CLARIN and FLaReNet workshop in Stockholm 25-26 Nov 2009

Steven Krauwer

Utrecht institute of Linguistics

UiL-OTS

What is CLARIN

- Common Language Resources and Technology Infrastructure (http://www.clarin.eu)
- Basic idea:
  - European federation of digital archives with language data and tools (text, speech, multimodal, gesture …)
  - target audience humanities and social sciences scholars
  - with uniform single sign-on access to the archives
  - with access to language and speech technology tools through web services to retrieve, manipulate, enhance, explore and exploit data
  - all languages are equally important
  - to cover all EU and associated countries

What is FLaReNet

- Fostering Language Resources Network (http://www.flarenet.eu)
- Developing a common vision and strategy in the field of language resources
- Short medium and long term objectives
- Plan of action for EC, national organisations and industry

Similarities and differences

Common points

- both strong focus on resources and technology
- both long term

but CLARIN

- addresses social sciences and humanities scholars
- and builds an infrastructure for them

whereas FLaReNet

- addresses EC, national organisations and industry
- and makes policy recommendations for them

Why this workshop?

- Most CLARIN players originate from text processing, speech is underrepresented
- Speech and multimodal resources are becoming more important, both for e.g. historians of the future and social scientists
- Sharing speech and multimodal resources pays off because of the high cost (in time & money) of annotation
- Consultation with the community, especially about standards, urgently needed
- Common interest of CLARIN and FLaReNet

Thanks, KTH for organising and hosting this event
- and let’s get started!

PS

- The following slides were not presented at the workshop but they provide some more general background on CLARIN, including the slides on sharing that were shown during the discussion

Who are the people

- At this moment a core consortium of 33 partners in 23 EU and associated countries (and more to join)
- Outside the consortium ca 140 contributing institutions in 32 countries in Europe
- Mostly academic institutions and a number of digital archives
- Contributions consist typically of data, technology, or expertise

When will it start (and what will it cost)

2008-10: *Preparatory phase*

- funded by the EU (grant 212230, 4.1 M€ , 33 consortium partners in 23 countries, plus over 140 other organisations in 32 countries), with (at this moment) additional funding from 19 national governments (> 14 M€, ranging 50K€ - 5M€)

2011-14: *Construction phase*

- to be funded by the member states (100 M€ needed, 5 M€ committed by 1 country, more to follow, 0 € from EC)

2015-…: *Exploitation phase*

- to be jointly funded by national governments, max 20% EC

2008-2018: estimated total cost ca 146 M€

Do we really have to wait until 2015?

- First small experimental prototype during this phase, but no real end user services
- If we get the green light (and the €€) for the next phase (construction) we may gradually start in 2011-2012
- Every country responsible for its own content, no central funding from EC foreseen
- What will be available (content and services) will depend on what countries do, and I don't expect them to start all at the same time

What are the main challenges or obstacles?

- We look at a few where you might be able to contribute to the discussions:
  - technical
  - linguistic
  - take-up
  - legal
  - business models
  - governance and funding

Main challenges
Technical

- Technical challenges:
  - Interconnecting existing archives that may use very different ways to encode and describe data
  - Ensuring that existing language technology tools made for material in archive A will also work for material in archive B, and will work together
  - Needed: common standards
- Current position: Support for a limited set of (de facto standards) for interoperability
- Action for you: Read our Standardisation Action Plan to see what we propose and participate in the standards discussion!

Main challenges

Linguistic

- Linguistic challenges:
  - Ensure that all languages are sufficiently covered in terms of data and tools
  - Ensure that we know what exists
  - Ensure that approach adopted fits for all languages
  - Needed: broad consultation (e.g. about standards) and verification (for each language)
- Actions for you:
  - Read standards document and protect the interests of your own language and of your research community
  - Register your resources and tools so that they become visible

Main challenges

Take-up

- Take-up by target audience:
  - aim at humanities and social sciences scholars
  - who have no technical background and who have very little tradition in using technological tools
- Special challenges:
  - discovering what they need
  - making them aware of the potential benefits of the infrastructure, e.g. to speed up or innovate their research
- Action for you if you are part of our target audience: Formulate your requirements and communicate them to us!

