



# CLARIN

## Common Language Resources and Technology Infrastructure

Daan Broeder  
Peter Wittenburg

# What is 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 for Language & SSH (Social Sciences & Humanities) researchers.

- Resources: Lexica, text corpora, multi-media/multi-modal recordings, ...
- Technology: applications, (web-)services, ..

# The problem



- Existence and location of resources only known to insiders
- Archives mostly unconnected islands
- Every archive has its own standards for storage and access
- Normally need to download first when processing resources
- Social sciences and humanities researchers are not language or speech technologists
- They are often not aware of the potential benefits of using language and speech technology
- Available tools are hard to use for non-specialist

# CLARIN overview

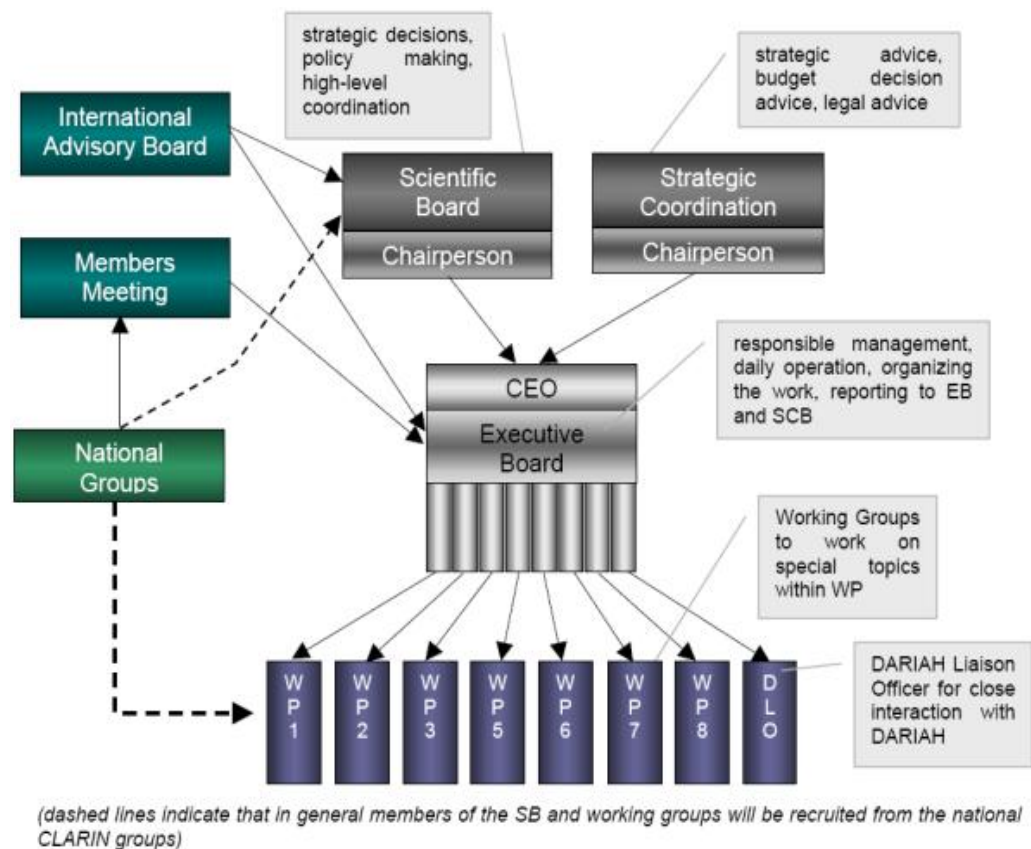


- CLARIN is an EU Infrastructure project with 4.2 ME funding for a 3 year preparatory phase
- Additional funding from national governments (at this moment at least 7 ME)
- The CLARIN consortium has now 32 partners from 26 EU countries
- The CLARIN community has 132 member organisations in 32 countries (mostly from NLP orgs.)
- CLARIN is based on 4 earlier initiatives with many participants: LangWeb, EARL, TELRI, LIRICS and more recent DAM-LR

