CLARIN

Common Language Resources and Technology Infrastructure

DAM-LR (CLARIN)

Initiative for a pan-European Research Infrastructure

Peter Wittenburg
MPI for Psycholinguistics

DAM-LR Meeting + CLARIN

- DAM-LR a small EC funded project is finishing
- goal: virtual integration of LR archives/repositories
- partners: MPI Nijmegen, INL Leiden, U Lund, SOAS London
- have achieved what we promised
- will report on details at this meeting

- turns out that DAM-LR infrastructure topics are basic layer for CLARIN
- therefore open discussion about CLARIN aspects and ideas
- not a formal CLARIN meeting but very good moment for brainstorming

- thanks for coming!!

- day 1: more general
- day 2: more technology detail

CLARIN

Common Language Resources and Technology Infrastructure

CLARIN Introduction

Initiative for a pan-European Research Infrastructure

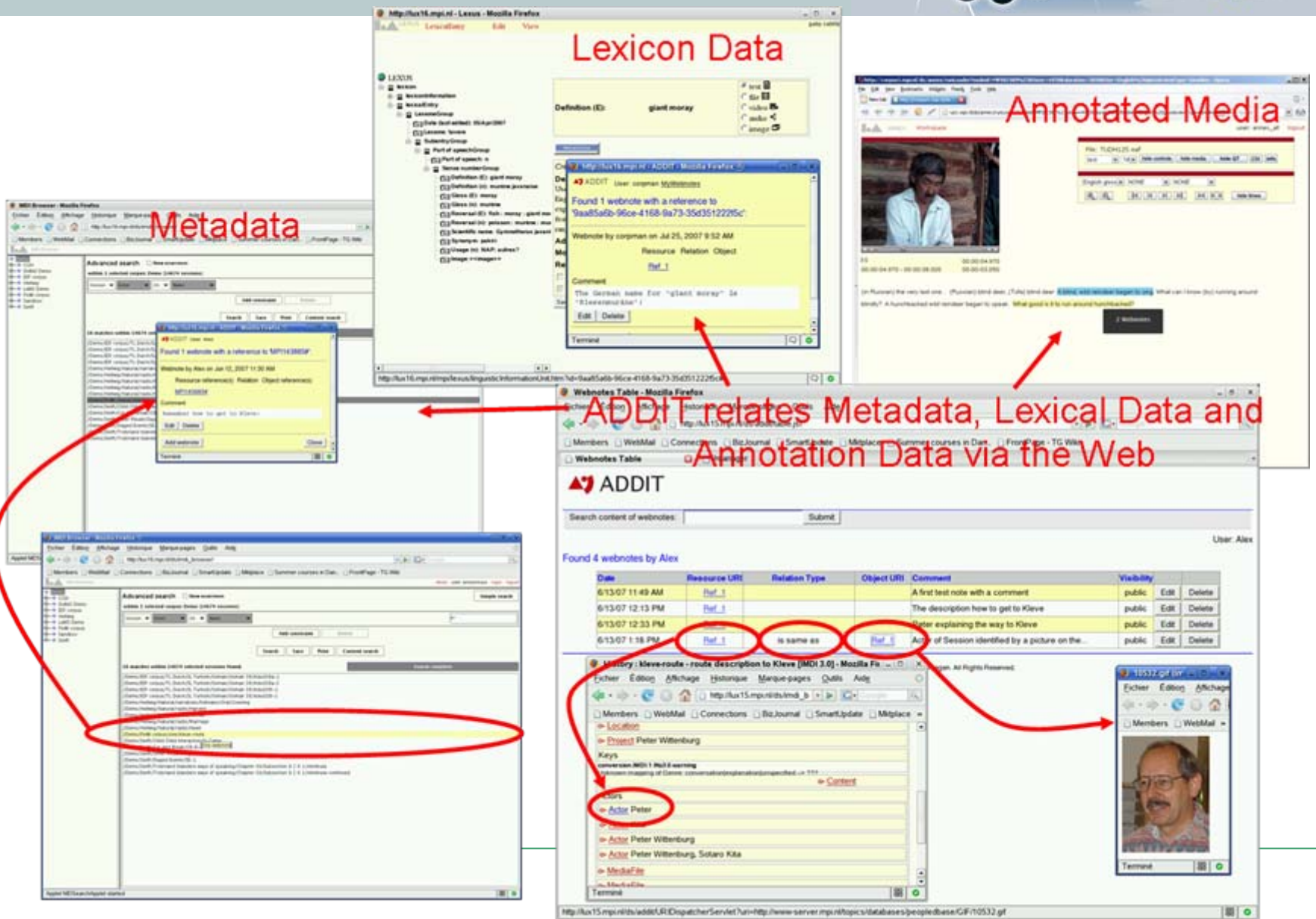
Current LR&T Situation

- Example 1:
  - Ethnologists have a recording and a transcription, they want to search for certain textual patterns and then return to the recorded dance fragments
  - to do this and to save time they would need to align media and transcriptions
  - there are “aligner tools”
  - but who is able to use them and will they work on the transcription format

- Example 2:
  - Historians want to access all material from physics, politics and sociology to understand the reasons for the marine dominance of the Venetian Republic
  - to do this they need to search for concepts in all material, extract summaries, relate fragments, add and exchange comments etc
  - they need to do this collaboratively
  - currently this involves a huge amount of handwork to overcome institutional, linguistic (morphological normalization, translation), semantic boundaries
  - but who is able to carry out such work, who can operate the tools

Can we improve the situation (stepwise)?

Example


Lexicon Data
Annotated Media
Metadata
ADDIT relates Metadata, Lexical Data and Annotation Data via the Web
ADDIT
Found 4 webnotes by Alex

Current LR&T Situation

- the amount of Language Resources and Technology components is growing extremely fast and including all media (texts, sound, video)
- digital era makes LR&T ***in principle*** easily accessible and enables data driven research
- but most researchers in humanities and even beyond are excluded from this data driven research
- LR&T is effectively only usable and visible by a few experts
    - LR&T is fragmented – it is scattered at various locations, it is available in many different formats and terminologies
