical and textual repositories.

1.1.1. Infrastructural initiatives

The set of these projects of the '90s can be seen as the beginning of a consistent and coherent realization in Europe of a well-thought plan to implement the badly needed infrastructure of LRs [4]. In addition to its "scientific" implications, this large intellectual and economic movement obviously entailed "strategic" considerations, and pushed towards the need to reflect on the situation in the area of LRs in Europe from a very broad perspective. Some of the LR projects, dealing with policy and meta-level issues related to LRs and standards, have been instrumental to define a coherent strategy for the LR field in Europe, and to give Europe a central position in the LR area, leading also to founding independent associations such as ELRA (European Language Resources Association), the European counterpart of the American LDC (Linguistic Data Consortium).

It was perceived as essential to define a general organization and plan for research, development and cooperation in the LR area, to avoid duplication of efforts and provide for a systematic distribution and sharing of knowledge. To ensure reusability, the creation of standards was the first priority. Another tenet was the recognition of the need of a global strategic vision, encompassing different types of (and different methodologies of building) LRs, for an articulated and coherent development of this field.

Even if LRs have a rather short history, they are nowadays recognised as one of the pillars of HLT, and a central and strategic component of the so-called "linguistic infrastructure" (the other key element being Evaluation), necessary for the development of any HLT system, application and product. The availability of adequate LRs for as many languages as possible is a pre-requisite for the development of a truly multilingual Information Society. They play a critical role, as a horizontal technology, in different areas of the EC 6th Framework Programme, and have been recognized as a priority within a few national projects around Europe.

1.1.2. Signs of the wide resonance of LRs

A few signs of the wide resonance LRs have acquired in the last decade can be found, among others, in a number of international initiatives: the LREC Conference (1000 participants in 2004 in Lisbon); bodies such as ELRA and LDC, or COCOSDA (International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques) and WRITE (Written Resources Infrastructure, Technology and Evaluation); the new international journal *Language Resources and Evaluation* [5]; not to mention the vital role of LRs in statistical and empirical methods, in evaluation campaigns, and so on. Moreover, there is a clear and growing industrial interest in the use of LRs and standards, in particular for multilingual applications.

On the one hand, such a solid position of the LR area must be maintained and reinforced, anticipating the needs

of new types of LRs and quickly consolidating (through EAGLES/ISLE-like standardisation initiatives) areas mature enough for recommendation of best practices and standards. A virtuous circle should be established between innovation and consolidation. On the other hand, however, much stronger initiatives are needed to achieve true interoperability (see e.g. the issue of open architectures below), for which I envision the need of a new paradigm – in the sense of Kuhn – for the area of LRs.

New types of initiatives are now underway, such as: a) the EC LIRICS (e-Content) project, aiming to provide ISO ratified standards for LRs & LT, b) the Unified Lexicon project – by ELRA and its Production Committee – linking the LC-Star and PAROLE lexicons to set up a methodology to connect Spoken and Written LRs, and thus establish common standards and new models of LR distribution, or c) the new NEDO Japanese project for developing international standards of LRs for Semantic web applications, specifically geared to Asian languages but with the cooperation of Asian and a European partner.

2. How to shape the future?

We must build on the set of accumulated experience – and data – we have gained so far, but – exactly because of the massive amount of knowledge and data we have been able to gather – we must also reflect if today situation does not require, to make a real step further, a deep change of perspective and a new vision.

2.1. Roadmap for LRs

In recent consultations about LRs, such as the ELSNET/ENABLER Roadmap workshops (Paris, 2003 and Lisbon, LREC2004), a first list of priorities which act as critical issues for the future of LRs was drawn:

- define and provide basic LR coverage for all languages (BLARK/ELARK concepts);
- significantly increase multilingual LRs;
- develop an "Open Source" concept for LRs;
- coordinate the design and creation of LRs (also across languages) with a view to interconnectivity and reusability, to enhance LR content interoperability;
- enhance metadata infrastructure and standards;
- give high priority to methods and tools to quickly develop LRs "on demand" (acquisition, annotation, merging, porting between domains or languages, …), a particularly important issue for industrial exploitation;
- develop LRs for evaluation purposes, and define validation methodologies and protocols for LRs;
- foster synergies between spoken and written areas and with neighbouring areas (e.g. terminology, Semantic Web);
- investigate IPR issues.

2.2. Some LR priorities and challenges

For a better organised field many challenges exist, at various levels of complexity and with various priorities and weights, both at technological and organisational level. I mention some and quickly touch a few:

- Overcome the usual *mismatch between advancement in LRs and in LT*.
- Design *lexicons as dynamic resources* whose content is co-determined by automatically acquired linguistic information from text corpora and from the web. We should push towards innovative types of lexicons: a sort of 'example-based living lexicons' that participate of properties of both lexicons and corpora.
- Eliminate the *lack of communication between the communities of LRs/LT and Semantic Web(SW)/Ontologies*. LT will highly benefit from the SW but the SW needs LT, otherwise there is a clear risk of 're-discovery' of what was done 20 years ago.

  Examples of relations from LRs/LT to SW:
  - *Semantic mark-up*: for the SW task of adding meaning to Web data and make it usable for automatic processing.
  - *LRs as the basis for knowledge representation and sharing*, for interoperability among knowledge based systems.
  - *Ontology learning, ontology design and evaluation of ontologies*: LT is mature enough to be a core technology for the extraction and creation of semantic content.

  Examples of relations from SW to LRs/LT:
  - *LRs/LT as web services, and use of SW representation formalisms*: the SW may crucially determine the shape of the new generation of LRs of the future, consistent with the vision of an open space of sharable knowledge available on the Web for processing.
  - *Open access paradigm, semantic interoperability, information integration*: this is – in my vision – the real target for the next decade for LRs, and implies a complete re-thinking of the current area of LRs.

I'd like also to mention a few types of LRs that should receive attention in the next years.

- New types of *"example-based" context sensitive LRs, Lexicon and Corpus together*, dynamically created.
- The *Web exploited as a multilingual corpus*.
- *Facts and commonsense knowledge*, built in distributed and collaborative way by the community.
- Common sense in *affective classification of text*.

And we cannot forget two issues often neglected:

- *Knowledge transfer across languages*, to take advantage of LRs built for few resource-rich languages and induce knowledge in languages with few LRs.
- *Maintenance of LRs* (updating, tuning, etc.): it is still a big issue that deserves to be organised.

3. LRs in the future HLT

Focusing our view into the future of LRs, a radical modification of perspective is needed, to facilitate integration of linguistic information resulting from all LR initiatives, bridge differences between various standpoints on language structure and linguistic content, put an infrastructure into place for content description and interoperability at European level and beyond, and make LRs usable within the emerging SW scenario [7].

3.1. A new paradigm for LRs

The need of ever growing LRs for effective multilingual content processing requires a change in the paradigm, and the design of a "new generation" of LRs, based on open content interoperability standards. SW developers will need repositories of words and terms, machine-understandable knowledge about their relations within language use and ontological classification. The effort of making available millions of 'annotated words' for dozens of languages is something that no single group

is able to afford. This objective can only be achieved when working in the direction of an integrated *Open and Distributed Linguistic Infrastructure*, where not only the linguistic experts can participate. It is already proved by a number of projects that lexicon building and maintenance can be achieved in a cooperative way. We claim that the field of LRs and LT is mature enough to open itself to the concept of collaborative effort of different sets of communities (e.g. spoken and written, LT and SW, theoretical and application oriented).

3.1.1. Open and distributed architectures for LRs and LT, interoperability, GRID technology

A new paradigm of R&D in LRs and LT is emerging, pushing towards the creation of open and distributed linguistic infrastructures for LRs and LT, based on sharing LRs and tools. It is urgent to create a framework – both technological and organisational – that enables controlled and effective cooperation of many groups on common tasks, adopting the paradigm of accumulation of knowledge so successful in more mature disciplines, such as biology and physics. This implies the ability to build on each other achievements, merge results and have them accessible to various systems and applications. This is the only way to make a clear leap forward. This means emphasizing interoperability among LRs, LT and knowledge bases. Standards are again unavoidable.

This may also mean application of *GRID technology* to tackle the problems of processing extremely large quantities of "facts and their relations", of development of unprecedented large-scale annotated LRs, and of their dynamic linking across many different sources. A difficulty and a challenge is how to coordinate different information sources.

A way to attain the optimisation of the process of production and sharing of (multilingual) LRs relies on a public and standardized framework ensuring that linguistic information is encoded in such a way to grant its reusability in different tasks and applications. The ENABLER [6] project promoted the compatibility and interoperability of LRs endorsing: i) ISLE/EAGLES (http://www.ilc.cnr.it/EAGLES96/isle/), for harmonisation of linguistic specifications, in particular for corpora and multilingual lexicons; ii) ISO TC37 SC4 WG4, to make European standards truly international Standards; iii) ELRA Validation Committee, for integration of standards in protocols for LR validation; iv) INTERA, for harmonisation of metadata descriptions; v) cooperation with Semantic Web communities, to encourage synergy between knowledge management/ontology and HLT/LRs.

3.1.2. Lexicons' integration and interoperability: concrete steps towards a cooperative model

The SW model of open data categories will foster LR integration and interoperability, through links to common standards. With the ISLE approach to lexical standards, and its definition of the MILE (Multilingual ISLE Lexical Entry) [8], new lexical objects can be progressively created and linked to a core set. An increasing number of linguistic data categories and lexical objects stored in open and standardised repositories will be shared and used by different types of users to define their own structures within an open lexical framework.

It will guarantee freedom for the user to add or change objects if that is deemed necessary, but will require an evaluation protocol for the core standard lexical data categories, and verification methods for the integration of new objects. This vision, enabled by MILE, will pave the way to the realisation of a common platform for interoperability between different fields of linguistic activity – such as lexicology, lexicography, terminology – and SW development. The lexicons may be distributed, i.e. different building blocks may reside at different locations on the web and be linked by URLs. This is strictly related to the adoption of SW standards (e.g. RDF metadata to describe lexicon data categories), and enables users to share lexicons and collaborate on parts of them.

In our group we have recently developed LeXFlow, an architectural and practical framework for dynamic semi-automatic integration of lexicons and LRs [9]. LeXFlow is a system – based on XML – that manages lexical workflows where the different agents can reside over distributed places, and thus enables new methods for cooperation among lexicon experts, through collaborative management on various lexicon operations.

4. Technical vs. organisational/strategic issues for a LR infrastructure

The approach to realise a true LR infrastructure requires the coverage not only of a range of scientific and technical aspects, but also organisational, coordination, strategic and political issues play a major – and maybe most critical – role, as was highlighted in the ENABLER project [10]. They in fact acquire a more and more decisive relevance with the growing maturity of the LR field. Existing experience in LR development proves that such a challenge can be tackled only by pursuing – on the *organisational* side – a truly interdisciplinary and cooperative approach, and by establishing – on the *technical* side – a highly advanced environment for the representation and acquisition of linguistic information, open to the reuse and interchange of linguistic data.

We should promote together the launch of a large initiative, comprising the major LR and HLT groups in Europe and world-wide, for the creation of an open and distributed infrastructure for LRs. The outcome of such an initiative could be the design of a completely new generation of LRs.

Linked to this idea, an important *Declaration on Open Access to LRs* was endorsed by all participants of an ENABLER/ELSNET Workshop held in Paris in 2003.

4.1. ELRA role in the field of LRs

The availability of LRs is also a "sensitive" issue, touching directly the sphere of linguistic and cultural identity, but also with economical, societal and political implications. This is going to be even more true in the new Europe with 25 languages. Coordination should be established between EC and member states, and strategies should be drawn in order to ensure a proper balance of language coverage in Europe. To this end ENABLER and ELRA have adopted and strongly supported the BLARK (Basic LAnguage Resource Kit) concept [11].

A Linguistic Infrastructure intends also to contribute to the structuring and integration of the European Research Area, addressing problems such as the fragmentation of its research base and the weakness in converting R&D results into useful economic or society benefits. To this aim, we claim it is necessary to pool together and build on many

different, but related, initiatives both for Spoken and Written LRs.

International cooperation will be certainly the most important factor for a coherent evolution of the field of LRs – and consequently of HLT – in the next years. A report produced by ELDA [12] presents an analysis of several organisational frameworks, focusing on funding and organisational procedures to provide LRs. ELRA [13], as a promoter of infrastructures for LRs, has in its mission also production and validation of LRs and promotion of standards. The Unified Lexicon project [14] of the Production Committee, defining common standards for spoken and written LRs, aims at overcoming existing barriers among independently built spoken and written LRs. It is the first step to pave the way to innovative methods of tailoring and acquiring LRs starting from available repositories, based on individual requirements. It can be seen as a contribution to solving the current fragmentation of LRs, while capitalising on and reusing results from previous European and national projects and standardisations activities.

4.2. Cooperation among communities

Technologies exist and develop fast, but the infrastructure that puts them together and sustains them is still largely missing. For example, the absence of a specific HLT action line in the European FP6 means not so much a change in the funding scene, but – more dangerous – lack of opportunities to discuss meta-level issues on HLT, difficulty in designing common global long-term strategies, with the risk of being just opportunistic in R&D choices. While there is a pressing need of international research infrastructures for LRs and LT, of bodies where to discuss a broad research agenda, priorities and strategic actions for multilingual and multimedia LRs and LT. To achieve this, cooperation must be enhanced among many communities acting now separately, such as LR and LT developers, terminology, SW and ontology experts, content providers, linguists, humanists. This is one of the challenges for the next years, for a usable and useful "language" scenario in the global network. The implementation of the notion of open distributed infrastructures for LRs and LT could act as a major technological and organisational challenge around which synergies (with other communities) can develop, and can naturally lead to the creation of an International Forum where to discuss about strategies and priorities.

A warning is due: such a language infrastructure may turn into being inherently market driven, since the most widely used language portions may become the best developed and supported. This deserves serious reflection for the political implications.