# CLARIN Organization



WP1	Management & Coordination, Steven Krauwer, OTS U. Utrecht NL
WP2	Technical Infrastructure, Peter Wittenburg, MPI Nijmegen NL
WP3	Humanities Overview, Tamas Varadi, Hung Ac. Sc., Hun
DLO	DARIAH Liaison, Martin Wynne, Oxford Univ. UK,
WP5	Language Resources & Technology Erhard Hinrichs, U. Tuebingen, Ger
WP6	Dissemination, Dan Crista, U. Iasi Rom
WP7	IPR and Business Models, Kimmo Koskiennemie, Univ. Helsinki Fin
WP8	Organizational Agreements, Bente Maegaard, Univ. Copenhagen, Dk



# Time plan

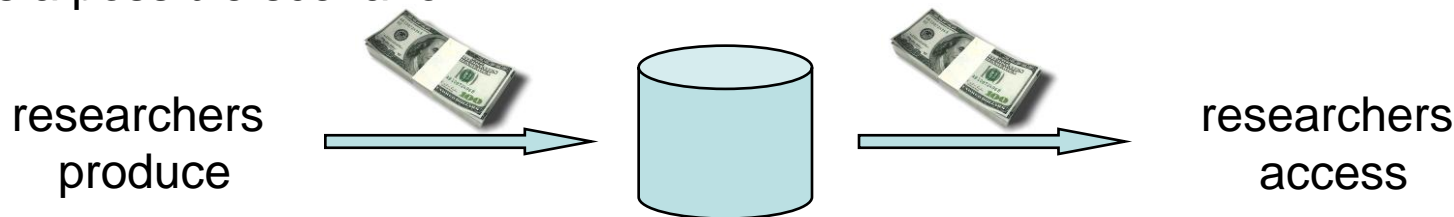


- 2008 - 2010 Preparatory Phase
  - Limited set of federated centers (10+)
  - Showcases, demonstrators
  - WP8: Investigate embedding in national funding schemes for construction phase & maintenance
- 2010 - 2020 Construction Phase
  - No important European funding
  - Depend on national project commitments (Germany, ...)
- 2020? - ... Maintenance Phase
  - Has to be cost efficient, we have to compete!

# The CLARIN Challenge I



- What are we aiming at? call it an LR&T Service Federation
- what is now the deal for primary and secondary research resources?
  - we slowly start understanding the digital dilemma
    - need to store digital research resources – but digital archives are living things
    - creating and maintaining access to digital resources costs money (!)
    - storing a conventional film master copy costs about \$ 1.000
    - storing a digital master copy costs about \$ 12.000
- here is a possible scenario

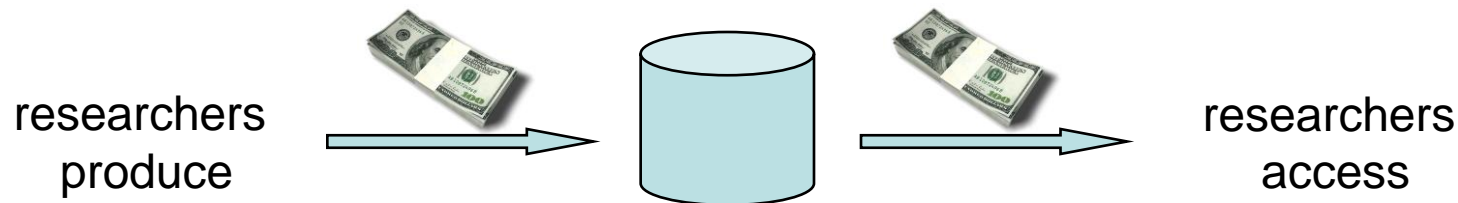


Google/Microsoft?/... store and give access  
who will have control in 10 years?  
what dependencies will emerge?  
which business models will emerge?

# The CLARIN Challenge II



- What are we aiming at? call it an LR&T Service Federation
- do we want this deal? NO or?
- here comes CLARIN



CLARIN Federation stores and gives access  
CLARIN research community will determine the rules!  
It will cost – and we need to keep the costs low!  
Will we manage? – don't know yet!  
We will be in a competitive situation!

EC promotes it – they want us strong and independent  
We cannot wait – need to start giving CLARIN services now!