    - LR&T is widely invisible since it is neither registered nor described, therefore it practically does not exist except for the experts
    - interdisciplinary work is still much too difficult to implement

Mission

CLARIN

- is committed to establishing an integrated and interoperable RI supporting easy access and use of Language Resources and Technology
- aims to overcome the current fragmentation and offer a stable, persistent and extendable infrastructure
- it will offer its services to researchers and scholars across a wide spectrum of domains in particular in the humanities and soc sciences
- knows that much education and training will be necessary to meet the goals and to get the researchers involved
- expects that altogether the CLARIN RI will boost humanities research in a multicultural and multilingual Europe (world)
- is aware that seamless access to very large collections and a sequence of tools will be required to tackle the complex questions about minds, societies, health and nature we will be faced with in future

The eScience Vision

J. Taylor

“eScience is about global collaboration in key areas of science and the next generation of infrastructures that will enable it”

CLARIN is establishing such a new generation of extended infrastructure.

Thus CLARIN is not about creating and building new language resources and technology, but making them available and accessible as services in a stable and persistent infrastructure to allow tackling the great challenges.

Part of a Large Game

- CLARIN is part of a larger game covering many disciplines
- at the EU level we have
  - the DRIVER initiative to harvest discipline metadata
  - 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, ...)
  - who knows what else ??
- at the international level we have
  - many attempts to collaborate at the infrastructure level
  - examples: persistent and unique identifiers, federation attributes etc
  - work in ISO TC37/SC4
- my conviction: we need a very sensitive balance between bottom-up approaches driven by the communities and top-down approaches

CLARIN Community

- Currently CLARIN
  - has about 90 member institutions from 31 European
  - includes most of the well-known researchers and technologists from our field
  - has commitment statements from 25 member states (growing)
  - two tier organizational structure
    - European level with EC support
    - national networks of CLARIN members (national coordinator)
  - members' voice is heard
    - annual plenary meetings foreseen and open working groups
  - has received statement of interests from various non-EU countries (Japan, Korea, China, US, South Africa, Australia, Argentina, Brazil, Russia)