The idea behind such (past and future) initiatives is to establish some sort of permanent coordination to build on parallel existing (national or international) initiatives. At the end everything is tied together, which makes our overall task so interesting – and difficult. What we must have is the ability to combine the overall view with its decomposition into manageable pieces. No one perspective – the global and the sectorial – is really fruitful if taken in isolation. A strategic and visionary policy for cooperation between various groups has to be debated, designed and adopted for the next few years, if we hope to be successful, but – inside this – a realistic and stepwise approach to solving well-defined and limited aspects must be adopted. To this end, the contribution of the main actors from the various areas involved is of extreme importance. This will be a must for our field to contribute, effectively and globally, to the big challenges of the 'knowledge-based society'. Some of the events of the last years are hopefully moving in this direction.

5. References

[1] Zampolli A., "Towards Reusable Linguistic Resources", *EACL 1991, 5th Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, 1991.

[2] Calzolari, N., "Lexical databases and textual corpora: perspectives of integration for a Lexical Knowledge Base", U. Zernik (ed.), *Lexical Acquisition: Exploiting on-line Resources to build a Lexicon*, Lawrence Erlbaum Associates, Hillsdale, NJ, 191-208, 1991.

[3] Calzolari, N., "An Overview of Written Language Resources in Europe: a few Reflection , Facts, and a Vision", Rubio, A., Gallardo, N., Castro, R., Tejada, A. (eds*.), Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada, Vol. I, 217-224, 1998.

[4] Calzolari, N., Zampolli, A., "Harmonised large-scale syntactic/semantic lexicons: a European multilingual infrastructure", *MT Summit Proceedings*, Singapore, 358-365, 1999.

[5] Ide, N., Calzolari, N., "Introduction to the Special Inaugural Issue", *Language Resources and Evaluation*, Springer, 39(1):1-7, 2005.

[6] Zampolli, A. et al., *ENABLER Technical Annex*, Pisa, 2000.

[7] Calzolari, N, "Computational Lexicons: Towards a New paradigm of an Open Lexical Infrastructure?", G. Willée, B. Schröder, H.C. Schmitz (eds.), C*omputerlinguistik. Was geht, was kommt?. Computational Linguistics. Achievements and Perspectives*, Gardez!, Sankt Augustin, 41-47, 2002.

[8] Calzolari, N., Bertagna, F., Lenci, A., Monachini, M. (eds.), *Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entry)*, ISLE CLWG Deliverables D2.2&D3.2, Pisa, 194 pp., 2003.

[9] Soria, C., Tesconi, M., Bertagna, F., Calzolari, N., Marchetti, A., and M. Monachini. 2006. "Moving to Dynamic Computational Lexicons with LeXFlow". *Proceedings of LREC2006*, Genova, Italy, 2006.

[10] Calzolari, N., Choukri, K., Gavrilidou, M., Maegaard, B., Baroni, P., Fersøe, H., Lenci, A., Mapelli, V., Monachini, M., Piperidis, S., "ENABLER Thematic Network of National Projects: Technical, Strategic and Political Issues of LRs", *LREC 2004 Proceedings*, Lisbon, 937-940, 2004.

[11] Mapelli, V., Choukri, K., "Report on a (Minimal) Set of LRs to Be Made Available for as Many Languages as Possible, and Map of the Actual Gaps", ENABLER Deliverable D5.1, Paris, 2003.

[12] Mapelli, V., Choukri, K., "Report Contributing to the Design of an Overall Co-ordination and Strategy in the Field of LRs", ENABLER Deliverable D5.2, Paris, 2003.

[13] Maegaard, B., Choukri, K., Calzolari, N., Odijk, J., "ELRA - European Language Resources Association. Background, Recent Developments and Future Perspectives", *Language Resources and Evaluation*, Springer, 39(1):9-23, 2005.

[14] Monachini, M., Calzolari, N., Choukri, K., Friedrich, J., Maltese, G., Mammini, M., Odijk, J., Ulivieri, M., "Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian". *Proceedings of LREC2006*, Genova, Italy, 2006.

Organization Models for Research Infrastructure and existing infrastructures

Bente Maegaard

Center for Sprogteknologi, University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen
bente@cst.dk

Abstract

Since the 1990es, various organizations have been taking care of the distribution of language resources for research and commercial applications. Technical developments recently have opened new possibilities; how do we organize ourselves in the future? do we need new organizations, or modifications to existing ones?

1. Background

Ever since computers were born, there has been a need to collect and analyze language resources. Most of the very first applications of computers were corpus investigations. Some early investigations worked on what was felt then to be pretty large corpora, e.g. the Brown Corpus of 1 million words (Kucera & Francis 1967), and Kierkegaard's works of 2 million words (McKinnon 1965).

At that time, dedicated researchers typed in text themselves, or rather, raised money to have text typed. When the corpus was available, it was often accessible to only one or a few researchers for their own work. But even if they wanted to share with others, the technical means, the infrastructure and the copyright problems were too important obstacles. The users were not only linguists, but also historians, philologists etc.

1.1. The Index Thomisticus

As an example of a very early resource project, which was not necessarily linguistic, let us consider the Index Thomisticus. In 1946 Father Busa planned the Index Thomisticus, as a tool for performing text searches within the massive corpus of Aquinas's works. In 1949 he met with Thomas Watson Sr., the founder of IBM, and was able to persuade him to sponsor the Index Thomisticus. The project lasted about 30 years, and eventually produced in the seventies the 56 printed volumes of the Index Thomisticus. In 1989 a CD-ROM version followed, and a DVD version is underway. In addition, in 2005 a web-based version made its debut, sponsored by the Fundación Tomás de Aquino and CAEL. This is an example of a huge amount of work which was sponsored from the very beginning, and which was shared first through printed books and then through CD distribution, - when the technical possibilities were available.

1.2. Infrastructure in the previous century – what is the problem?

The infrastructure problems mentioned above include e.g. the fact that as soon as a resource is to be distributed, it needs to be in good shape: a clean version has to be made, and it has to be accompanied by documentation in some widely known language etc. Still today, the reason that many resources are not distributed is that it takes some energy to prepare them for distribution, and this is work that has to be done by those who produced the resource and therefore know it. However, the efforts that are needed to prepare a resource for sharing are much smaller than the effort to build the resource again, so researchers should be encouraged to make this last investment in their resource.

For commercial applications, the problems are different. If a company has built a resource they do not necessarily want to share it with others, as the resource may provide a competitive advantage. Below, we are focusing on research use of resources.

For research, shared resources provide benefits that 'private' resources do not, apart from the fact that more researchers can use the same resources. Shared resources also permit replication of published results, support fair comparison of alternative algorithms or systems, and permit the research community to benefit from corrections and additions provided by individual users.

2. Existing infrastructures

As an answer to this arising understanding of the possibilities in shared resources, two organizations were established in the 1990es, LDC and ELRA. We first present LDC shortly, and then go into more details with ELRA.

2.1. LDC

The Linguistic Data Consortium (LDC) was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that now includes more than 100 companies, universities, and government agencies

The Linguistic Data Consortium is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes. The LDC was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by grant IRI-9528587 from the Information and Intelligent Systems division of the National Science Foundation.

2.2. ELRA

In Europe, the European Language Resources Association (ELRA) was founded in 1995.

Antonio Zampolli was the main driving force behind the creation of ELRA. The starting point was the realisation that the development of language technologies

was crucially dependent on the capability of processing large quantities of 'real' texts and on the availability of large-scale lexicons. This gave rise to the so-called 'reusability' notion which was at the basis of many initiatives for establishing standards and best practices.

This trend arose also from the increasing interest of national and international authorities in the potential of the so-called 'language industry'. The path went through a wide range of language resources (LR) projects, most of them financed by the European Commission (EC), both projects that aimed at developing LRs, and projects that were of a more political and coordinating nature. Within the EC Language Engineering (LE) program there was a very fruitful combination of LR, language technology and application projects, recognising the natural links among these aspects and need for them to proceed in parallel, in synergy, and in a coherent way.

Zampolli clearly delineated the major strategic lines of activity:

- elaboration of consensual standards,
- creation of the necessary LRs,
- distribution and sharing of LRs,
- creation of synergies among national projects, European and international projects, industrial initiatives.

In order to carry out such a strategic analysis, A. Zampolli, together with a large number of key players in the European language technology field proposed to the European Commission to launch a project called RELATOR - A European Network of Repositories for Linguistic Resources (1993-95). The project aimed at defining a broad organisational framework for the creation of the LRs, for both written and spoken language technology, which are necessary for the development of an adequate language technology and industry in Europe. It also aimed at determining the feasibility of creating a coordinated European network of partners that would perform the function of storing, disseminating and maintaining such resources.

The major outcome of RELATOR was the creation of ELRA as well as the initiation of several Language Resource production projects (e.g. SpeechDat family, PAROLE/SIMPLE, POINTER, etc.). The RELATOR project presented final recommendations for establishing a collaborative infrastructure that would act as a collection, verification, management and dissemination centre, built on the foundation provided by existing European structures and organisations. RELATOR proposed the foundation of a European Association for Language Resources, which was registered in Luxemburg (**ELRA - European Language Resources Association**) in February 1995. ELRA was established as an independent, not-for-profit, membership-driven association. ELRA was supported by the European Commission through project funding in the first years, but has been self-supporting since 1998.

ELRA's initial mission was to set up a centralised not-for-profit organisation for the collection, distribution, and validation of speech, text, terminology resources and tools. In order to play this role of a central repository, ELRA had to address issues of various nature such as technical and logistic problems, commercial issues (prices, fees, royalties), legal issues (licensing, Intellectual Property Rights), and information dissemination. ELDA (Evaluation and Language Resources Distribution Agency) was established as the operational unit of ELRA.

The mission of ELRA is to promote language resources and evaluation for the Human Language Technology (HLT) sector in all their forms and all their uses, in a European context. Consequently the goals are: to coordinate and carry out identification, production, validation, distribution, standardisation of LRs, as well as support for evaluation of systems, products, tools, etc..- related to language resources.

3. New challenges

Lately, the field of computational linguistics has seen a number of new developments.

New types of resources are needed for language technology research and applications, e.g. multimodal resources. At the same time other fields of application than computational linguistics and language technology are seeing the advantages of computational access to resources, - history, philosophy, music, literature etc. This means that other types of resources have to be made available. Knowledge of the field is necessary to make the right resources available in the right form.

Another development is the presence of the Internet with masses of data. The Internet has become a major source of data for many researchers, and this will certainly continue. However, even if for some applications this type of data is acceptable, the data do not come with quality assurance, and e.g. free lexica on the Internet are not of the quality needed for most applications. Also, there are copyright issues to be solved when data are taken from the Internet. It can be assumed that quality of what is available on the Internet will grow as it has done until now, but to solve the copyright issues a political effort is necessary.

The Internet and GRID technologies also provide new possibilities for distribution of data. ELRA has e.g. almost exclusively been using CD, because many resources are too large to be downloaded through the web, - but new technology will change this.

4. Organizational models

The existing structures, LDC and ELRA, are different:

LDC is a consortium that an organization (university, company) may join by paying a subscription fee. The organization then receives all resources built during the year of subscription.

ELRA is a member-driven association. Members pay the membership fee, and may purchase resources at reduced prices. A good deal of the research resources are extremely cheap, but ELRA also provides resources for industry which are more expensive.

The difference between LDC and ELRA are 1) the consortium vs. association, 2) the fee structure.

LDC and ELRA have more similarities than differences: they both provide a legal framework for copyright and licensing issues, they both maintain a catalogue of available resources, they both support the development of and adherence to standards, they both ensure some kind of quality in their resources. Both entities also identify new interesting resources for their customers.

ELRA has set up formal procedures for validation of resources and made the validation manuals public. ELRA

is promoting the concept of validation, also the internal validation at universities or in companies.

ELRA has also been working on a 'universal catalogue'. ELRA's catalogue contains information about the resources provided by ELRA, whereas the universal catalogue contains information about resources identified that might be of interest to the community. The universal catalogue is at present a membership advantage.

Organizational models need to take into account that there is a cost to pay for the management of resource identification, archiving, .licensing, distribution and validation. For some resources some of these items can be free, or almost free, - e.g. the management of free resources can be dealt with in a very light way, by enabling access to free resources etc. This is one of the developments ELRA is considering.

ELRA is open to collaboration with other organizations, sharing the acquired expertise in an active partnership. E.g. a collaboration with the proposed CLARIN initiative should be explored.

5. HERA

As a very last point, we should mention the European HERA initiative (Humanities in the European Research Area). The text below is taken from the HERA project description at the EU CORDIS web site.

"During the ERA-NET Specific Support Action in 2004, The European Network of Research Councils in the Humanities (ERCH) has taken several initial steps towards large-scale cross-border coordination of research activities within the humanities. The network has now in cooperation with the European Science Foundation decided to continue the efforts under a new name: Humanities in the European Research Area (HERA). Building on the ERCH work, the HERA Coordination Action will be an extension of the network, in scope as well as depth. Firstly, the Consortium is being extended from three to fourteen members and, secondly, the range of activities is being widened to cover coordination of research activities, including the setting-up of joint research-funding initiatives. The main tasks of the CA will be:

- Consolidation of the network by establishing new network structures and integrating new members.
- Exchange of information and best practice on issues such as peer review, programme management, quality and impact assessment, and benchmarking.
- The development of research infrastructures within the humanities, which will pave the way for greater efficiency and enable new perspectives by ensuring accessibility and availability for of data and information in the widest sense.
- The ultimate objective of the CA-proposal is to coordinate research programmes in a cumulative process leading to the initiation of joint research-funding initiatives.

By applying comparative perspectives to humanities research and enabling new transnational funding schemes, it will be possible to transcend the traditional, national focus of humanities research."

It seems that it will be beneficial to explore the possibilities of cooperating with the HERA initiative, if a larger initiative covering the humanities is to be explored.

6. Acknowledgements

This paper draws upon work done in the ELRA Board, of which I am the president. In particular I want to mention contributions by Jan Odijk, Nicoletta Calzolari and Khalid Choukri.

7. References

Busa, Roberto (ed.): *Index Thomisticus,* Stuttgart, 1974, 56 volumes.