# CLARIN Holy Grail User Scenario



A researcher authenticates himself with his own organization and creates a “virtual” collection of resources from different repositories. He does this on the basis of browsing a catalogue, searching through metadata, or searching in resource content. He is then able to use a workflow specification tool and process this virtual collection with possibly a mix of home grown and remote service components. Resulting data can be added to the origin repositories with proper access rights and the “virtual” collection specification can be stored for future reference.

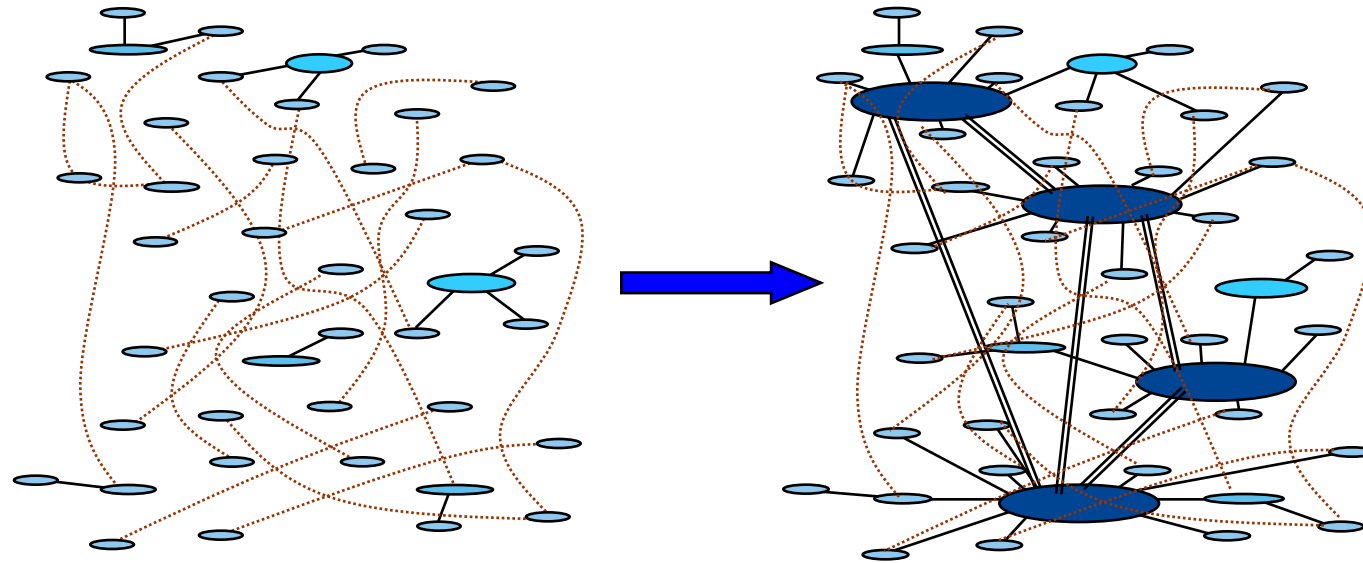
This is very ambitious and challenging, but even a partial realization is worthwhile!

# Dimensions of Infrastructure



- **Service Centers**
  - what kind of services will we offer – which centers can offer them
- **Federation Infrastructure**
  - there is middleware to be set up to get the LR&T federation working
- **Registry & Catalog Infrastructure**
  - if we want to build the big market place of LR&T we need to find ways to organize it so that users find their way
- **Web Services Infrastructure and Workflow Mechanisms**
  - we need to have mechanisms allowing anyone to easily integrate their applications – the LR&T components
- **Some demo applications**
  - need to show the potential and how combination of components is done
- need to make **cost estimates** for the construction and operation phases