EC funded Consortium

- some principles formulated by the EC
  - funds only for preparatory phase for a limited set of institutions qualified to do the preparatory work
  - letter of support from National funding authorities
  - attested competence/resources deemed necessary for the preparatory phase
- CLARIN principles
  - every national group with national funding backing should be represented to represent languages
  - in 3 years we will see who the strong contributors are
  - open to new countries/partners to join (contract restrictions)
- total EC budget support of 4.1 Mio € for 3 years prep phase
- 32 of the members from 22 states are consortium partners
  - concrete negotiations in several members states about national programs (in D we expect a budget of 3 Mio € for 11 members)

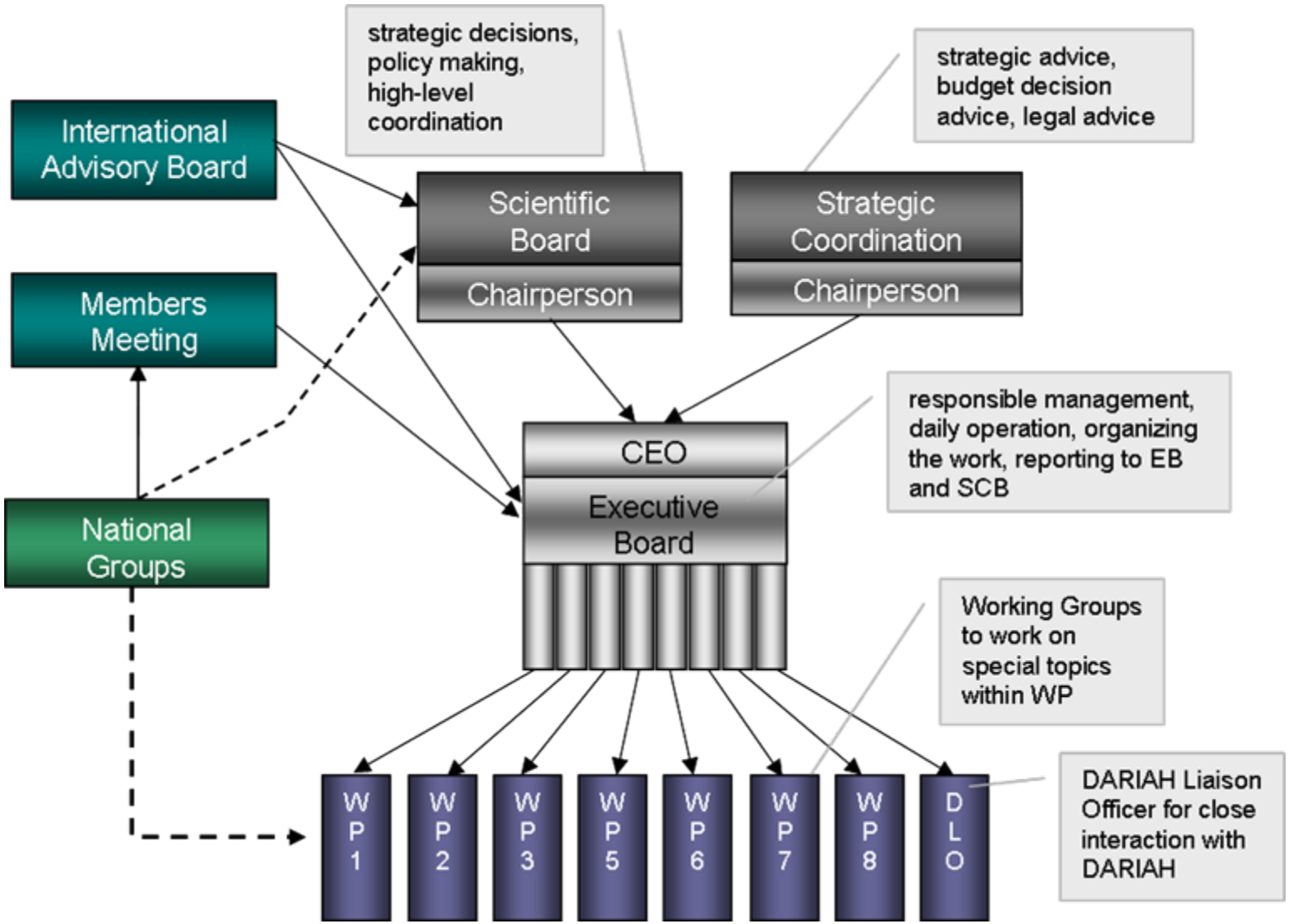
CLARIN Work Packages

WP1	Management & Coordination	Steven Krauwer, Utrecht, NL