Kucera & Francis: *Computational Analysis of Present-Day American English*, Brown University Press, 1967

LDC web site, http://www.ldc.upenn.edu/

Maegaard, Bente, Khald Choukri, Nicoletta Calzolari, Jan Odijk: ELRA – European Language Resources Association – Background, Recent Developments and Future Perspectives. In: *Language Resources and Evaluation vol. 39*, Springer 2005, pp. 9-23.

McKinnon, Alastair: Computational Analysis of Kierkegaards Samlede Værker, Vol IV, Brill, Leiden, 1975

The relevance of standards for research infrastructures

Gil Francopoulo[1], Thierry Declerck[3],

Monica Monachini[2], Laurent Romary[4]

[1]INRIA-Loria: gil.francopoulo@wanadoo.fr
[2]CNR-ILC: monica.monachini@ilc.cnr.it
[3]DFKI: declerck@dfki.de
[4]INRIA-Loria: Laurent.Romary@loria.fr

Abstract

In this paper, we show the importance of standards as an essential aspect for any research infrastructure in the humanities. In the context of the current activities within ISO committee TC 37/SC 4 (Language Resource Management), we show in particular how important it is to provide means to compare linguistic representations through the use of a shared semantics for elementary descriptors. This is further exemplified by describing the ongoing work to define a central *data category registry*, which aims at being a reference point in the language resource community, in conjunction to the definition of basic standards for linguistic annotation, as illustrated with the current work that is being carried out in the domain of morpho-syntactic categories.

1. Standards: are they at all needed ?

For many years, the language resource community has been the place of numerous projects (see Cole et alii, 1997) that have aimed to produce resources and tools to facilitate the study or automatic processing of language. Still, we have all faced the issue of ensuring long-term availability of the corresponding results, with the consequence that researchers still have to carry out technical tasks of corpus gathering, lexical description or tool implementation that others are supposed to have achieved beforehand, and above all that should be the duty of shared research infrastructures working for the benefit of all.

One of the key issues to define such research infrastructures is our ability, as a mature scientific community, to be able to identify that new research results should be based upon the stabilization of shared knowledge by means of a range of internationally agreed upon standards. Such standards would obviously bring the following benefits:

- Ensure wide accessibility of data in space (between research sites) and time (in the perspective of providing long-term preservation of data). Standards are there to provide a stable representational basis as well as maintained documentation, that researchers are not able to produce on their own;
- Facilitate the reusability of software by making it independent from the actual proprietary data formats an implementer might use;
- Guaranty that research results are comparable, by, for instance, making sure that the same underlying data has been used in the context of the elicitation of statistical results;
- Create communities of practice that will share the knowledge of such standards and create new concepts on the basis of this common culture.

As a matter of fact such benefits have already been observed in the context of the wide deployment of the Text Encoding Initiative guidelines, which have both been the basis of numerous projects worldwide[1], but also have been the basis of a shared understanding of basic textual descriptions that now leads to the explorations of new textual types or phenomena[2].

Still, the language resource community requires even more standards to cope with both the variety of linguistic phenomena that have to be taken into account as well as the diversity of human languages. This is why, a the International Organization for Standardization[3] has put together a new committee dedicated to language resources, known as ISO/TC 37/SC 4 and started to foster several standardization projects to deal with what has been identified as priorities for the progress of the management of language resources.

In the remaining sections, we first provide a few elements related to the role we think research infrastructures should play with regards standards. We then outline the working agenda of ISO/TC 37/SC 4 and we present our opinion concerning standards when applied to Research Infrastructure (RI). Then, as an illustration, we present the work in progress within ISO-TC37/SC4 on the morpho-syntactic profile of the data category registry (DCR).

2. Research infrastructures and standards

As we have seen, standards are an essential component of any language resource related activity. In this context research infrastructures should consider standardization as one essential point of their activities. More precisely we consider that at least the three following missions should be allocated to research infrastructures:

- They should contribute the wide dissemination of standards by initiating training sessions and providing teaching materials and samples on line;
- They should actually implement available standards in all their activities, with the constant objective of

[1] See the TEI projects page under http://www.tei-c.org/Applications/
[2] See the P5 edition of the guidelines: http://www.tei-c.org/P5/
[3] http://www.iso.org

long-term availability of the data or tools they produce (see above);

- They should be at the forefront of standardization activities by explicitly reviewing existing standards, contribute to their evolution and even participate to the definition of new standards when needed by the corresponding research community.

3. Work in progress within ISO-TC37

ISO committee TC 37/SC 4 is dedicated to the specification of a full family of standards for NLP and language resources. These standards can be categorized according to two levels:

Low level standards, describing the linguistic constants. More precisely, this is a pair:

a) revision of ISO-12620 that specifies the rules for describing and maintaining data categories.

b) data category registry

There are also some other important low-level standards that we can use: the standards for character encoding (ISO/IEC 10646 i.e. Unicode), language codes (ISO-639), script codes (ISO-15924), country codes (ISO-3166) and dates (ISO-8601).

High level standards, describing structural models (sometimes called meta-models) that specify how to represent linguistic resources. The structural model provides classes (in UML terminology) and the relations between classes together with a textual usage description for each class.

The registry provides the needed attributes and values that are used **to adorn the classes**. The structural models being currently developed deal with word-segmentation, morpho-syntactic annotation (aka MAF), syntactic annotation (aka SynAF) [1] and lexicon (aka LMF) [2].

4. Objective

The objective is to propose to the user and developer of language resources a coherent family of standards. All these standards have the following property: they allow the definition of a model of linguistic resource by combining structural elements with constants taken in low-level standards. All the resources share thus the same set of constants, supporting our goal of providing interoperability between segmentation, annotation and lexicon.

5. Roadmap

As said before, the duration for defining an ISO standard is rather long. It takes around four years. So, instead of defining low-level standards then high level standards (or the contrary), the various ISO groups works in parallel with a closed collaboration between them.

6. Some basic definitions

6.1. A data category

A data category is a linguistic constant. A data category is either an attribute name like /partOfSpeech/ or a value dedicated to populate an attribute. An example of value is /noun/.

6.2. Profiles

A profile is a specific set of data categories in the DCR.

The current profiles are:

For Terminology within TC37/SC3

One profile

For NLP within TC37/SC4

Three profiles:

Meta-data

Morpho-syntax

Semantics

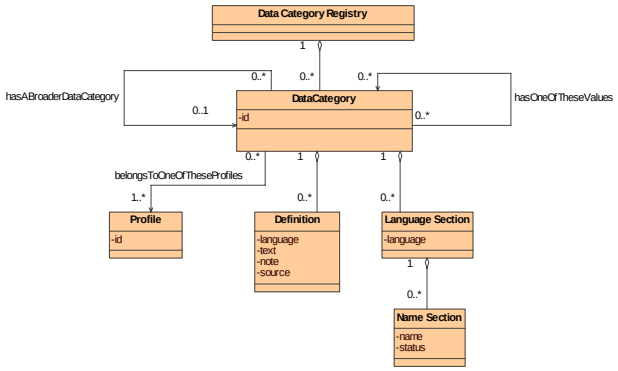
You can notice that to ensure interoperability in NLP between word-segmentation, annotation and lexicon, the distinction between each profile is made according to linguistic criteria and not according to the resources. Another point to mention, is that a data category may belong to several profiles but we try to avoid this situation in order to avoid conflicts.

6.3. The data category registry

The registry is the union of all data categories.

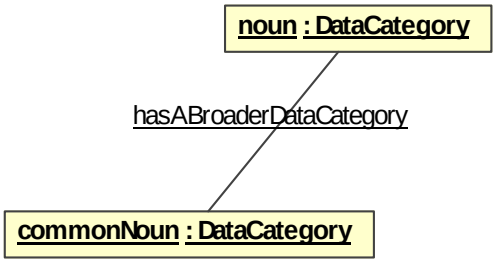
7. Morpho-syntactic profile

The DCR structure is specified by the ISO-12620 revision. In the morpho-syntactic profile we restrict ourselves for the time being to the following features:

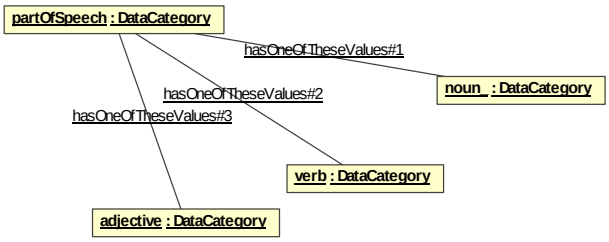


We differentiate between the notion of /broader/ relation and the notion of /conceptual domain/.

The /broader/ link allows a hierarchy of constants to be defined. Example: a common noun is a more specialized value than noun.



The notion of conceptual domain allows a set of valid values to be identified. Example: noun is a value for partOfSpeech.



We proceeded in three phases:

Phase-1: collect

Phase-2: group, structure and write a first draft of the definitions

Phase-3: revise

An initial long and flat list of data categories has been collected from:

- Current ISO-12620
- Eagles and Multext-East

8. What has been done in the morpho-syntactic profile?

- A couple of values for the NLP sections in LMF

The ISO-12620 constants are general purpose values like /language/ or /derivation/ and cover only terminological resources. For instance, for /partOfSpeech/, the only values are /noun/, /adjective/ and /verb/. By comparison, in NLP, we need much more values including /preposition/ and /pronoun/ etc.

We propose a set of constants according to the following criteria:

- broad linguistic coverage within the morpho-syntactic perimeter
- no semantic overlap
- good choice of a name associated with a good textual definition

9. What has been recorded so far in the DCR?

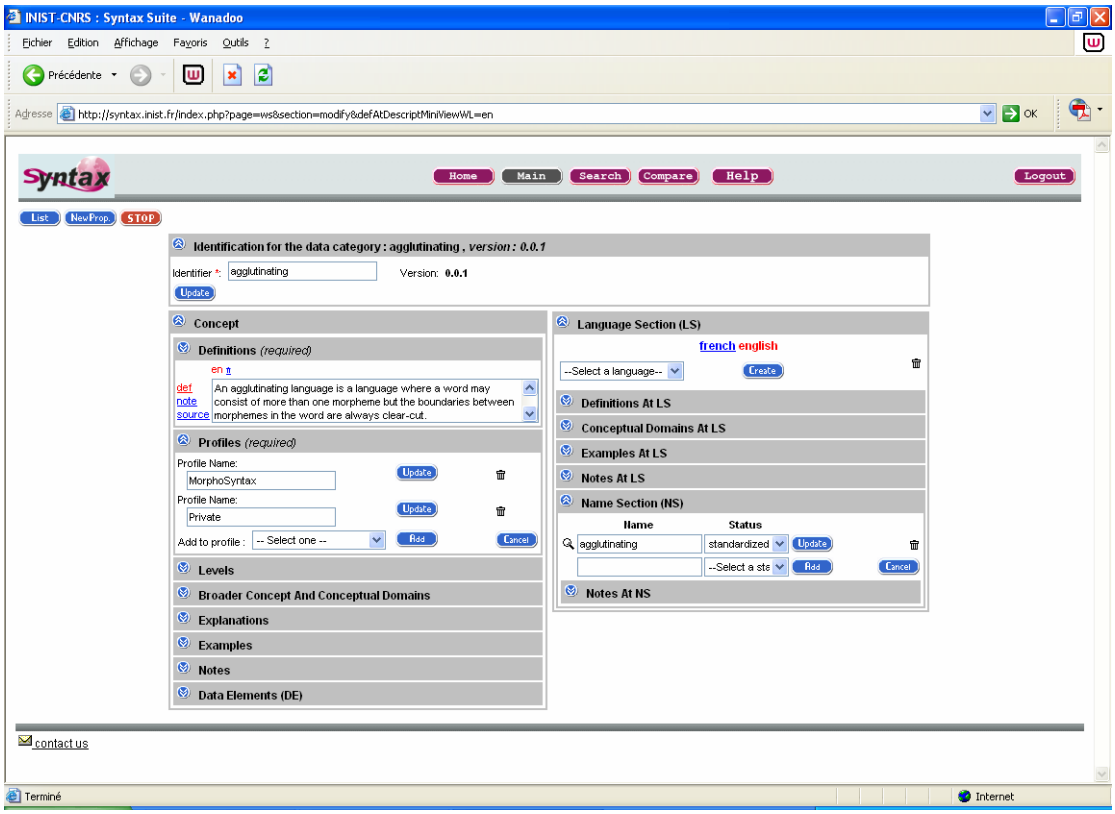
The list being rather huge we created 11 directories within the Syntax software (see next section) in order to help data category organization. It easier to work on medium sized list than on a list with 300 items.

In each directory: one or several attributes names and related values are recorded.

Basics		29
	These are general purpose linguistic constants, like: comment, derivation, elision, foreignText, label.	
Cases		33
	Examples of values: ablativeCase or dativeCase.	
FormRelated		33
	These are constantes for the specifications of forms like: spokenForm, writtenForm, abbreviation, expansionVariation, transliteration, romanization, transcription, script.	
Language Typology		4
	An attribute is languageTypology and values are agglutinating, inflectional and isolating.	
Morphological Features excluding cases		72
	Attributes are for instance grammaticalGender, mood and tense. Values are for instance feminine, indicative, present.	
Operations		8
	The constants are for instance addAfter, addBefore, copy etc.	
Part of speech		93
	The part of speech values are structured with a top level set composed of 10 values like noun or verb. A very precise ontology is specified for grammatical words. Most of parts of speech are common to lexicons and annotations but two set of values (i.e. punctuation and residual) are specific to annotation and are not usually used in lexical descriptions.	
Reference		5
	The constants are anaphora, antecedent, cataphora, coreference, endophora and referent. This is some doubt to maintain these constants in the morpho-syntactic profile.	
Register, dating and frequency		19
	The constants are slangRegister or rarelyUsed.	
Semantically motivated		16
	The constants are agent, intensive. This is some doubt to maintain these constants in the morpho-syntactic profile.	
Syntactically motivated		36
	Attributes are function or voice. Values are subject, activeVoice for instance.	
Total		348 items

10. Software

We use the Syntax software hosted by CNRS-INIST in Nancy (see http://syntax.inist.fr) in order to edit the data categories. This is a server based on a relational database with a set of PHP programs in order to manage the interaction. Here is a screen dump:



11. APIs

In order to allow programs to access to the DCR, a set of Application Programming Interfaces are being specified and implemented by Max Planck Institute for Psycholinguistics of Nijmegen, INRIA-Loria and University of Sheffield.