Main challenges

Legal and ethical

- Legal challenges:
  - making a light access and licensing system for the users
  - protecting owners' rights and interests
  - respecting national IPR legislation
- Special problems:
  - transnational access and diversity of national legislation
  - repurposed data (e.g. using novels or TV news for linguistic studies)
  - ethical & privacy considerations (e.g. use recorded phone calls to train speech recognition systems)
- Action for you: Read our documents (to appear) about licenses and check whether they cover your needs (also in relation to national IPR legislation)

Main challenges:
Business models

Expectations depend on your role in life:

- *Everything should be available for free*
- *I want to be reimbursed for the extra effort to make my data and tools accessible through CLARIN*
- *I don't want others to use my results to make a profit*
- *Funders should not pay for the creation of tools and data that can be bought on the market*
- *Funding infrastructures is a primarily a national responsibility*
- *We fund you for now but we expect you to become self-sustaining*
- *Creation of data and tools is the responsibility of the infrastructure*

Question: who should pay for what to whom, and why

Main challenges

Business models

- Building and maintaining an infrastructure costs money, but where should the money come from?
- Current position:
  - Every country pays for its own CLARIN construction and operations
  - All countries together fund central generic operations and overall coordination
  - After construction the EC may also contribute
  - Researchers in participating countries have free access to the whole infrastructure

Governance challenges

Future shape of the infrastructure

Some features of the RI as we see it

- networked digital infrastructure with one or more centers in most participating countries:
  - data centers (24/7 availability)
  - service centers (24/7 availability)
  - centers of expertise
  - other centers (more loosely connected to the infrastructure)
- all based on or hosted by existing centers
- small head office for general coordination
- no major initial investment in physical installations or buildings required, but …
- … construction never ends (new data, new tools)

Action for you: Read our document about centers and decide about the role your center wants to play

Governance challenges

Organisation and funding

- Find a legal form that allows 23 or more countries
  - to jointly build and operate an infrastructure distributed over all these countries
  - to jointly fund the construction and exploitation in a sustainable way
- At this moment we feel inclined to adopt the new legal entity the EC has just created: ERIC (European Research Infrastructure Consortium; members are governments, not universities)

Intermezzo

- How to get hold of the documents I mentioned:
  - If your organisation is a CLARIN member you can get an account on the site www.clarin.eu and get access to all documents
  - If they are not a member, ask them to join
  - If they don’t want to join or don’t qualify for membership contact me: s.krauwer@uu.nl
- How to participate in discussions:
  - Join a Working Group, open to all staff from member sites

Sharing

- What can you share through CLARIN:
  - anything that might be relevant for our user community, and
  - that satisfies certain quality criteria, and
  - that you are legally allowed to share (raw or annotated data, transformed data, tools, programs, expertise, etc)

Why share at all?

- For the researcher
  - Idealism
  - Becoming famous
  - Hope that others will share with you
  - Because your funder tells you
- For the funder
  - Better return on their investment: reusability

Why not share?

- It may involve an extra effort (adapting to representation or interoperability standards, creating metadata, documenting)
- Others may do brilliant things with your data that you would never have thought of
- Others might criticize your stuff

Sharing

- Our position: every resource that has been created on the basis of public funding should in principle be shared
- How can you share: by depositing your material at one of the registered CLARIN Centers
- See Centers document for more details

What makes sharing difficult in CLARIN

- What you share has to work with what others share
  - Interoperability standards
- What you share will have to be usable by non-technical people (SSH scholars)
  - Much effort on training and user-friendliness
- Sharing has to be technologically sustainable
  - Need for flexibility and adaptivity as the world changes
- Sharing requires long-term preservation
  - Federation of trusted archives, curation
- The whole infrastructure has to be financially sustainable
  - Should be owned by governments

Concluding remarks

- CLARIN is still full of challenges and needs your input on many issues
- Remember that if you don't take care of your language no one will!
- CLARIN is not about content creation, but about providing access to what exists (or will exist)
- We have a number of interesting discussions ahead of us about non-content issues such as business models – please participate!
- In brief: never a dull moment in CLARIN!

THANKS