# Service Centres

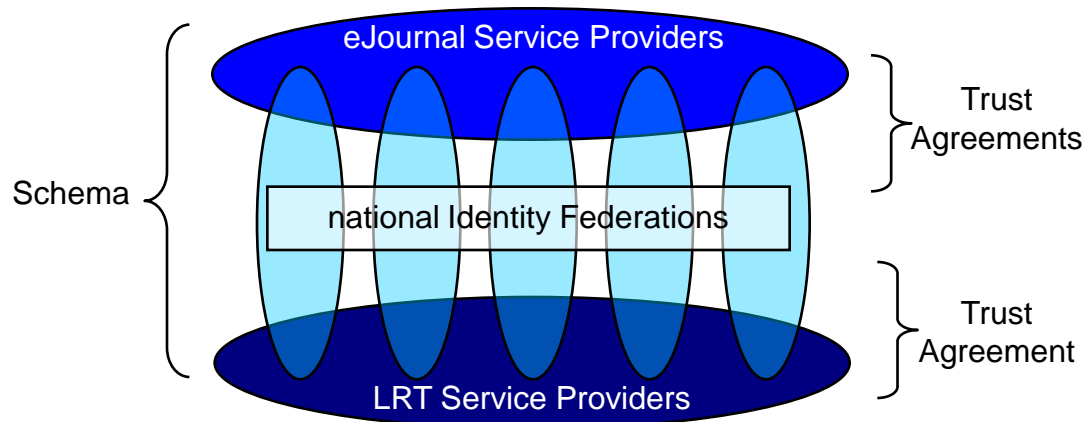


- need to add a persistent infrastructure layer on top of the landscape formed by accidental and temporary collaborations that is easily accessible for everyone and that offers high availability so that people can rely on it
- will be different types of centers dependent on the service
- centers need to change their attitudes – they have to acquire a service mentality
- need a strong national support for many years
- looking at the above, its clear not all existing centers can participate on all levels

# Federation Infrastructure I



- support for virtual collections with resources from different archives based on metadata, profile matching, scientific searches, etc.
- “middleware” pillars
  - joint metadata catalogue for resources and tools partly based on existing practice
  - common way of referencing electronic resources in the federation
  - single sign-on/identity middleware
  - trust agreements with national identity federations
    - CLARIN needs to build a service federation based on simplified and unified rules for licensing, accessing, user authentication etc (WP2 + WP7)
  - all communication based on trusted and signed certificates



# Federation infrastructure II



- AAI & Identity Federation issues
  - Shibboleth for SSO browser access
  - ... but applications & services need to have access too
  - Will we only interact with national IDFs, what are the attribute schemas and the encodings. Use of TERENA SCHAC work.
  - Interaction with GRID AAI
- PID framework for resource reference and citation
  - Have experience with Handle System – robust, scalable, flexible, but some functionality is missing
  - Build extra services for resource citation, virtual collections, ...
- Metadata harvesting
  - OAI protocol, XML harvesting (IMDI)
- Repository system
  - Need upload service with versioning, access service
- Resource duplication
  - Respect owner authorization rules
  - Synchronize auth. information amongst archives housing resource copies.

# Catalogues & Registries



- joint metadata catalogue for resources and tools based
  - **Metadata itself is an increasingly important research resource**
  - learned a lot from IMDI, DC/OLAC, TEI etc during the last 8 years
  - how to describe tools? DFKI's LT registry experience
  - how to support national maintenance – hierarchical architecture?
  - how to present large domain of LR&T – which taxonomy tree? (WP5)
  - personalized dynamic tree extraction
- new modular and more flexible metadata schema to be worked out based on a comprehensive taxonomy of LRT (WP5)
  - important is that everyone can create his/her own schema!!
  - everyone can use localized and sub-discipline terminology
  - re-use of existing categories registered in the public ISOCat concept registry is a **MUST**
  - need a new infrastructure (definitions, registry, tools, geo browsing, gateways, ...)
- support for virtual collections with resources from different archives