12. Acknowledgements

The work presented here is partially funded by the EU eContent-22236 LIRICS project[4], and by the French TECHNOLANGUE program[5].

13. References

Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V. (Eds.) 1997. *Survey of the State of the Art in Human Language Technology*, First Edition – 1997, Cambridge University Press.

Declerck T. 2006 SynAF: towards a standard for syntactic annotation LREC Genoa.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF). LREC Genoa.

[4] http://lirics.loria.fr
[5] http://www.technolangue.net

DAM-LR as a Language Archive Federation: strategies and prospects

David Nathan and Remco van Veenendaal

SOAS, London; INL, Leiden
djn@soas.ac.uk

1. Introduction

The DAM-LR partners are well on the way to forming a federation. The term 'federation' has at least two quite different meanings and it is important not only to distinguish them but also to put our own stamp on what kind of federation we create.

'Federation' has a specialised meaning in information technology, referring to bringing information resources together via information management and networking techniques. It also has an organisational meaning, referring to agencies and their aims and strategies for collaboratively dealing with identities, resources, and services. In this paper, we refer to 'federation' in the first domain as federationT ("federation technologies") and in the second domain as federationA ("federation agency/-ies").

2. FederationT: a background

Federations in the IT sense go back to the earliest days of electronic networking. For example, in 1967 the Online Computer Library Center (OCLC, http://www.oclc.org/) at Ohio State University started sharing bibliographic authority files with fellow libraries, and has long been involved with the issues that still face us now: standards, metadata, quality levels, technology, membership and collaboration. The OCLC now has 9,000 members who share 65 million records to assist in their librarianship work. By the mid 1990s the term 'federated database' was well known. Dempsey et al, for example, describe a "federating solution … [that] allows services to develop autonomously while projecting a single unified image to the user". The motivation for federating resources is to provide value to users by providing a larger metadata set with a correspondingly greater ability to "relieve … potential users of having to have full advance knowledge" of the existence or nature of resources (Dempsey et al). According to this definition, search engines such as Google also, in a weak sense, federate all the web pages that they index.

Lynch (1998) refers to Dublin Core (DC) – also with its roots in libraries – as a tool for federating existing resources: "networked information discovery and retrieval [through] federating disparate, independently maintained databases … [via] a common semantic view of the various databases involved". DC was intended to enhance resource discovery in an open networked world, i.e. in a world without librarian operated-catalogues where quality and consistency are principal values and practices. Dublin Core provides, then, a low-density, lowest-common-denominator but unified method for description and discovery in a unified search domain (the www) by a wide variety of professionals, data-creators and ordinary users. To achieve this, the DC consortium dealt with issues such as (i) syntactic standards e.g. for data and metadata encoding (ii) semantics, e.g. ontologies, semantic web etc. (iii) strategic goals, e.g. selection of the lowest-common-denominator approach to lower the costs and other barriers to coding.

Note that computing power here is a catalyst rather than a central factor; most of the activity is done by humans. FederationT in the sense discussed here contrasts with its use elsewhere to refer to linking networks or grids of computers in order to provide a scaling up of computational power. Here, we seek to scale up resource discovery, retrieval, and preservation, rather than processing.

More recently, parts of the linguistics community have been working in similar areas – OLAC, which was similarly centred on strategic goals for resource discovery, and GOLD ontology, which focussed on mapping out the concept territory of linguistics, to enable linguists to cross-map their varied terminologies (i.e. to bridge between author-created metadata and unified metadata formalised by a body of professionals). OLAC has been moderately successful, although more in terms of raising awareness about issues in language data handling than in unifying resource discovery across language data repositories, possibly because of its broad but ambiguous ambit ranging from endangered languages to multimedia to any language data. GOLD has been motivated by the putative needs of the "endangered languages community" (http://emeld.org/workshop/2003/paper-terry.html), but has mainly drawn interest from typologists and computationalists.

Ultimately, resource discovery has not, at least so far, been a foreground problem for most linguists. In other areas, web search engines have provided alternative solutions, and various areas of industry and commerce have been unobtrusively implementing EDI systems.

A conclusion one might warily draw is that the linguistic community has not (at least yet) found a clear need for such resource discovery and ultimately

federalism among repositories. On the other hand, however, linguists will benefit from previous and current work when the day comes that they do find such needs. Progress is likely to be sudden rather than evolutionary, when, at some point, linguists find that not only their tools (email, word processors, databases) but also their modes of expression are electronic (most likely this will occur among the forthcoming generation that will have been fully imbued with electronic communications of all kinds). Once enough linguists' decide to disseminate their own resources via electronic repositories, then federated electronic repositories will become a major locus for searching for other linguistic materials.

3. Opportunities

The current environment for language and technology and the nature of the DAM-LR partners suggest a number of opportunities that can guide strategy for collaboration. Our archives have relatively clear conception of our aims, holdings, and audiences, enabling us to exploit the valuable insights from specific linguistic (and related) subdomains, such as specialised corpuses, endangered languages, sign languages, the collection and implementation of protocol, new genres of data and presentation, new modes of access, and recognition of the new client groups for whom language data is crucially important.

Federating offers us important opportunities, because our repositories hold data that is typically fragmented, not published (or not conventionally publishable), and rare (in fact, it is the fragmented, data-oriented nature of our materials that unifies them as much as the fact that they are linguistic resources). Federation will provide increased dissemination opportunities and therefore add value to our individual collections.

In addition, we have a focal client group, depositors, to whom we need to offer substantial services in order to live up to our manifesto for "Live archives" (DAM-LR). While we do see depositors as a class of archive users, depositors have particular needs, for example to prepare and maintain their materials. The kind of interoperability typically provided by federation is based on use of a single SQL-like query to interrogate multiple repositories, which is centred on the information seeker rather than the information manager, which depositors are becoming. MPI's Lamus is a tool that is offering support in this direction. Another concern of depositors, to attain recognition of archive deposits as significant intellectual contribution on par with conventional publishing, can be greatly aided through successful federated dissemination of materials.

Finally, federation allows us to pool and share our strengths, for example, MPI's IMDI infrastructure and programming strengths, INL and Lund's corpuses, and SOAS' expertise in endangered languages.

4. Federating the domain

The goal of federationT is interoperability, the effectiveness of which is traditionally evaluated by the information retrieval measures precision and recall. Precision and recall are improved by using constrained metalanguages. The more lowest-common-denominator the approach to descriptive metadata (and therefore federationT), the less the specialities of participating agencies are reflected. For agencies that wish to serve users more thoroughly, metadata that drives resource discovery needs to be richer and domain-oriented. However, the mere sharing or overlapping of domains does not guarantee a shared semantics or vocabulary. Colomb (1997) shows that inter-database semantics or metadata mapping is a significant problem, even for simple domains. Agents within a federationT are faced with problems of semantic heterogeneity across their databases. Semantic heterogeneity can be a result of differences not solely between data categories, but between participant's understanding of their meanings, interpretations or usages (Sheth and Larson 1990, quoted in Colomb). It can be about differences in formal data models, system or project goals, or as a result of evolution of these over time.

Language archives face quite different data semantics from business and industry. Business data is anchored in well-defined concepts such as quantification, currency, and product codes; these are clearly-understood abstractions, widely agreed to represent key attributes and whose relation to the real world are not subject to interpretation. Libraries also enjoy conventionality of most of their descriptive attributes: well-understood concepts of author, title etc; in addition, these data are typically provided by authoritative publishers, and, as mentioned above, are available to individual libraries from centralised bibliographic sources.

In this sense, the language data world is a quite distinct one, with its descriptive categories, rather than being predetermined and centrally provided, needing to be derived bottom-up from our widely varied data and methodologies. A nomenclature of linguistics exists, but language data does not consist of measurements or key attributes, but speculative and contestable interpretations.[1] Thus, the apparent paradox that linguistics seems to guarantee non-interoperability arises due to the nature of language data (which is already metadata, i.e. we do not have agreed-upon data that will "ground out" the metadata semantics), and due to other factors such as that

[1] For example, a transcription might be changed as the linguist better understands a language's structures. Chomsky's aim was to lay foundations of a linguistic theory that would ground out this problem but it has not been overwhelmingly influential in our areas.

human languages are different from each other in arbitrarily complex ways and that individual linguists seek to emphasise or differentiate aspects of their data or analysis.

Repositories can federate with varying degrees of retention of their "design autonomy" (Colomb), i.e. different levels of change to their information systems to meet the needs of the federation. This is an important issue for DAM-LR. While all the partner agencies hold language data with common but specialised characteristics (e.g. sensitivity; identifying particular persons; emphasis on sound/video in binary formats), they are nevertheless quite specialised. Indeed for most it is a central mission to make a distinct contribution, manifested by creating new infrastructures (e.g. IMDI in the case of DoBeS); others (such as INL) have areal specialisation, or, like ELAR at SOAS, policy specialisation such as collection and implementation of protocol data. In addition, the nature of linguistic data itself is changing and diverging rapidly as the new paradigm of language documentation (a response to language endangerment) grows. For DAM-LR, some concepts are likely to be especially difficult to unify across partners, especially those related to granularity, such as the meanings and cross-mappings of bundle, collection, session etc., and categories of access rights.

5. FederationA: organisational and strategic aspects

The key to dealing with the issues in the preceding section is that the standardisation that enables federationT "is not primarily a computing process" (Colomb); it requires people-based structures, communication channels, and significant resources to maintain these and to enable these to be harnessed towards effective and ongoing development. It is the task of these federationsA to create and host an ongoing, evolving universe of negotiation, knowledge models, and transactions, not merely technical interoperability of terms.

Agencies aiming to form a federation need to be clear about a number of matters, from the semantic ones discussed above, to their purpose and scope, membership, and other strategic, organisational and legal questions. Purpose and scope could range from very broad[2] – to very narrow e.g. 17th century American visual culture (Ninch 2000). These in turn help to create informed and realistic user expectations; i.e. the federationA aims must provide both a forum for sharing and negotiation and a vehicle for disseminating. A co-ordinating body is needed to provide this forum, and to make decisions and strategy, especially in a period of rapidly advancing technology, and where the technology influences what services are expected and provided to users, and have significant financial implications for members.

Therefore, the core of federationA consists of a membership, and its goals. This is totally unlike perceptions of a federationT that consist only of technical standards broadcast from a central agency (the same lesson was learnt in the early development of Z39.50). One could go as far as requiring some form of membership even for users, who must ultimately (for a specialised domain) become part of the community of understanding of the federated metadata and its relationships.[3]

We do have special concerns. For example, conventional authentication systems (such as Shibboleth) exchange minimal data about users, and leave detailed gatekeeping up to individual repositories handling access. However, many linguistic resources have access conditions that associate resources with users, rather as if particular books in a library are not only borrowed under different terms by staff and students, but may be only borrowable by particular named individuals. In our specialist and changing area, federation is not only about searching multiple repositories but about identifying a range of user groups and their needs. This in turn will be enhanced by experience and feedback that a federal forum can incorporate into ongoing strategy.

6. Resourcing and legal aspects

Federation inevitably involves standards, which means formulating rules about implementing them, and, in turn, enforcement through either "incentives or penalties" (Colomb). The mechanism of membership needs to be clear, so that members are signatories to relevant statements of practice, with and formulations of what counts as compliance. Depending in the scope of its activities, a Federation might also be responsible for compliance (and reporting) with various legal requirements (such as data protection, privacy etc.) on behalf of members. These various requirements – heightened by the specific sensitivities and potencies of our holdings – mean that initial statements about trust, ethics etc need to be roundly discussed and formulated as a code to which members assent.

Some of our specialisations create limits to the extent that repositories can be federated. For example, one way of making two data sources comparable is to lose some specificity of the more constrained field – i.e. a "lossy" merge that nevertheless allows users to retrieve the relevant data under most queries. However, where a data attribute has legal or ethical implications (e.g. related to intellectual property, access restrictions, or privacy), then the option to manipulate the appearance, content, or granularity of such data is not open. In

[2] Which can raise problems, such as OLAC's adoption of DC-type scope while appealing to language endangerment for its motivation, thus diffusing its clarity of purpose.

[3] Although we should try to avoid the abuse that the term 'community' currently suffers, such as the "Windows user community" or the "open-source community".

this example, one can see that ultimately federationA is inseparable from federationT, because technologies must reliably implement the policies of members as legal entities with legal and ethical responsibilities and liabilities.

A federation will need a forum or body that can answer the questions that a legal mind will ask; questions such as: Who owns what? What are the risks and who is responsible for them? Where are the boundaries between agencies? How are differences across jurisdictions handled? Who is accountable? Who can communicate on behalf of the federation? For example, privacy legislation require that someone meet an individual's requests to examine data held about them, which would need to be handled initially at the same level as the "seamless interface" that federationT implies to the wider world. Ultimately, such legal and organisational aspects probably need be formally modelled and integrated into the implementation – again, we see the co-dependence of federationA and federationT.

The activities described in sections 4 and 5 above cannot take place without resources. However, in some cases, the resource base can be hidden or go unnoticed; for example, where participants are (a) performing tasks that are part of their core remit i.e. for which local resources can legitimately be expended; (b) public institutions such as libraries that are expected to develop public infrastructure; or (c) in a homogenous, stable, and well-integrated domain, so that benefits from investment could reasonably be assumed to accrue to all participants. Many of these conditions do not hold for the DAM-LR partners and their domains. Therefore, developments are dependent on obtaining sources of funding together with negotiations about the dedication of local members' resources to the federation's benefit. Again, this will place constraints on the processes for membership.

People resources are also needed: Ninch suggest that a federation may need access to a number of types of skills not only on the IT side (e.g. systems analysis, user interface, programmers) but also linguistic, archive, IP and legal experts, representatives of user groups.

