Collaborative Infrastructures to enable e-Science

Peter Wittenburg
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
CLARIN Research Infrastructure
Content

- relation e-Science and infrastructures
- big and small challenges
- ESFRI process as consequence of the debates in Europe
- eco-system of infrastructures
- collaborative data infrastructure
- EUDAT initiative

- will not discuss at length what CLARIN is intending and what we did so far -> http://www.clarin.eu
Given our human capabilities to change our conditions of life in all aspects we cannot simply continue with the old paradigms in research.

John Taylor:
“e-Science is about global collaboration in key areas of science and the next generation of infrastructures that will enable it.”

As for building new fast trains we need new tracks, new signaling options, etc.
in all major areas we see grand challenges:

• how to come to a **stable climate** in which next generation can survive?
• how to solve our eminent **energy problems** given the enormous effects on the environment?
• how to maintain a stable **health** given all environmental changes and influences?
• how to maintain stable societies given the globalization affecting our cultures and languages?
• how to maintain stable minds given cultural changes and increasing technological innovation?
• etc.
major scientific break-throughs were achieved by the small groups driven by scientific curiosity

• so let’s not forget these “small challenges”

• in our domain of languages and mind:
  – how does our human brain/mind process language?
e-Science - the “small” challenges
major scientific break-throughs were achieved by the small groups driven by scientific curiosity

- so let’s not forget these “small challenges”

- in our domain of languages and mind:
  - how does our human brain/mind process language?
  - how have the 6500 languages still spoken developed over time?
e-Science - the “small” challenges

according to this dependency tree Taiwan is at the root of Polynesian languages.
major scientific break-throughs were achieved by the small groups driven by scientific curiosity

• so let’s not forget these “small challenges”

• in our domain of languages and mind:
  – how does our human brain/mind process language?
  – how have 6500 languages still spoken developed over time?
  – how to improve automatic access to multimedia information to make them usable for research purposes?
e-Science - the “small” challenges
major scientific break-throughs were achieved by the small groups driven by scientific curiosity

• so let’s not forget these “small challenges”

• in our domain of languages and mind:
  – how does our human brain/mind process language?
  – how have 6500 languages still spoken developed over time?
  – how to improve automatic access to multimedia information to make them usable for research purposes?

  – many more of these challenges (in all disciplines)
Impact of J. Taylor

European Strategy Forum on Research Infrastructures (ESFRI)

- more than 40 research infrastructures started working
- all aimed to create persistent services to the researchers
European Strategy Forum on Research Infrastructures

Social Science and Humanities

Impact of J. Taylor

CLARIN is where my group is engaged in fully distributed domain
• do all these 40+ RI have to solve the same basic tasks?

- Within Community Services
  - CLARIN

- Domain Services
  - SSH in preparation

- HPC Services (DEISA->PRACE)
  - available - being extended

- Data Services
  - in preparation (EUDAT)

- Grid/Cloud Services (EGI)
  - available - in discussion

- Network Services (GEANT)
  - available - being extended

- e-Infrastructures

• no of course not - this would not be efficient
• need to build on common services where possible
• but finding a good mutual understanding is not simple
Example 1: trust federation

State CLARIN SPF
- 4 German centers
- Meertens, INL, MPI
- Nancy
- U Helsinki
- CSC
- U Vienna
- CU Prague
- DANS
- U Copenhagen
- U Bergen
- U Gothenburg
- U Oxford
- U Lancaster
- U Aix en P
- about 10 more to come
Example 1: trust federation

Contracts with IdPs
- Finland
- Germany
- Netherlands
- Sweden
- Norway
- Denmark
- Iceland
- France
- Austria
- Czech Republic
- UK
more countries to come

now: large number of researchers who can operate on virtual collections using sso
now: potential of large number of users to execute processing chains
Example 1: trust federation

German National Identity Federation

European Identity Federation (GEANT/eduGain)

will become the CLARIN ERIC

User

Depositor

CLAIRN Service Provider Federation
Riding the wave

How Europe can gain from the rising tide of scientific data
a vision for 2030

Report der High Level Expert Group on Scientific Data from 6. October 2010
Collaborative Data Infrastructure

A Collaborative Data Infrastructure - a framework for the future

several communities have a proper data organization solution i.e. what is the right, abstract interface?

- CLARIN
- DARIAH
- CESSDA
- LifeWatch
- ENES
- etc.

- EUDAT
- D4Science
- etc.

Workbenches
Portals
Web Apps etc.
How to organize CDI

“bottom up” from Communities

“top down” from IT

• need a dual approach
• there are waves
  (a) where particular solutions are in focus to get scientists on board
  (b) where IT experts start to generalize

• currently we see a move towards bottom up i.e. different languages, different solutions, etc.
Two “data issues” from CLARIN

1. How to take care about long-term curation and preservation of patrimonial data?

2. How to ensure that workflow chains on stored data can be executed by everyone?
   – in our domain capacity computing

• **EUDAT** wants to address these topics
  many communities and data centers on board
  a long bi-directional interaction as basis
Safe Replication for LTP

- since 2004 a LTP strategy in Max Planck Society
- yet no systematic European solution!!
- 80% endangered
- yet no safe and rule-based replication!!
- using EPIC services and iRODs
- in addition 13 regional archives worldwide to help human heritage to survive (10 requests)
Distributed Workflow Execution

- Complete chaining system in operation - running on departmental servers
- Needs to be available for all interested researchers in Europe (+ beyond)
• yet no robust solution for attribute delegation for web services

• have joint projects with Dutch Grid colleagues, but must be a service for everyone in Europe (and beyond)
Technology Issue II

• why is CLOUD interesting for us?
  – it’s a technology - it does not solve all our problems
    • it does not solve long-term curation/preservation
  – it allows to store much data and protect access from outside
    • but that’s not the only issue
    • issue for researchers is internal data access and flow control
  – it allows easy service deployment
  – it caters for scalable capacity computing

• as Community we don’t care too much which technology is used
  – robustness, persistence
  – decent level of security
  – not forcing us to change our data organization
In Europe we seem to be on a very good way to

- build research infrastructures “bottom up” for advancing science/research
- infrastructures = not projects
- build and interact with horizontal e-Infrastructures to create an eco-system of infrastructures
- will take some time until all of this will work seamlessly and cost efficient

Thanks for your attention.