WP2	Technical Infrastructure	Peter Wittenburg, Nijmegen, NL
WP3	Humanities Overview	Tamas Varadi, Budapest, Hun
DLO	DARIAH Liaison Officer	Martin Wynne, Oxford, UK
WP5	Language R&T Overview	Erhard Hinrichs, Tübingen, Ger
WP6	Dissemination	Dan Cristea, Iasi, Rom
WP7	IPR and Business Models	Kimmo Koskiennemi, Helsinki, Fin
WP8	Organizational Agreement	Bente Maegaard, Copenhagen, Dk

- WP4 was devoted to reach out to the humanities community – cut by EC
- still central in CLARIN: the LRT community should offer services to humanities etc
- WP3 with concrete humanities projects very important as well as the link to the DARIAH initiative (represented by Laurent Romary, Peter Doorn, Sheila Anderson)

- work will be organized in Working Groups
- Working Groups are open to experts from everywhere

CLARIN Organization

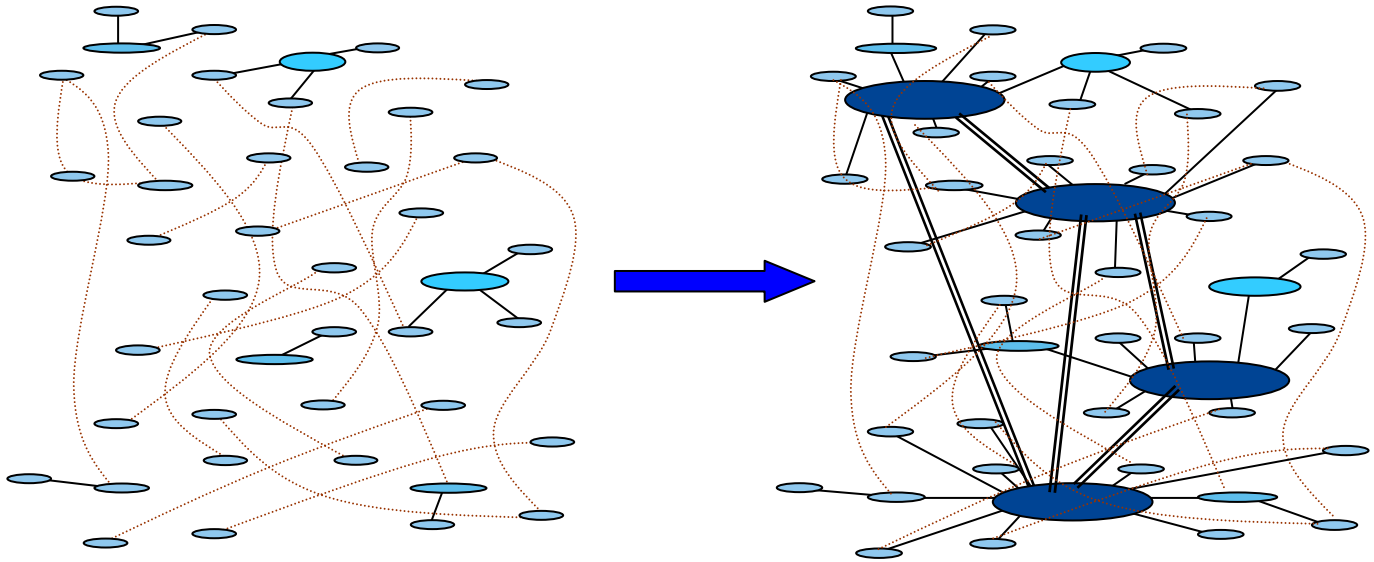


WP2/5 Overview/Standards/...