7. Conclusion

DAM-LR is providing a useful testbed for the development of a federation of language resource archives, which could be extended to other nascent groups, such as DELAMAN. It already meets several of the considerations discussed above; in particular, we (i) have clear and constrained tasks and membership; (ii) there is a project and funding scenario within which our tasks are negotiated and resourced. On the other hand, it would be misleading to ignore the diverse and distinct organisational, strategic and implementation issues, and to conflate them all under the one term 'federation'. This paper has shown that a federation will weave together aspects of federationA and federationT.

The function of a federation, then, is to:

- supply services to particular communities (cf. OAI "designated communities")
- to supply those services from allocated resources, i.e. federations must *choose* the communities they will serve (for which there needs to be a forum for negotiation and evolution)
- supply services that take advantage of its members' resources, priorities and values
- to manage its membership and resources in support of the above

References

Colomb, R.M. 1997. "Impact of Semantic Heterogeneity on Federating Databases" *The Computer Journal* Vol 40, No. 5, pp. 235 -244.

DAM-LR (partners). 2006 *Live archives*. Pamphlet.

Dempsey, L., Russell, R., Heery, R. 1997. "Discovering Online Resources. In at the Shallow End: Metadata and Cross-domain Resource Discovery." http://ahds.ac.uk/public/metadata/disc_07.html

Lynch, Clifford 1998. "The Dublin Core Descriptive Metadata Program: Strategic Implications for Libraries and Networked Information Access." In ARL 1998 (196), Association of Research Libraries

NINCH 2000. "Federating Digital Image Repositories and Interpretive Information." *http://www.ninch.org/bb/proposals/visual2.html*

Integrated Services for the Language Resource Domain

Daan Broeder, Peter Wittenburg, Alex Klassmann, Freddie Offenga

Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen
{daan.broeder,alex.klassmann,freddy.offenga,peter.wittenburg}@mpi.nl

Abstract

Integrated services for the Language Resource domain will enable users to operate in a single unified domain of language resources. This type of integration introduces Grid technology to the humanities disciplines and allows the formation of a federation of archives. The DAM-LR project, will establish such a federation, integrating various European language resource archives. The complete architecture is designed based on a few well-known components and some integrated services are already tested and available.

1. Introduction

Creating integrated services and sharing resources between like minded archives for language resources as described by the "Live Archives" document [1] looks like an attractive proposition.

The aim is to benefit the user by creating an environment that allows access to all archives as one single virtual archive. It will benefit the participating archives as well by allowing them to better serve their users, allow pooling resources and development efforts and improving the basis of long term preservation.

The integration and sharing technologies used for such an effort are often referred to as "Grid" technologies [2], and in the world of hard science they are a popular subject for forming cooperative groups of institutes and archives called "federations". In the humanities especially so in the language resource domain such initiatives are rare. The work described here is largely developed within the DAM-LR [3] project that is one of the few that aims at establishing such a federation in the domain of language resources. While Grid technology solutions in the hard sciences were mainly driven by the typical compute bound tasks, leading to the development of middleware such as the Globus Toolkit [4], the humanities interests are more in-line with Data Grid solutions mainly inspired and coming from the Digital Library community.

In this paper we will not go into the organizational, legal and other non-technical aspects of forming such federation but leave it with mentioning that trust embodied in some kind of organizational form is required to make it all work.

2. Integrated Services for Language Archives

In many cases when we use the words "integrating" and "sharing" we actually are talking about interoperability. Users see a single domain of searchable metadata but the metadata format itself can be implemented differently for different archives. There is, however, a gateway that connects and translates to the agreed format so a single integrated "shared" domain is presented to the users.

Services that can be shared or integrated between language archives that present substantial advantages to the users are:

1) Sharing a single metadata domain for searching and browsing. This allows users to formulate one single query for "interesting" resources and obtain results of all cooperating archives. The required precision for such queries determined by the research questions also requires a domain specific metadata set. For more general queries more general metadata sets, shared by possibly other domains as well, can be used.

2) Sharing a scheme for persistent identifiers for resources. This is an issue when supporting references to resources stored in archives. It is well known that URLs are not the ideal means to do this. Different schemes for supporting persistent identifiers have been developed in the librarians' domain: Persistent URLs (PURL) [5] and the Handle System (HS) [6]. Sharing the persistent identifier scheme allows archives to easily reference each others resources and exchange resources with embedded references.

3) Secure authentication of archive identity. When sharing resources it is important to be able to establish the partners' identities. Without this, agreed access policies for instance, can not be guaranteed.

4) Single sign-on domain. Language Resource archives cater for the same user community. It would be very welcome if a single user identity can be established requiring a user to identify him only once when accessing resources from different archives.

5) Shared access policy or authorization. For reasons of efficiency it can be advantageous to copy resources between archives. It is important that the access policies of the originating archive for that resource are maintained. If also a single user identity domain is shared (see the previous point), this authorization information can be specific at the level of access by individual users.

The above enumeration of shared services does not imply that all of these should be actually shared between all the members of a federation. Indeed an opt-out for some difficult to maintain services can be desirable to also allow the participation of partners not able to maintain such a service. This requires an architectural framework where these shared services are as much independent as possible.

This independence is not to be confused with the possible organizational requirements where for instance it may be required to actually support a specific way of authentication, one that is trusted by the partner institutions. Or a service can be essential to the goals of a federation or project such as supporting a metadata infrastructure so the resources will be visible via a central portal.

The choice for a particular technology to implement the shared services is usually a matter of pragmatics. One of the partners can already have an installed base that can relatively easily be extended and used by other federation partners. However, it is always sensible to agree on the definitions of the exchange protocols rather than defining the implementation technologies. This allows individual archives the freedom in choosing the actual implementation while concentrating on the interoperability issue.

3. DAM-LR integrated services

In accordance with principles mentioned above, the DAM-LR project emphasized agreeing about the use of certain protocols for interoperability, leaving the partners free to choose a different implementation where possible. However the Max-Planck Institute for Psycholinguistics (MPI) agreed to further develop its archive management solution as a "reference implementation" demonstrating the integrated DAM-LR functionality. Some additional Grid components like the HS for persistent identifiers, were chosen especially because of an existing robust and dependable implementation and its already existing user base.

Prerequisite for all accepted solutions is that any integration component can only be accepted when it is distributed and redundant so that every archive can also function completely autonomous. In the following we will introduce the key pillars of the DAM-LR architecture that is also summarized in figure1.

3.1. Integrated Metadata Domain

With respect to metadata interoperability the following principles were agreed upon:

1) The IMDI metadata infrastructure [7],[8] will be supported for browsing and searching either by using the actual IMDI metadata format for storing metadata or by creating them on the fly from a local format or database. At least two portals will be made available with full functionality of metadata browsing and searching.

2) The Open Archives Initiative's (OAI) PMH [9] protocol is supported to allow harvesting metadata also in DC record format allowing interoperability to the outside world at the level of OAI service providers.

How the different partner archives make use of the integrated domain of IMDI metadata is a matter of choice, the "reference implementation" developed at the MPI and adopted by a number of the partners is described in 4.1.

3.2. Persistent Resource Identifiers

The DAM-LR archives will use persistent resource identifiers or URIDs (Unique Resource Identifiers) to enable stable references for their resources. The problems pertaining to the use of URLs are well known. Previous discussions have shown the advantage of using the Handle System over its contender PURL; the other widely used persistent identifier system. The Handle System of the CNRI [10] provides a highly available service for resolving URIDs to actual URLs. The HS is well known in the library community, adopting it will guarantee stable references from non-local resources (stand-off annotations) and also from publications.

The archive at MPI currently has a HS available for resolving references to its resources. The HS is integrated with other archive services in such a way that it is not an essential service but a highly desirable one.

The DAM-LR partners have agreed to host replications of each others handle service revolvers so this will be a distributed highly available service within the DAM-LR federation. Currently, the simplest scheme was chosen where one partner, possibly the MPI, has copies of all other Handle Systems.

3.3. Secure Archive Identification

All confidential communication between DAM-LR servers and services has to be secure. The interaction between peer components such as for instance those involved with user authentication are based on the existence of a domain of trusted servers and services and each component has to make sure that it is provably identified to be the one that it claims to be. As a means of implementing such a trusted domain, the TACAR list [11] of mutually agreed certificates was created, based on the principles of EUGridPMA [12]. In this implementation, national bodies declare that they will accept certificates form each other, with a Public Key Infrastructure [13] used to sign certificates. Every federation member has to apply to their national Certificate Authority to request the status of a Registration Authority, if the appropriate university is not already a Certification or Registration Authority. Once recognized as a Certification or Registration Authority, sites can issue or request certificates that will be accepted within the EUGridPMA domain.

3.4. Distributed User Authentication

Although all the cooperating archives aim at self sufficiency, several share a group of (potential) users that would like to access resources housed at different places without maintaining different user accounts. Therefore, it would be advantageous if the archives should accept each others identification and authentication of users. An accepted solution for this is the Shibboleth system [14].

The Shibboleth concept is primarily aimed at situations where users can be described by attributes such as "member of university class X". The authentication of the student is left to the student's home institution and the others grant access to individual resources based on the attributes associated with his identity. However, for individually operating researchers this scheme does not work as every individual needs still to be identifiable at each site when access rights are determined. In spite of this mismatch of required user specificity, Shibboleth

brings the advantage of user authentication being performed at the users home institution and transmitting in a secure way only limited and agreed user information over the internet.

Other possibilities have been considered such as the AAA toolkit [15] that emerged from the Grid community discussions as were also solutions based on a shared LDAP [16] domain. Shibboleth, however, looks to become the most widely accepted standard and might even become a requirement imposed by national libraries, government institutions or funding agencies.

Basically, the partners agree that user management should be done by the home site and that privacy sensitive information such as passwords will not be exchanged. Instead a user will be identified by a unique key that will be transmitted together with a limited number of user attributes between the partners. This key will be used in authorization records when associating resource access policies with users.

3.5. Access Authorization

The access authorization is different from user identification and authentication; it links resource access policies with user and/or group identifiers. If we consider the possibility that archives store copies of each others resources we have to make sure that the access policies remain the same irrelevant of the place where the copy of the resource is stored. Therefore, it seems a natural fit that the authorization records are coupled together with the resource's URID record in the HS. The HS allows to add such user defined record to every handle and thanks to the HS high availability, the authorization record will be available even when the "owner" archive is off-line in the same way as its URID will be.

An access manager component has to be developed or integrated that will match the Shibboleth provided identity with the policy stored in HS record, this can perhaps be achieved by extending Shibboleth's default access manager.

As already stated, the authorization records contain access policies mapped to Shibboleth provided and proven user identifiers and maybe some group access policies, however, Shibboleth does not provide archive managers with authorization records where none yet exists. If a user requests access to a resource this request has to be processed such that the unique federation wide user identifier is confirmed and suitable records can be produced if the archive manager approved the request. Such a resource request management system needs to be developed separately from Shibboleth.

4. Additional functions and Specific Implementation Issues

The following functions and applications are not part of any proscribed DAM-LR integrated service. However, they are essential for running a useful and consistent archive.

4.1. Metadata Utilization.

Within DAM-LR different portals will be established that allow utilization of the integrated metadata domain so users can find relevant resources searching all the partner archives simultaneously. The DAM-LR partners are free to develop their own solution for this, but the majority has chosen to adopt the IMDI infrastructure that allows the

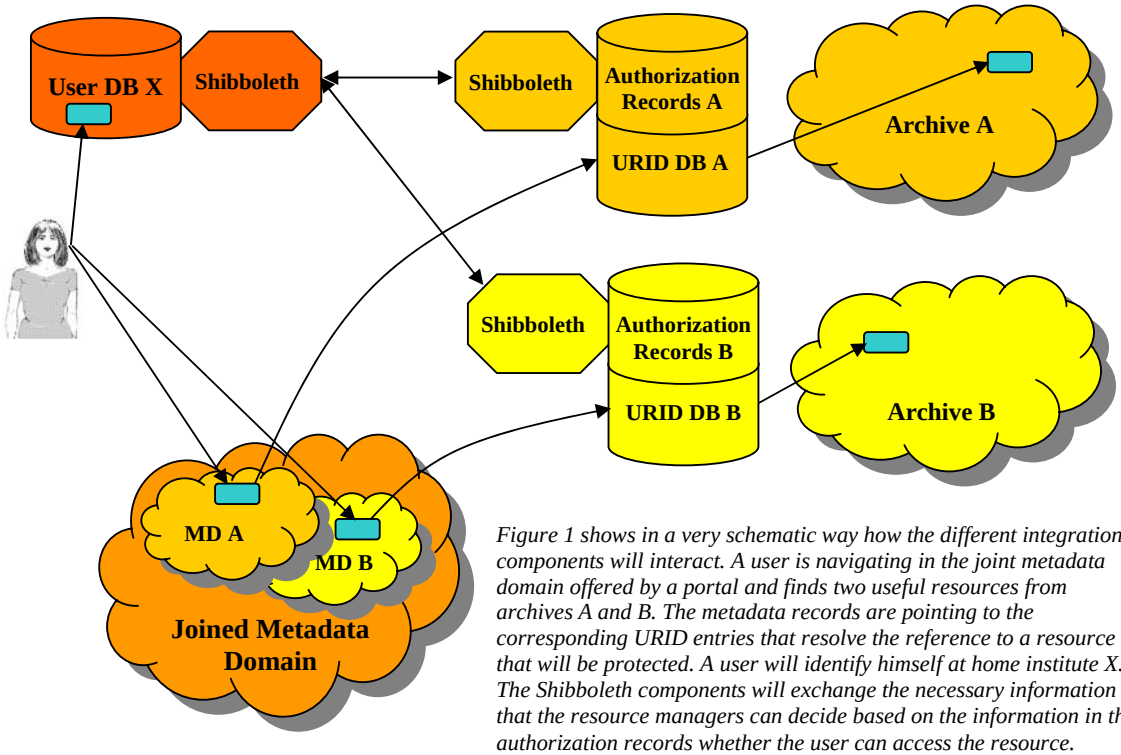


Figure 1 shows in a very schematic way how the different integration components will interact. A user is navigating in the joint metadata domain offered by a portal and finds two useful resources from archives A and B. The metadata records are pointing to the corresponding URID entries that resolve the reference to a resource that will be protected. A user will identify himself at home institute X. The Shibboleth components will exchange the necessary information so that the resource managers can decide based on the information in the authorization records whether the user can access the resource.

following functionality:

(1) Browsing. This is similar to clicking through a local file system where the directories are replaced by linguistically relevant groupings (sub-corpora). The approach is aimed at users familiar with or quickly able to grasp the underlying logical organization. A component allowing geographic browsing is also available.

