eScience in Linguistics

Eiríkur Rögnvaldsson
University of Iceland

eNoria Workshop on eScience in Higher Education
Uppsala
October 7, 2008

HÁSKÓLI ÍSLANDS

David Lodge: *Small World*

- „Well, I did my research on Shakespeare and T.S. Eliot," said Persse.
- „I could have helped you with that," Dempsey butted in. […] „It would just lend itself nicely to computerization," Dempsey continued. „All you'd have to do would be to put the texts on to tape and you would get the computer to list every word, phrase and syntactical construction that the two writers had in common. You could precisely quantify the influence of Shakespeare on T.S. Eliot."

DAM-LR Goals

- The goals of the DAM-LR project are to create an integrated and unified domain of:
  - trusted servers and services
  - deep metadata for research purposes
  - stable and unique resource identifiers
  - user management and authentication
  - exchange of user credentials for access authorization
  - longer-term potential for exchanging resources to strengthen preservation purposes

DAM-LR and eHumanities

- The DAM-LR (Distributed Access Management for Language Resources) project aims at virtually integrating various European language resource archives that allow users to navigate and operate in a single unified domain of language resources. This type of integration introduces Grid technology to the humanities disciplines and forms a federation of archives. It is the basis for establishing a research infrastructure for language resources which will finally enable eHumanities.

CLARIN

- The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable. CLARIN offers scholars the tools to allow computer-aided language processing, addressing one or more of the multiple roles language plays (i.e. carrier of cultural content and knowledge, instrument of communication, component of identity and object of study) in the Humanities and Social Sciences.

CLARIN Key Points

- The CLARIN initiative offers:
  - Comprehensive service to the humanities disciplines with respect to language *resources* and technology.
  - Technology overcoming the many boundaries currently fragmenting the resources and tools landscape as it is given by institutional, structural and semantic interoperability problems.
  - Tools and resources that will be interoperable across languages and domains, thus addressing the issue of preserving and supporting the multilingual and multicultural European heritage.
  - Comprehensive training and education programs that include university education in the different member states.
  - Improvement and extension of web-based collaborations, i.e. creating virtual working groups breaking the discipline boundaries.
  - Development or improvement of standards for language resource maintenance.
  - A persistent and stable infrastructure that researchers can rely on for the next decades.

CLARIN Key Technologies

- To achieve these challenging goals CLARIN will be built on and contribute to a number of key technologies coming from the major initiatives advancing the eScience paradigm:
  - It includes Data Grid technology to connect the repositories as being implemented in the DAM-LR pilot project and web services the various centres provide;
  - It builds on ideas launched by the Digital Library community to create Live Archives, and will further such initiatives;
  - It incorporates, and contributes to, Semantic Web technology to overcome the structural and semantic encoding problems;
  - It incorporates advanced multi-lingual language processing technology that supports cultural and linguistic integration.

The CLARIN Architecture Consists of

- a high capacity network layer provided by the European GEANT initiative strong enough to transfer even high resolution video streams
- language resource and technology repositories that archive language resources and offer language technology
- service registries that allow to register all sorts of services and expertise centers that help users to make use of the services
- an integration layer that can best be described by Grid type of services and an interoperability and access layer offering services that will enable Semantic Web type of services
- an application layer that makes use of all services, registries and repositories to allow users to tackle the grand challenges in the humanities including a semantically rich humanities domain

Kubrick: 2001: A Space Odyssey

- Dave Bowman: Open the pod bay doors, HAL.
- HAL: I'm sorry Dave, I'm afraid I can't do that.

What would it take to create at least the language-related parts of HAL? Minimally, such an agent would have to be capable of interacting with humans via language, which includes understanding humans via **speech recognition** and **natural language understanding** (and, of course, **lip-reading**), and of communicating with humans via **natural language generation** and **speech synthesis**. HAL would also need to be able to do **information retrieval** (finding out where needed textual resources reside), **information extraction** (extracting pertinent facts from those textual resources), and **inference** (drawing conclusions based on known facts).

