CLARIN
Common Language Resources and Technology Infrastructure

Complex Computing and Data Infrastructure Challenges – the Case of Language Based Materials

Bente Maegaard, University of Copenhagen, Centre for Language Technology
CLARIN ERIC Vice director

Member of the ESFRI Social and Cultural Innovation SWG

What is CLARIN?

- *Common Language Resources and Technology Infrastructure*
  - CLARIN is an ERIC since February 2012
  - RI for researchers in humanities and social sciences
  - Goals: to make it possible for humanities and SS researchers to take advantage of technology, and to share data and tools
  - Methods
    - to create a European federation of digital archives that include language based data (text, audio, video, multimodal)
    - to integrate services and tools to manipulate, explore, enhance and exploit the data
    - to give easy access
    - to cover as many languages as possible
    - to collaborate widely

List of challenges

- Users
- User interfaces
- Repositories
- Services
- Legal issues
- Sustainability
- Computing
- Knowledge sharing, education
- Cooperation

Users

- Characteristics
  - Wide spread user base
  - Scattered over hundreds of universities and research institutions (CLARIN PP project had 215 member sites in 33 countries)
- Challenges
  - Need for well-scaling, easily deployable middleware and standards
  - Usability for the non-expert user
  - Reach users that do not know themselves that they are users
  - Difficult to avoid heterogeneity – both in data and metadata
  - Risk that the wheel is re-invented

User interfaces

- Characteristics
  - Users are humanities researchers
  - They need to search in and manipulate very large repositories
  - They need to be able to design their service workflow
- Challenges
  - They need to search in hundreds of thousands of documents with different structures
  - They need to search in metadata mapping to thousands of concepts
  - We should not demand that they master XML
  - We should give them automatic tools for designing their workflows in the most efficient way

Examples of data and services

- Text, old and modern
- Literature, language for special purposes
- Parallel texts for translation studies
- Videos, - audio and gestures
- Newspapers, news on other media
- Parliament debates
- Tomb stones

- Add annotation (e.g. morpohology, lemma, analysis of gestures)
- Search all occurrences of the same gesture
- Find the most common pattern of xx
- Find all names in historical texts
- Find all different pronunciations of the letter ’a’ in Danish and their frequency
- Find positive or negative expressions relating to islam in NL and DK newspapers between 1980 and 2000.

Repositories

- Characteristics
  - Large volumes, continuous expansion
  - Heterogeneous material: text, audio, video, multimodal
  - Heterogeneous types: texts, recordings, grammatical annotations, dictionaries, ontologies
  - Many standards for data and metadata.
- Challenges
  - Users should see one single collection
  - Agreements on standards, for data and metadata (semantics, ISOcat)
  - We have 700 data categories, many with definitions, in 17 languages
  - We have 53 metadata profiles, based on 235 components, resulting in 425,000 metadata records in our VLO at present
  - Granularity: how to model our repositories – fedora-commons cannot cope with more than 100 mio entries

Services

- Characteristics
  - Not just **metadata** search and retrieval, but also **content** search and
  - Applications of advanced language and speech technology tools as services in a service oriented architecture
    - Located in different places
    - Operating on distributed non-uniform collections
    - Combinable into workflows
- Challenges
  - Interoperability standards to make everything fit together
  - Usability for the non-expert user
  - Accommodation of new tools and technologies

Legal issues

- Characteristics
  - Heterogeneous access and licensing systems
  - Differences in national IPR legislation
  - Access is too limited
- Challenges
  - Uniform access and authentication
  - Lightweight general licensing system – CLARIN has made a first step in this direction
  - Overcoming legislation issue
  - Overcoming national legislation issues
  - Need to have more open access

Sustainability

- Characteristics
  - No systematic registry system for data and tools
  - Legacy material
  - Legacy tools
  - Digital sustainability is relatively new for many centres and users
- Challenges
  - We need persistent identifiers
  - We need a persistent descriptive apparatus, persistent interoperability standards
  - We need to be able to follow the technological evolution
  - And, computing centres need to be able to issue long term guarantees on web services, and data repositories

Computing

- Characteristics
  - Many new resources are demanding (more than just text)
  - Many new tools and techniques are statistical (e.g. data mining) or require training on large volumes of data (e.g. speech recognizers)
  - Data intensive web services (video processing etc)
  - Data and tools may be distributed over Europe, so demands an efficient network
- Challenges
  - Performing statistical operations on large distributed, non-uniform data collections
  - Example: Uploading several gigabytes to a REST or SOAP web service

Knowledge sharing, education

- Characteristics
  - The CLARIN national teams have different expertise
  - Important to share knowledge and upgrade teams
  - Important to teach students, to make sure next generation is well prepared
- Challenges
  - Teams are very scattered
  - Can we find enough researchers to take this teaching on
  - Decide how important it is to know XML, regular expressions and other technical stuff

Cooperation

- Characteristics
  - Many research infrastructures (an similar bodies), in the EU and nationally
  - Also many counterparts internationally
  - CLARIN is now an ERIC and can set up collaboration better than on a project-project basis
- Challenges
  - Ensuring linking capabilities between RI at the national, EU and international level
  - Maybe think about broader collaboration platforms at the European and international level?
  - However, - the CLARIN ERIC is very lean at the ERIC level, all finances for technical infrastructure is at the national level

CLARIN requirements

- CLARIN is an ERIC
- Currently 9 members (Austria, Bulgaria, Czech Republic, Denmark, Dutch Language Union, Estonia, Germany, Netherlands, Poland)
- Funding for CLARIN comes from national sources
  - Fee
  - In kind contributions
- One of the national in kind contributions is a data and service centre (or more centres, may be distributed)
  - Hosting the data
  - Providing the services 24/7
  - Responsible for the authentication, authorisation, access (DK: WAYF); entering the federation
  - Giving long term guarantee – 50 years?
- In some countries universities have such centres, in other countries there are dedicated data and service centres

Concluding remarks

- Our technical challenges are important, but probably not HPC
- Sustainability is important
- Longterm collaboration preferred