(2) Structured search over the whole domain as well as within selected parts of it. With this type of search every metadata element can be addressed individually and the search for different elements can be combined into one query. Queries can be formulated with high precision required by research interests. Yet, the user has to know the terminology used by the metadata set in order to achieve a high recall. Furthermore, structured search is restricted to elements with closed or open vocabularies and does not cover elements with free text.

(3) Unstructured search over the whole domain or selected parts of it. Users can enter words or regular expressions into a free text field (Google-like). Any metadata element including the free text descriptions that contains matching strings will produce a hit. The recall with this method can be expected to be higher compared with structured search however, the precision will be poor.

4.2. Versioning of Resources.

The "stable identifier" issue addressed in 3.2 makes no sense if the resource itself is modified. Therefore, the original resource should never be deleted from an archive and always be accessible (although it need not be immediately). Also when we have a reference to a resource, we would like to be able to have access to older and newer versions if they exist. So when new resources are put into the archive and the depositor specifies they are to replace existing ones, the old resources are to be suitably marked and moved to the archive's "attic".

Discussions on the visibility in views on the archive of the old versions are complicated, but for the moment we have decided on allowing only access to older versions on the basis of a direct reference to it or via a reference to another version of it. This divides the "viewable" archive in two dimensions: (1) the set of all latest versions of all objects in the archive and (2) on the basis of a selected archive object we have access to its older versions.

4.3. Access Management System

Needed is also an efficient way to generate the authorization records for resources of whole corpora at once. Such a system should also allow archive management to delegate this task of setting access permissions to the depositor of the resource or somebody else responsible for the corpus.

At the MPI such a system is currently available although not yet integrated with Shibboleth and HS. This access management system is not DAM-LR prescribed and every partner archive can choose to implement its own version.

5. Conclusions

The DAM-LR project is an excellent test-bed for integration and sharing technologies for the Language Resource domain and even beyond for the humanities. Also the project partners are convinced that archive federations are an essential step on the way to realize an eScience scenario for linguistics and the humanities as is indicated in figure 2. Federations will be an utterly important part of a research infrastructure that will lend services not only to linguists in the broad sense, but also to other disciplines in the humanities. They will also link up to archives that house for example ethnological, historical resources and many others. Due to the virtual integration aspect of archives it is obvious that federations will bring an added value to the researcher.

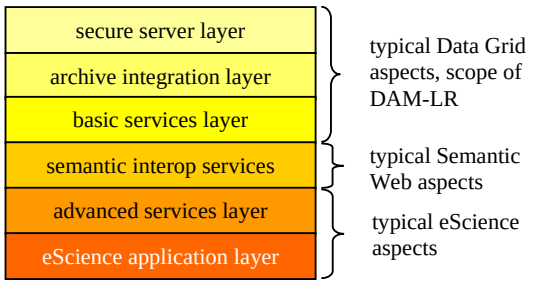


Figure 2 indicates the typical layer hierarchy where Grid solutions take care of typical integration aspects, Semantic Web solutions address the problems associated with interoperability in particular at the semantic level and eScience solutions provide advanced applications such as semantic weaving and web-based collaboration on top of the other layers.

Since DAM-LR is – as far as we know – the first project in the humanities that applies Grid-type of technology on a supra-national scale, it will have a great impact on establishing stable research infrastructures in the humanities. Therefore, we feel that it is important that all DAM-LR documents be made openly available and a training program be created to actively inform other interested parties. Also DAM-LR was purposefully setup as a small project with initially a few partners, but, given the architectural simplicity of the solution found, it is our intention to scale DAM-LR up to possibly up to 20 European partners if enough interested resource archives can be found that can offer well organized documented resources.

6. References

[1] live archives, http://www.mpi.nl/dam-lr/live-archives
[2]GRID forum, http://www.gridforum.org
[3] DAM-LR project, http://www.mpi.nl/DAM-LR/
[3] GTK, http://www.globus. org/
[4] PURL, http://www.purl.org
[5[HS, http://www.handle.net
[6] http://www.mpi.nl/IMDI
[7] Wittenburg, P., Peters, W., Broeder, D. (2002). Metadata Proposals for Corpora and Lexica. In M. Roriguez Ganzalez & C. Paz Suarez Araujo (eds.), Proceedings of the 3rd International Conference on

Language Resources and Evaluation. Paris: European Language Resource Association. pp 1321-1326
[8] OAI/PMH http://www.openarchives.org/OAI/openarchivesprotocol.html
[9] CNRI. http://www.cnri.net
[10] TACAR. http://www.tacar.org/
[11] EUGRID, http://www.eugridpma.org/
[12] PKI, http://www.pki-page.org
[13] http://shibboleth.internet2.edu/
[14] http://www.science.uva.nl/research/air/projects/aaa
[15] http://www.openldap.org

Common Infrastructure for Finite-State Based Methods and Linguistic Descriptions

Anssi YLI-JYRÄ[1,2] and **Kimmo KOSKENNIEMI**[2] and **Krister LINDÉN**[2]
[1] CSC - Scientific Computing Ltd., P.O. Box 405, FI-02101 Espoo, Finland
[2] Department of General Linguistics, P.O. Box 9, FI-00014 University of Helsinki, Finland
{aylijyra,koskenni,klinden}@ling.helsinki.fi

Abstract

Finite-state methods have been adopted widely in computational morphology and related linguistic applications. To enable efficient development of finite-state based linguistic descriptions, these methods should be a freely available resource for academic language research and the language technology industry. The following needs can be identified: (i) a **registry** that maps the existing approaches, implementations and descriptions, (ii) managing the **incompatibilities** of the existing tools, (iii) increasing **synergy** and complementary functionality of the tools, (iv) persistent **availability** of the tools used to manipulate the archived descriptions, (v) an **archive** for free finite-state based tools and linguistic descriptions. Addressing these challenges contributes to building a common research infrastructure for advanced language technology

1. Introduction

Finite-state methods have been adopted widely in computational morphology and related tasks of natural language and speech processing, including segmentation, tokenisation, shallow parsing, name entity recognition, normalization etc. To enable efficient development of finite-state based linguistic descriptions, the underlying methods and the lexicons should be a freely available, **common and growing** resource for academic language research and the language technology industry. The idea of a common finite-state based methodology is not new, but it has not been easy to implement in large scale.

The purpose of this article is to identify some needs that are faced when we try to reach this goal, and to propose some helpful approaches to their satisfaction. These needs are discussed in Sections 2 – 6.

2. Specialized Software Registry for Finite-State Based Resources

We need to **register** finite-state tools and linguistic resources. An open registry, **FSMREG**, currently located at http://www.ling.helsinki.fi/users/aylijyra/FSMREG will be pre-populated with the entries in our local database. After the necessary extensions, this registry

- will be a locator service for commercial and non-commercial finite-state based resources
- will map file formats and algorithms that are in use in the existing resources
- will contain hypertext links to a distributed collection of examples and stub grammars that can be used as starting points for benchmarking, testing and teaching.

According to our investigations, there are at least 70 languages to which some finite-state based methods have already been applied. Moreover, we have constructed partial registry entries about a few dozen finite-state based tools (including *ALE-RA, Amore, ASTL, BELLEx3, Carmel, DFKI FSM, FIRE toolkits, FAdo, RWTH FSA, FSA (Gdansk), FSA (Groningen), fskit, fsmlibrary, GFSMNT, grmlibrary, ifsc, Intex, KIMMO, lexc, lextools, MAP (Alvey), MIT FST, MMORPH, OMAC FSM, PC-KIMMO, SFST, twolc, UCFSM, Unitex, Vaucanson, wfsc, wfst, X2MORF, xfst*).

We welcome contributions of new or corrected entries in the registry. In the future, we plan to move the registry to a collaboration environment using the wiki technology, and to present a version of the registry as a survey article or technical report.

3. Common Formats and Formalisms for Finite-State Resources

We need to manage the divergence of the existing finite-state tools. Different finite-state tools should be capable of **exchanging** various types of data: finite-state objects as well as grammar source files created in finite-state based formalisms. Currently, many finite-state based formalisms can be parsed only with a proprietary compiler. To create interoperable tools and industry standards, we need

- an open forum for reviewing idiosyncratic features of finite-state based rule formalisms
- a generic XML-based exchange format for finite-state based rule formalisms
- converters that rewrite formalisms into system specific regular expressions (For example, *xfst2fsa* (Cohen-Sygal and Wintner 2005) converts a large subset of the Xerox finite-state formalism in *xfs*, to expressions of the *FSA* utilities from Rijksuniversiteit Groningen.)
- XML-formats (such as proposed by the Vaucanson group http://www.lrde.epita.fr/cgi-bin/twiki/view/Vaucanson/XML) for exchanging small finite-state objects
- open libraries that can exchange huge finite-state objects in various binary formats

4. Complementary Modules of Finite-State Methods

We need to increase **synergy** in building new finite-state tools. Earlier, proprietary and private implementations of finite-state methods have been in-house tools for building certain natural language and speech processing applications. As a result, similar finite-state toolkits have been reimplemented several times in different places. Now that a few proprietary finite-state toolkits are available under commercial licenses, there is a great need for **complementary tools** that would help in tasks where flexibility is more important than high performance.

- We need open source tools that can be mutated and exploited more freely
- We need compilers that can be linked with different finite-state libraries:
    a. a pre-compiler for compiling linguistic descriptions into regular expressions
    b. regular expressions would be compiled by a separate program into finite-state objects

It is surprising how little the flexibility and modularity of widely available finite-state compilers has developed during the course of last 20 years. Earlier, when finite-state tools were written in the Lisp programming language, it was convenient to implement rule compilers and pre-compilers (see *e.g.* Karttunen *et al.* 1987) also in Lisp. Today, some pre-compilers for regular expressions have been implemented with XML-based techniques (Piskorski *et al.* 2002). The software package *fskit* developed by the first author employs a further pre-compiler and macro expansion method.

5. Encouraging Open Source Development of Finite-State Resources

We need an **action plan** that increases the free availability of useful finite-state based methods and descriptions. Currently, some tools for creating linguistic resources are available under incompatible or closed-term license models. The action plan would

- encourage compatibility with such research networks that build free finite-state based descriptions (including the RELEX network and OpenOffice-related projects)
- encourage the use of open source or creative commons licenses that allow linking to software covered by GNU's copyleft license as well as to proprietary software
- recognize the need for a manageable negotiation procedure in the exceptional cases where the terms of the default license is not compatible with a desirable combination
- discuss the possible need for joint copyright systems

There is a trade-off between the commercial relevance for widely spoken languages and the common good for communities of less-studied languages and the research community. This opposition has wide practical implications that make it especially complicated to build a common, standardized infrastructure for finite-state based methods and applications.

For example, the free availability of some finite-state based formalisms is perhaps not even possible due to potential patent risks. In other words, patents and proprietary programming languages are problematic from the viewpoint of persistent archiving and sustainability. They may involve risks if the value of the infrastructure of language resources is dependent on the availability of the software needed to maintain the resources.

6. Archiving

All the finite-state resources need to be archived and stored somewhere. We believe that storage is not a problem for open-source resources, but the main problem is to keep the resources maintainable and exploitable. This involves, in addition to the maintained finite-state compilers for the resources, sufficient documentation on the metadata and the used codes for each stored linguistic finite-state resource.

7. Conclusion

The better interoperability of high-end proprietary tools and freely available, sustainable tools is crucial requirement for multi-lingual language technology industry that would support diversity and development of language technology for minority languages (Yli-Jyrä 2005, Koskenniemi 2006). Open source language technology resources such as finite-state based methods and finite-state based linguistic descriptions

- create a basis for further experimental research on finite-state methods
- increase the availability of basic utilities needed in many small language technology projects
- support the development of complex applications on top of basic methods
- increase the efficiency and flexibility of commercial and academic research and development.

If the repeated investments in basic finite-state based resources could be avoided, new development efforts could concentrate on less-studied languages, research collaboration, more complex applications and the production of end-user products.

8. References

Beesley, K. R. (2004). Morphological analysis and generation: a first-step in natural language processing. In *Proceedings of the SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages. Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*, Lisbon, Portugal, May.

Cohen-Sygal, Y and S. Wintner. (2005). XFST2FSA: comparing two finite-state toolboxes. In *Proceedings of the ACL-2005 Workshop on Software*, Ann Arbor, MI, June 2005.

Karttunen, L., K. Koskenniemi and R. M. Kaplan. (1987). A compiler for phonological rules. In M. Dalrymple, et al. (eds). *Tools for Morphological Analysis. Center for the Study of Language and Information*. Stanford University, Palo Alto.

Koskenniemi, K. (2005). White paper: "Multilingual Europe – How to get there?" Presented in the Workshop "Machine Translation and Human Language Technologies", European Commission, Luxembourg, February.

Piskorski, J., W. Drożdżyński, O. Scherf and F. Xu. (2002). A flexible XML-based regular compiler for creation and conversion of linguistic resources. In *Proceedings of the 3rd International Conference on Language Resources an Evaluation (LREC'02)*, Las Palmas, Canary Islands, Spain.

Yli-Jyrä, A. (2005). Toward a widely usable finite-state morphology workbench for less studied languages – part I: desiderata. *Nordic Journal of African Studies*, 14(4): 479 – 491.