- we need a comprehensive overview about LRT, their characteristics and their state – much analysis and selection work
- we need a comprehensive taxonomy of all LRT
- we need a new flexible metadata schema covering all LRT
- we need generic standards for resource structures
  one example is certainly Lexical Markup Framework
- we need stable services for terminology and relations amongst them
- we need to understand architectural designs of the tools
- we need to work out typical workflow chains to design the SOA
- we need to identify gaps in LRT
- what about quality control – very difficult if not impossible

- this is all a lot of non-trivial work ☺
- have to start now

WP2 Strong Centers



- 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 offer a true service mentality and a new form of openness and technical accessibility
- need a strong national support for many years
- resource centers need to have a suitable repository/archiving system

WP2 Type of Centers

- LR Centers
  - uploading and integration of new resources/versions
  - long-term data preservation
  - curation of LR
  - allow access to LR
  - all based on proper repository/archiving systems

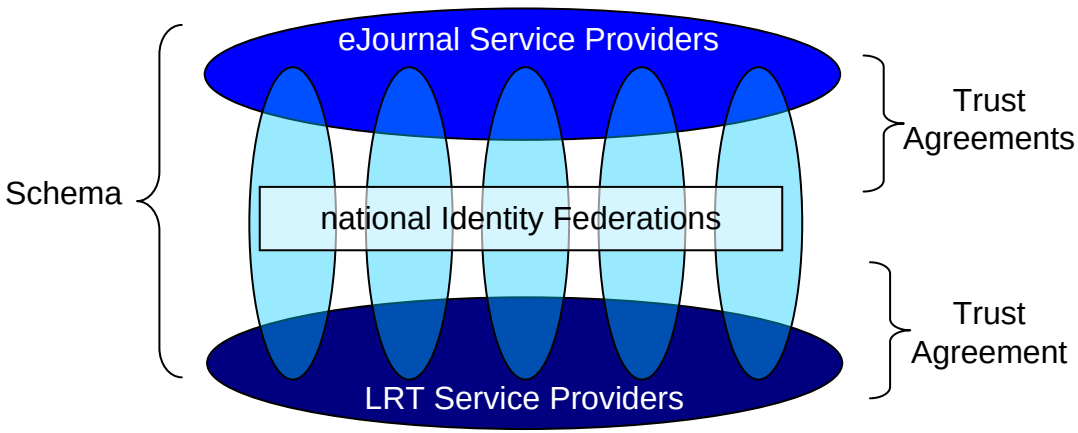
- LT Centers
  - offer web services to execute LT
  - offer advice

- Infrastructure Centers
  - ISOcat service
  - URID service
  - MD service
  - registries

- Advisory Centers

WP2 LRT Federation

- CLARIN needs to build a federation based on simplified and unified rules for licensing, accessing, user authentication etc
- joint metadata registry for resources and tools based on long experience
- support for virtual collections with resources from different archives
- unique way of referencing electronic resources in federation
- single sign-on/identity principles in federation
- trust agreements with national identity federation



WP2 Type of Federation

- architecture for the federation (agreement with national structures, MPI, ...)
- which are the attributes and how to use them (mostly def by EduPerson etc)
- what are the trust agreements in federation – also a matter of licences
- criteria for centers to participate in federation (strong enough, nat. support, ...)
- check the situation of all applicants
- install and integrate Shibboleth
- maintain WAYF
- make agreements with national identity federations

- grid integration -> Daan

WP2 LRT Registry

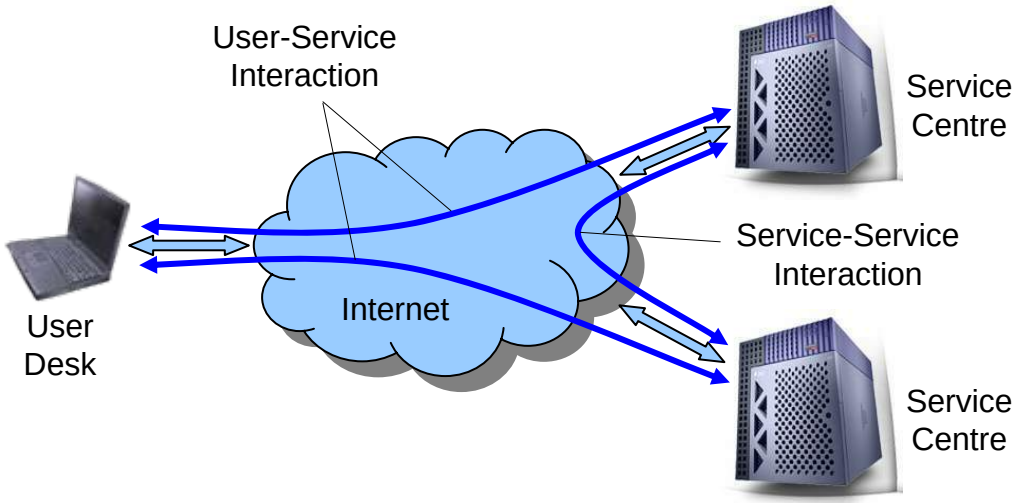
- joint metadata registry for resources and tools based on long experience
  - is it ONE registry?
  - how to describe tools?
  - how to support national maintenance – hierarchical architecture?