NGSLT

- Nordic Graduate School of Language Technology (NGSLT)
  - More than 30 participating institutions
  - In all Nordic and Baltic Countries and NW Russia
- Financed by NordForsk, 2004-2008
- Two types of courses:
  - Short courses (usually five days)
  - Longer courses with intensive lecture periods and distance learning

“Vismansrapporten”

- In 2005, the Nordic Council of Ministers commissioned a ten-year plan in the form of an expert panel report (= vismansrapport) for making the Nordic Countries a leading region in language technology (LT).
- The aim of the report was to identify the common key areas which need to be addressed when making the Nordic countries into a leading region.
- The report highlights key areas, magnitudes of investments, suggested partners, modes of cooperation and some initial key actions.

LT Training and Education

- More cooperation is needed in academic training among the universities in the Nordic/Baltic region. A sufficient number of highly skilled PhDs and Masters ought to be trained with the best possible LT skills and all countries and language groups should be participating, including minorities and small language communities.

Action Plan

- To implement the goals and to further specify the areas and their time-frames in the 10-year plan, we suggest the following steps in allocating resources:
- Establishing NEALT and its working groups
- Commissioning BLARK reports for the Nordic languages
- Nordic funding for cooperation on LT training and education
- National funding of medium-term applied research projects involving university and industrial partners

NMPLT

- Nordic Master Program in Language Technology (NMPLT)
- Proposal submitted to the Nordic Council of Ministers in 2007
- Participants: 7 universities in Iceland, Denmark, Finland, Norway, Sweden, and North-western Russia
- Unfortunately, the proposal was not successful

Objectives of NMPLT

- To strengthen the cooperation in LT among Nordic, Baltic and NW Russian teachers and scholars and enable the deepening of expertise by division of labor. The present environment does not encourage specialization in many areas with great potential because much capacity is absorbed in routine education.
- To provide better quality teaching to the Nordic, Baltic and NW Russian students of language technology and a wider array of possible areas of specialization for students. Different institutions may join their forces in building better materials, assignments and methods for courses.
- To attract talented and well motivated students to the field from within the region and from other countries. The proposed program would be very exceptional in the breadth of teaching it offers.

Proposed Organization of NMPLT

- The Nordic Master Program in Language Technology is organized as a network of masters' level teaching in language technology where the program defines a framework of the master's degree and provides shared distant learning courses and supervision for the students. A student completes the degree in his/her own university according to the requirements accepted by the local home university faculty.
- The network approach allows for great flexibility in offering fields of specialization. In this way the program can offer all areas of specialization to all of its students, including the possibility to write one's master's thesis under a distant supervisor. Each home university (or faculty) is responsible for granting the master's degree according to its own regulations and national legislation.

The Need for Nordic Cooperation

- Resources for teaching language technology (LT) within the region are limited and expertise of different branches of LT is concentrated to different sites. Sharing the resources through distant learning and joint development of courses would result in better quality teaching and wider selection of courses for the students. Departments could use their teaching resources more efficiently by delivering only a part of the courses and relying on the other members of the consortium in some courses.
- The project would make the Nordic/Baltic region more visible in master's level studies of LT by international cooperation in building the course materials and by attracting students from outside the region. The NMPLT is inspired by and builds on the experiences of NGSLT.

Proposed Structure of NMPLT

- The NMPLT starts as a network of masters' level teaching in language technology where the Nordic program defines a framework of the master's degree and provides shared distant learning courses and supervision for the students. A student completes the degree in his/her own university according to the requirements accepted by the local home university faculty. The framework sets requirements for such degrees in order to qualify as a NMPLT degree and thus guarantees the necessary uniformity and compatibility among the members of the consortium.

Why the Network Approach?

- This approach of a *network rather than a strictly administrated joint studies degree* was chosen because of several reasons. A strict program where the requirements for master's degree are identical in all universities would be time consuming to establish and difficult to. To avoid time consuming and possibly difficult harmonizing efforts of these less important features, it was decided to start this program with a network type of a structure with much lighter bureaucracy.
- Another reason for choosing the network approach was that it allows for greater flexibility in offering fields of specialization. In this way the program can offer all areas of specialization to all of its students, including the possibility to write one's master's thesis under a distant supervisor.

The Current Nordic Situation

- NGSLT is in its final year
- The Action Plan in “Vismansrapporten” has not been implemented
- NMPLT was not funded
- The number of universities offering programs in Language Technology and/or Computational Linguistics has decreased