Knowledge Center for Processing Hebrew

Alon Itai
Head of the Knowledge Center for Processing Hebrew
Department of Computer Science
Technion, Israel Institute of Technology
32000 Haifa, Israel
Email: itai@cs.technion.ac.il

Introduction

To preserve cultural diversity it is necessary to preserve underrepresented languages. Such languages suffer from the dominance of English. This is aggravated by the Internet and the personal computer whose tools are tailored to English. Many tools cannot handle native alphabets, and these languages lack specific tools, such as spellers. Thus, for example, email and chats are often conducted in (pidgin) English. Moreover, even users of languages which use the Latin alphabet need language specific tool such as spellers, word processors and web searchers; these in turn often require linguistic tools such as morphological parsers, syntactic parsers, lexicons and bilingual dictionaries. Since the number of native speakers is relatively small there is little economic incentive for commercial companies to develop the tools. Thus there is a real danger that the more sophisticated users will abandon their native tongue in their professional work. It is the role of national and regional governments to carry the burden of preserving the native languages and to that end develop computer tools: software and databases.

It is also important to conduct linguistic research in such languages and order to do so one needs publicly available linguistic tools with open access source code software programs. For example, it is argued that one cannot meaningfully search documents in a language with complex morphology without using a morphological analyzer. If there is no publicly available morphological analyzer then every researcher has to reconstruct such a tool. Moreover, in order to build a high quality morphological analyzer one needs a high quality lexicon. Thus every researcher conducting corpus linguistics has to invest in a morphological analyzer and lexicon before starting her research.

This paper describes the Knowledge Center for Processing Hebrew whose aim is to make computer tools and databases for Hebrew available to the public and thus enhance the linguistic research of Modern Hebrew in both computational and theoretical linguistics, and to promote the commercial usage of NLP systems for Hebrew.

The Knowledge Center Model

In 2003, the Israeli Ministry of Science and Technology established a Knowledge Center for Processing Hebrew. Its aim was to develop products (software and databases) for processing Hebrew and make them available to the public, both in academia and industry. Researchers from four universities are involved with the Center's activities.

Before the establishment of the Center, the lack of standardization and centralization caused much duplication of effort. For example, several morphological analyzers of Hebrew were developed by different teams, using different methodologies and different output formats, and based on different lexicons (Choueka and Shapira 1964, Ornan 1987, Lavie et al. 1988, Bentur et al. 1992, Segal 1999, Yona and Wintner 2005). Since their output was different, and the source code was not available, it was impossible to compare them or reuse their resources. Furthermore, many of the developed tools were unavailable to the entire community.

Much of the Center's efforts are dedicated to transform software developed in academia for research purposes into tools available to the public. In research one wants to prove a concept, not to provide a commercial tool. Providing documentation, user interface and making the programs platform independent require a lot of work with little academic reward. Thus most often tools created in academia cannot be reused. The Center upgrades software and other tools created by academia and private researchers to make them reusable. The Center also provides a depository, thus researchers know from where to download tools.

The ministry's aim was to make the center self sustainable, i.e., the revenues from selling products and services should provide funds to maintain the Center. However, since the market is small, such revenues proved to be insufficient. Furthermore, had there been a commercial market there would have been no need to establish the Center.

Since our aim was to make the products available to the entire community in order to encourage research we have made all our products available under the GPL (Gnu public license http://www.gnu.org/copyleft/gpl.html), including the source code of software. This license allows free use but requires that all products that embed GPL products also be under GPL thus limiting the commercial use of our products. In order to enable commercial development, we allow the commercial use of products that do not contain embedded GPL components. This use is non-exclusive, i.e., the same products are also available for free use under the GPL.

Modern Hebrew

Modern Hebrew is one of the two official languages of the State of Israel, spoken natively by half of the population and fluently by virtually all the (seven million) residents of the country. The language is strongly related to (though linguistically distinct from) biblical Hebrew, and thus has raised the interests of both linguists and religious scholars.

Modern Hebrew exhibits clear Semitic behavior. In particular, its lexicon, word formation and inflectional morphology are typically Semitic. Its morphology is inflectional and highly productive and consists mostly of suffixes, but sometimes of prefixes or circumfixes. Often connectives and prepositions are prepended to words.

In the standard Hebrew script, like Arabic, most of the vowels are not represented. Thus Hebrew texts are highly ambiguous. 55% of the tokens are ambiguous; some tokens have up to 13 analyses, while the average number of analyses is over 2.

Thus a major difficulty in processing Hebrew is to morphologically disambiguate the text, i.e., choose the right analyses according to the context.

Products

The development of the products was motivated by the following principles:

Portability: The format should be platform independent

Readability: The representation should allow for easy production of annotations, easy parsing and processing of the annotated data, by both machines and humans;

Standardization: Processing of the annotated data should be supported by a wide variety of environments (information processing tools, programming languages, etc.);

Reversibility: The original data should be easily extracted from the annotated version if desired;

Openness: The tools used to produce the resources and the production steps of the annotated data should be publicly available, to allow the recreation of the data or further development;

Suboptimal Efficiency: The resources and tools are not meant to compete with industrial products but instead to be easy to understand, easy to use and easy to expand. Thus, the resources and tools we provide are not always optimized for space and time.

Our linguistic databases are represented in Extensible Markup Language – XML (Connolly 1997) according to schemas that enforce structure and are also used for documentation and validation purposes. The output of the morphological analyzers and taggers is also in XML format. Thus we can use the software modularly and compare the outputs of different implementations.

The products include

1. XML standards for representing lexicons and corpora.
2. Segmentizers: Tokenizer, sentencizer (a program that partitions the corpus to sentences), word-segmentizer (a program that partitions the word into morphemes).
3. Morphological analyzers and taggers: The analyzers list all the possible analyses, whereas the taggers attempt to find the correct analysis in context.
4. Part of speech taggers (Bar Haim, Sima'an and Winter 2005).
5. Corpora: 20 million word corpora of printed press, 17 million words of Parliamentary proceedings, 1.3 million word corpora of printed press with partial niqqud (diacritical marks for vowels). All these corpora appear in XML format, and include morphological analysis and automatic tagging. A 6000 sentence morphologically manually tagged corpus is also available.
6. Graphical User Interfaces (GUI) for tagging and preparing lexicons.
7. Tree bank: 6000 syntactically parsed sentences (Sima'an 2001).
8. Lexicon: A full lexicon of Modern Hebrew containing over 21,000 entries (Itai, Wintner and Yona 2006).
9. Tools for processing phonemic script (Ornan 1987).
10. Speech analysis databases.

A full list of products is available at
http://mila.cs.technion.ac.il/website/english/resources/index.html

Conclusions

The Knowledge Center for Processing Hebrew was created for the sole purpose of developing a research infrastructure for language resources. It is a good example of a government-funded entity that functions as a language resource center and focuses on defining and enforcing standards, as well as developing and archiving linguistic databases (such as corpora and lexicons) and tools (such as morphological analyzers). It facilitates easy access to and sharing of resources through an open-source policy. The products developed at the Center have so far proved useful both for commercial applications and for linguistic, psycholinguistic and literary research.

The Knowledge Center provides a model that can be applied to other languages. Some of our products are language specific, whereas others can be adapted to other languages. However, this is not a solution for languages with very few speakers, since the cost of establishing a center is large and it is essential to have skilled professionals — linguists and programmers.

Acknowledgement

The Center for Processing Hebrew was funded by the Israeli Ministry of Science and Technology. We wish to thank the Center's members: Yoad Winter (Technion), Shuly Wintner (Haifa University), Michael Elhadad and the late Arnon Cohen (Ben Gurion University), Yoram Singer and Eli Shamir (Hebrew University) and to the numerous graduate students who were involved in the development of the resources. We also wish to thank the technical staff: Dalia Bojan, Adi Cohen-Milea, Danny Shacham, Shira Schwartz, and Shlomo Yona.

References

Bar-Haim, Roy, Khalil Sima'an and Yoad Winter 2005: Choosing an Optimal Architecture for Segmentation and POS-Tagging of Modern Hebrew. *ACL 2005 Workshop on Computational Approaches to Semitic Languages*.

Bentur, Esther, Aviella Angel, and Danit Segev, 1992. Computerized Analysis of Hebrew Words. *Hebrew Linguistics* 36, 33-38. (In Hebrew.)

Choueka Yaacov 1998, (Chief Editor, with the Rav-Milim team), *Rav-Milim - the complete dictionary of contemporary Hebrew*, Steimatzki, C.E.T. and Miskal, Tel-Aviv.

Choueka, Yaacov and M. Shapiro 1964, *Machine analysis of Hebrew morphology: potentialities and achievements*, Leshonenu (Journal of the Academy of the Hebrew Language), Vol. 27, 1964, 354 -372 (Hebrew).

Connolly, Dan 1997. XML: Principles, Tools, and Techniques. O'Reilly.

Itai, Alon, Shuly Wintner and Shlomo Yona 2006. A Computational Lexicon of Contemporary Hebrew. In *Proceedings of LREC-2006*, Genoa, Italy, May 2006.

Lavie, Alon, Alon Itai, Uzzi Ornan, and Mori Rimon. 1988. On the applicability of two-level morphology to the inflection of Hebrew verbs. ALLC June 1988, Jerusalem, (TR 513, Department of Computer Science, Technion, 32000 Haifa, Israel).

Ornan, Uzzi 1987. Computer processing of Hebrew texts based on an unambiguous script. *Mishpatim* 17(2) 15-24. (In Hebrew.)

Segal, Erel 1999 Hebrew Morphological Analyzer for Hebrew undotted texts *M.Sc. Theses.* Dept. of Computer Science,Technion, Israel Institute of Technology Haifa, Israel. October 1999.

Sima'an, Khalil, Alon Itai, Yoad Winter, Alon Altman and Noa Nativ 2001. Building a Tree-Bank of Modern Hebrew Text. *Traitment Automatique des Langues*, 42, 347-380.

Yona, Shlomo and Shuly Wintner 2005. A finite-state morphological grammar of Hebrew. In *Proceedings of the ACL-2005 Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, MI, June 2005.

Design Features for the Collection and Distribution of Basic NLP-Resources for the World's Writing Systems

Oliver Streiter 1*, Mathias Stuflesser 2†

*National University of Kaohsiung, Taiwan
ostreiter@nuk.edu.tw
† European Academy of Bozen Bolzano, Italy
mstuflesser@eurac.edu

Abstract

Most of the world's 7000 languages are still lacking freely available language resources. This lack of resources forms a major bottleneck in the processing of those languages and prevents them form being more widely used. To overcome current limitations, researchers might profit from studying the cooperation in free software projects or Wiki-projects. In this paper, we explore such models of 'organic' cooperation for the collection and elaboration of free NLP-resources. We describe the database XNLRDF which has been set up for this purpose. Storing NLP-data for hundreds of languages, we gradually refined and extended the idea of what kind of information has to be included in such a database, how the information is to be stored and how such data might be created in an organic cooperation. A principled distinction we make is that between the data structure used for development and the data structure used for distribution, a relational database and XML-RDF respectively. Taking the advantages of XML for granted, we explain the advantages of a relational database for the development and maintenance of collaboratively developed data. Within the data structure, the notion of 'writing system' functions as pivot. A writing system incorporates a set of metadata such as language, locality, script, orthography, writing standard and assigns them to the NLP resources provided in XNLRDF. An overview over the first data we collected and an outlook on future developments will conclude this paper.

1. Old and New Research Traditions

From the 7000 languages of the world, about 1000 are estimated to have shown up on the Internet[1]. This high number reflects the pride of people in their culture and their willingness to use their language in electronic medias for communication and learning. It also represents the economical and ideological interest in most languages as a means to contact, inform or persuade people. However, many languages are not supported in their digitalized form. Computer users might be able to input the characters of a writing system, sometimes with difficulties (Uchechukwu, 2005), but overall there are no spell-checkers, grammar checkers, information retrieval systems or translation dictionaries.

This deplorable situation is not an exception for one remote language spoken far away, but is reality for more than 99% of the world's languages, a fact not taken into consideration by 99% of the NLP community. This distortion, one might call it even a caricature, is not due, as one might assume, to a lack of money, a lack or scientific interest, a lack of commercial interest or a lack of linguistic knowledge. In fact, many languages have been scientifically described with great care. In addition, each speaker of a language is a potential client for a soft drink, a political movement or a religious community and could thus best be addressed in his or her native tongue. Instead, the misery is rooted in our research tradition.

This research tradition will change however under the influence of free software projects, blogs, Wikis and creative commons licenses (Streiter, 2005; Streiter et al., 2006): Academical hierarchies, the distinction between affiliated scientist, enthusiast and partisans, the attribution of a scientific work to a researcher or a research body, the search for research topics in predefined academic fields and modular models of cooperation in research projects will become less pervasive and thus might pave the way for new models of scientific cooperation.

To explore the potentials of this new modes of research and to bridge the gaps between a) the needs of languages users, b) affiliated research and c) the potential contributions of non-affiliated researchers, we started to create an environment for an organic cooperation through the Internet with the aim of collecting and elaborating NLP-resources for the world's 8000 languages. The created NLP-resources are available in hourly builds under the GNU public license[2] and intellectual insights related to the development of the resources are available under the Creative Commons License[3]. Currently, the discussion of data structures and the collection of the first data is still done by a small circle of volunteers. But we hope that the circle of interested people might gradually enlarge, to open up finally for a free Wikipedia-like cooperation.

2. More is Different

The project XNLRDF (Natural Language Resource Description Framework) thus develops, breadth-first, an NLP-infrastructure for the world's writing systems, and, not tackled for the moment, the world's speech systems. Exploring the world's writing systems in all their differences and particularities, we hope to define a stable framework which can accommodate the most unusual cases without having to redefine the basic model or to compromise the data. While registering 23.000 writing systems, 8200 languages, 150 scripts and textual examples of 700 writing

[1] http://www.guardian.co.uk/GWeekly/Story/0,3939,427939,00.html

[2] http://140.127.211.214/research/nlrdf_download.html

[3] http://xnlrdf.wikispaces.com

systems, we have been forced to rethink and adapt our notions of (i) linguistic metadata, (ii) the nature of basic NLP-data and (iii) the way data are created and managed.