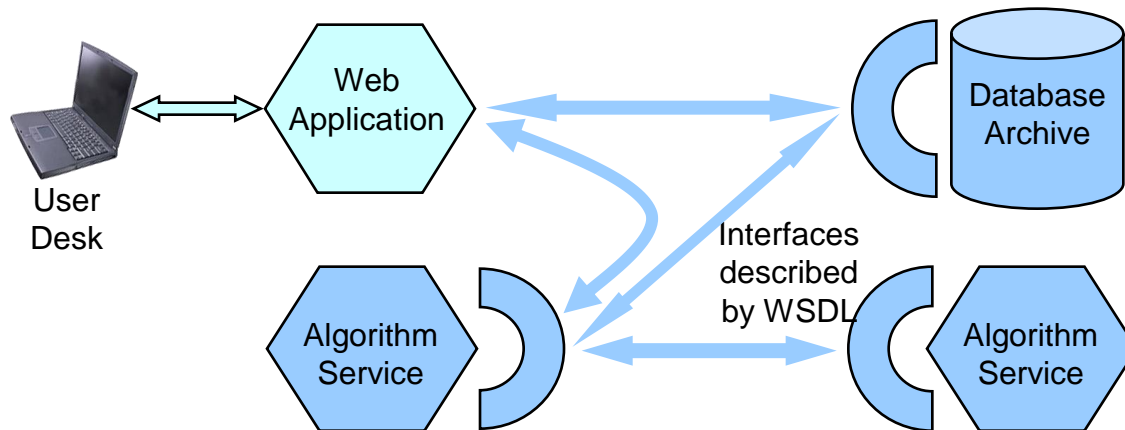
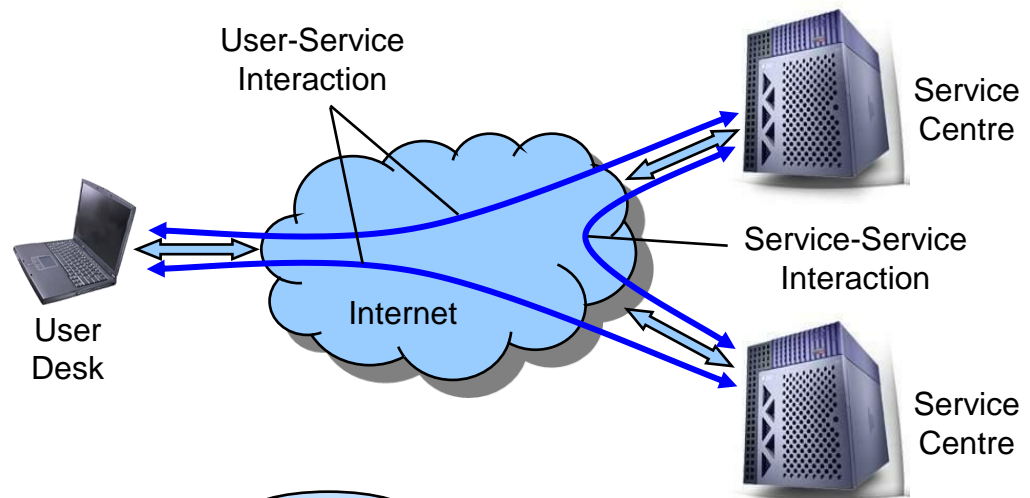
# Service Oriented Architecture



still live in a down-load first and not in a cyberinfrastructure world

current way of interaction:

- user interacts with a web-site
- receives intermediate result
- manipulates this result and
- sends it to the next web-site
- etc



better way of interaction:

- users interacts with an application
- the application makes use of different services without bothering the user
- user receives the final result

- SOA not at all simple to achieve, but is the only architecture scalable and flexible enough
- standardization and harmonization is required to realize workflow mechanisms

# Part of a Larger Game



- CLARIN is part of a larger game covering many disciplines
- at the EU level we have
  - the DRIVER initiative to harvest discipline metadata and publications
  - the DARIAH initiative to act as an harmonization umbrella for the humanities
  - the ALLIANCE initiative to harmonize between actually all disciplines at various levels (centers, IPR issues, standards, ...)
  - TERENA: SCHAC (Schema Harmonization) and TACAR (CA repository)
  - who knows what else ??
- at the global level we have many attempts to collaborate at the infrastructure level
  - PID frameworks
  - work in ISO TC37: ISOCat, CitER
- We need a very sensitive balance between bottom-up approaches driven by the communities and top-down approaches.

# Infrastructure after 3 Years



- Service Centers
  - should have a network of > 10 dedicated centers
- Federation Infrastructure
  - these centers should support all middleware pillars
- Registry Infrastructure
  - should have a ready made registry infrastructure, ISO process and a critical mass of entries for LRT components
- Web Services Infrastructure and Workflow Mechanisms
  - need to have a clear specification, wrappers & encapsulation tools, a critical mass of integrated LRT components and a first graphical WF framework
- some demo applications
  - should have some nice demos ready
- cost estimates for the construction and operation phases to be ready
- and proposals how to embed CLARIN maintenance into national & EC funding schemes



# The End

## Thank you for your attention

More info: [www.clarin.eu](http://www.clarin.eu)

Please take the flyers & Newsletter 😊