2.1. Different Metadata

In NLP, metadata identify the most suitable resources for the processing of text documents. In document formats like HTML, the language of a document is considered the most important kind of metadata. However, as the same language might be written at different times or in different places, before or after a spelling reform, before or after the adoption of a new alphabet or a new script, NLP-resources and text documents shouldn't use *language* as metadata, but a more specific notion, the ***writing system*** of a language.

To distinguish the estimated 100.000 writing systems of the world and to assign the most suitable resources, metadata have to be much more specific than what is currently used in the HTML header and even more specific than what has been suggested in the framework of OLAC (Simons and Bird (eds.), 2003). In addition, in case the processing of a document needs a resource of a given type, but no such resource has been explicitly assigned to the writing system of the document, inheritance principles are required that allow to assign to the document the most suitable resource from a related writing system. We thus currently define a *writing system* by the n-tuple of the more elementary metadata *language*, *locality*, *script*, *orthography*, the *writing standard*, the *time period* and a *reference* to another writing system. Each of these elementary metadata is identified by an arbitrary ID and ISO-codes, if available. Natural language names for these metadata are provided for convenience or in case standard codes are not defined[4] or ambiguous[5]. The choice of the natural language designators is not crucial as long as they are not pejorative or ambiguous. Designators in XNLRDF are provided in many languages (more precisely writing systems). One of the designators in the writing system *'late modern english_united states of america@latin'* is selected as being the default for generation, e.g. when generating a pick-list of language names.

Supporting evidence for the necessity of these elementary metadata comes from cases like Abkhaz: Abkhaz (*language*) has not only been written with two different Cyrillic alphabets (*script*), but also with two different Latin alphabets (*script*), one between 1926 and 1928 (*time period*) and one between 1928 and 1937. One might want to distinguish these writing systems by their name (the *standard*) or by the *time period*. In such cases, we do not exclude the first solution, although there is frequently no name for the standard. If possible, we prefer the time period, as it offers the possibility to calculate intersections with other time constraints, e.g. on the production date of the document, or the foundation or disintegration of a country or region.

The *writing standard* is best explained with the help of the different, concurring, isochronic writing standards for Norwegian (*language*): Nynorsk, Bokmål, Riksmål and Høgnorsk are different conventions (*writing standards*) to represent basically the same language[6].

[4]e.g. the Ladin variety of Gherdëina Valley

[5]e.g iso-639-1 codes of groups of languages

[6]cf. http://en.wikipedia.org/wiki/Norwegian

The *orthography* is best illustrated by the spelling reform of German with the new orthography coming into force in different *localities* (*'Germany', 'Austria', 'Liechtenstein'* ...) at different times and overlapping with the old spelling for a different number of years. In this case, disposing of the *time period* is a nice feature but it does not allow to dispense with the category of *orthography*. Unfortunately, orthographies, also frequently lack a standard name and are referred to as '*new*' in opposition to '*old*'.

The *reference* is a necessary metadatum to represent transliteration systems, i.e. transliterations in the strong meaning as one-to-one mapping, but also as one-to-many or many-to-many mappings. *'Braille'* is such a transliteration system which changes with the spelling reforms and standards of the referenced writing system. Thus, there exists one Norwegian Braille derived from Nynorsk and a second Norwegian Braille derived from Bokmål. By the same principle, Braille of the new German orthography is different from Braille based on the old German orthography.

Braille changes also when the *locality* of Braille is different from the *locality* of the referenced writing system. For example, Spanish Braille in a Spanish speaking country is different from the Spanish of a Spanish speaking country represented as Braille in the USA. This complexity can be handled when we allow writing systems to refer to each other. Thus Braille, as other transliteration systems, is represented as writing system with its own independent *locality, script* and *standard*, (e.g. '*contracted*' and '*non-contracted*'). The *language* and *time period* of the transliteration and the referred writing system are however the same.

A transliteration is thus marked by a reference to another writing system and mapping tables between the two systems, e.g. between Bokmål and Bokmål Braille. Mappings between writing systems are a natural component in the description of all writing systems, even if they do not represent transliterations of each other, e.g. mappings between *'hanyu pinyin', 'wade-giles'* and *'zhuyin fuhao'*.

Writing systems and, in the future, speech systems are identified by an arbitrary ID. They can be recognized by programmers through the concatenation of default natural language designators of the elementary meta data. Unspecified data are omitted, e.g. *'Uighur@Cyrillic', 'Uighur_Uzbekistan@Latin', 'Norwegian_Norway@Latin#nynorsk'*. NLP-resources are then described with respect to their *function*, their *encoding*, their *copyright*, their *URL*. NLP-resources accumulated within XNLRDF are encoded in UTF-8, distributed under the GNU Public license and associated with one or more writing system. The writing system is thus the pivot notion, which connects metadata and NLP resources.

2.2. Different NLP Data

Given the huge number of writing systems created by mankind, no property of a writing system is universally valid. E.g. the function of a white space, dash or dot varies between scripts, but also between languages of the same script. Writing systems differ also by the characters representing word/syllable boundaries, ciphers and number words, the writing direction (e.g top to bottom left to right for Mongolian) and the sorting of characters in a wordlist.

Unicode which is generally assumed to cover this information fails to provide this NLP relevant information. Unicode refers only to scripts and ignores the notions of language or writing system. Unicode thus assigns properties at a level of the *script*, where these properties can only be understood at best as a default for a writing system.[7] We thus observe a huge gap between what Unicode defines and want NLP resources normally assume to be defined, This gap has to be bridged by XNLRDF.

To test XNLRDF we create, in addition to the data, small applications which are supposed to work for all or most writing systems that have a minimum of data. The set of applications currently includes a language guesser and a spell checker. These and all other applications to follow use only the data available within XNLRDF and thus show whether or not all necessary data types and tokens are included. For instance, while creating the web-interface of the spell checker, we recently discovered that the writing direction has to be provided explicitly by XNRLDF and cannot be left to the discretion of Unicode, the word processor or the Web-browser. Sorting, the function of uppercasing and the relations between characters and writing directions (some Chinese characters change their shape when written vertically or horizontally) are further examples of what kind of writing system-specific NLP information is needed beyond what is traditionally included in NLP.

XNLRDF however will provide more than these very elementary data. We will try to create word lists, dictionaries, corpora, stemmers, morphological analyzers and taggers for each writing systems. The challenge will be to find uniform representations and procedures which can correctly handle the great variety of languages, and, of course, to find or create the necessary data.

2.3. Different Ways of Data Creation

Traditionally, coordinated research is funded by a body which, more often than not, wants its money to be invested in what it perceives to be relevant for the financial resources of that body. Thus, research in France, payed by French tax payers is more likely to create NLP-resources for French than for Khamtanga. This, as natural as it seems, creates however a distortion of the relation between actual requirements and funding. As a consequence of this self-centered perception, those languages, which have the smallest gaps receive most funding.

A second feature of traditional models is that the cooperation between research units is organized in modules. Research units are thus autonomous within their modules and interact with other modules through specific interfaces, standards or protocols. In this way, intellectual properties can be easily assigned to a research unit. In addition, the consistency and coherence of the data within one module seem to be manageable. However, this model cannot take direct advantage of closely overlapping, complementary intellectual competences.

In models of organic cooperation however, volunteers, which may be experts or not, cooperate on the realization of some content, be it software, lingware, translation, images or a new texts, despite the absence of any funding (Bey et al., 2005). The only criterion for setting the research topic is the perceived relevance by the volunteers who, although not free of any self-centered perception, can accommodate more easily to an unbiased view than a funding body can do. Thus, while in the institutional cooperation, no language resources are created for Khamtanga, except in Ethiopia itself, researchers from France and many other countries would contribute to the development of Khamtanga NLP-resource in the model of organic cooperation. As a consequence, the gap between actual needs and research activities becomes smaller.

The cooperation in projects of organic cooperation is not necessarily modular. Different people might work on submodules where the function of the submodule cannot be defined on the basis on its own. This way different knowledge resources can be merged and software can be used to minimize friction and inconsistencies. Especially promising thus seem relational databases as they allow for a maximal fragmentation on the one hand, but guarantee on the other hand consistency and coherence through the usage of uniqueness constraints, references and triggers.

In XNLRDF we attempt to follow a model of organic cooperation, for a number of reasons. Firstly, the sheer amount of data we aim at is far beyond what one even large research team can achieve. Secondly, the many different competences required can only be brought together in an open model of broad cooperation. Third, the cooperation of professional linguists and volunteer experts can help to improve the database infrastructure, keeping it simple at the interface, yet complex and coherent in the data model.

3. A Relational Database as Backbone

While metadata and the organization of linguistic data in XNLRDF is determined bottom-up, we have principled ideas about the overall project design. As backbone for data development serves a relational database, whereas XML is used for the exchange of data in RDF (Powers, Sh., 2003), hence the name XNLRDF. The database can already be downloaded as database dump or as a one-to-one representation of the database in XML. An RDF will be designed which, in order to avoid bulky downloads, will allow for extracts for single languages and writing systems.

The relational database, installed with one command (in `Linux`) and configured with a few clicks in `Webmin` offers a set of features which can hardly be matched by XML. A relational database is integrated in a client-server architecture and designed for collaborative work. A battery of off-the-shelf interfaces is available for different purposes and can be used over the Internet.

A further advantage are the internal checks for data-types, uniqueness, coherence and consistency at a level below the interface so that these checks are effective in all interactions with the database. These checks will be primordial for the quality of the data when a great number of people cooperate blindly on the same database. The checks can can be defined to any level of complexity using *triggers* and *functions*. For example, changing the time period of Middle English in XNLRDF will change the time period for Old English and Early Modern English as well (thus assuring

[7] For a more detailed discussion of the shortcomings of Unicode see (Streiter and Stuflesser, 2005).

the coherence). Any attempt at placing e.g. the writing system of Proto-Norse in the former GDR, however, is most likely to fail due to temporal or local constraints associated with the language and the locality. Organizing data into a network makes singular incorrect data modifications difficult or impossible. *Freezing* an ever growing amount of validated data in this network, will make the space for incorrect modifications smaller and smaller.
Creating ambiguous metadata becomes impossible through *uniqueness constraint*. *References* make it impossible to delete central data, e.g. a language referred to by a writing system. The inclusion of *false positives*, e.g. pejorative language names, marked as deleted, make it impossible to insert or inherit the same value again through the effect of *uniqueness constraints*. Overall, XNLRDF forsees the following hierarchy of collaborators.

All users can enter new data. In this work, users are guided by the XNLRDF-browser[8] which already assisted in the creation of the currently available data in XNLRDF. Step by step the browser will evolve into a Wikipedia-like workbench where linguists can store, elaborate and test linguistic data.

A group of experts in language, linguistic subfields, language groups etc has the power to 'delete' incorrect entries, i.e. to move them into the false positives, or to assign the status of 'unchangeable' to cornerstone data. These experts thus complete and guide the set of control mechanisms provided by the system by controlling the validity of the data.

A third group of language and database experts defines the constraints and inheritance mechanisms to account for the completeness and coherence of the data. All of this is fairly easy to realize within relational databases but probably unreachable for XML.

4. Achievements

We have created a basic architecture for the development of fundamental NLP-resources for the writing systems of the world that might be fit for a model of organic cooperation. The potential of these resources starts to get visible with an automatic writing system (language) recognizer for currently about 700 writing systems and an spelling-checker for more than 700 writing systems. Most of the texts collected in 700 writing systems in XNLRDF are parallel texts, providing a means to create translation dictionaries for thousands of language pairs. Pointers to websites which can be freely downloaded for corpus construction are available for over 150 writing systems. In addition, the database contains a first set of about 2000 number words in 29 writing systems and 900 function words in 25 languages.

5. Further Outlook

While currently the database still requires a password for most modifications, a number of minor modifications have been opened to be changed freely by everyone. A Wikipedia-like cooperation of researchers is thus getting more and more likely. Opening the system step by step we explore techniques for checking new data and the automatic creation of message to the controlling linguists. At the same time we try to estimate the impact of erroneous entries on the quality of the data.
The data collection has focused until now on finding textual examples. We will proceed to a linguistic analysis of these examples to prepare the creation of corpora, stemming, morphological analysis and tagging. This work will be supported by small tools which propose different analysis solutions to be selected by the linguist.
Through the integration of simple applications, which among others test and show the potential of XNLRDF, we want to motivate researchers to enter the required data e.g. to insert open-licensed texts of a language to download shortly later a simple spell-checker, or to enter morphemes to download a better morphological analyzer. Some applications, like the spell checker provide for an inherent feedback function through which more linguistic data can be collected, e.g. the confirmation of unknown words. In addition, as suggested to us by Trond Trosterud, linguists might use integrated parsers or morphological analyzers to test their theories and produce at the same time word lists, classified morphemes and formal linguistic rules.
We hope that the creation and collection of data will speed up and extend to more languages once the system has been opened for organic cooperation. But even now, collaboration, advice and assistance of any kind, related to datastructure, metadata, the creation of applications, designing the final RDF, contributing data etc. are more than welcome.

[8] http://140.127.211.214/cgi-bin/gz-cgi/browse.pl

6. References

Y. Bey, K. Kageurat, and Ch. Boitet. 2005. A framework for data management for the online volunteer translators' aid system QRLex. In *Proceedings of PACLIC 19, the 19th Asia-Pacific Conference on Language, Information and Computation*.

Powers, Sh. 2003. *Practical RDF*. O'Reilly.

G. Simons and St. Bird (eds.). 2003. OLAC metadata set. Technical report.

O. Streiter and M. Stuflesser. 2005. XNLRDF, the open source framework for multilingual computing. In *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*.

O. Streiter, M. Stuflesser, and Q. L. Weng. 2006. Models of cooperation for the development of NLP resources: A comparison of institutional coordinated research and voluntary cooperation. In *Proceedings of the LREC workshop "Strategies for Developing Machine Translation for Minority Languages"*.

O. Streiter. 2005. Implementing NLP-projects for small languages: Instructions for funding bodies, strategies for developers. In *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*.

Ch. Uchechukwu. 2005. The Igbo language and computer linguistics: Problems and prospects. In *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*.