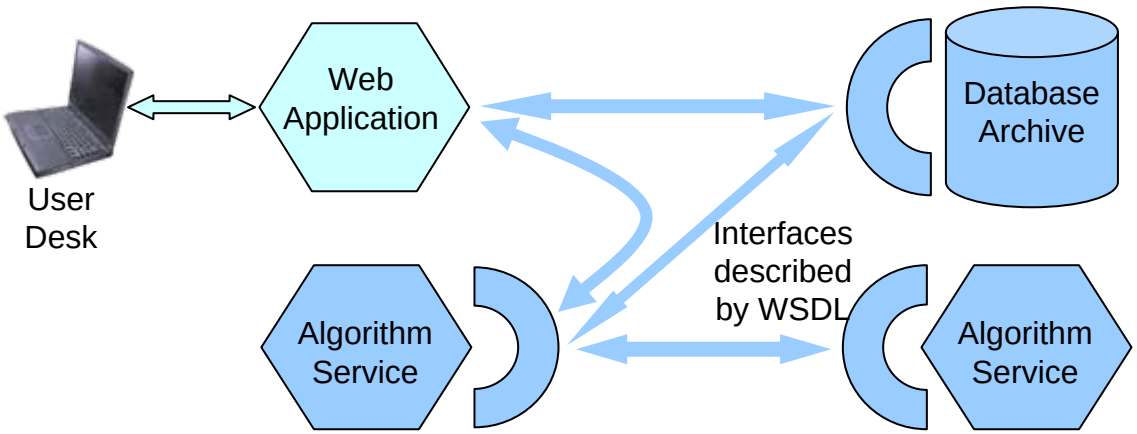
- new modular and more flexible metadata schema to be worked out based on a comprehensive taxonomy of LRT
  - important is everyone can make his own schema
  - everyone can use localized and sub-discipline vocabulary
  - re-use of existing categories registered in public ISODCR
  - pointer to interface / resource

- need a new infrastructure (definitions, registry, tools, geo browsing, gateways, ...)
- support for virtual collections with resources from different archives
- LT-World integration

WP2 Service Oriented Arch

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 only architecture scalable and flexible enough
- standardization and harmonization is required

WP2 Service Oriented Arch

- description of the problem
  - there are things such as streaming services etc
  - what are the needs for WSDL
  - support for REST services
  - what is the I/E problem

- have already some services (LMF API, DCR API, GATE, RACAI, ...)
- all for machines and for humans

- study of some typical workflow chains
  - take a text – execute tokenizer – execute POS - ...
  - LREP survival
  - ...

- first ideas about workflow language

WP2 Basic Services

- domain wide searches
  - metadata
  - content (which architecture?, which rights?, ...)
  - combined

- LREP service

- needs to evolve during the project

Risks and challenges

- Scale of effort
  - danger of fragmentation
  - difficult to coordinate various interests (EU – national)

- Outreach to the SSH research community
  - extensive work needed to generate critical mass in interest
  - strong collaboration with DARIAH envisaged

- National contribution - the hidden dimension
  - many partners have only a nominal share in EC support, hence success depends already in the preparatory phase on national support
  - encouraging signs from negotiation with funders from some countries
  - but also readjustments of national commitment to match EC support
- Continued support by ESFRI to influence national funders necessary

Thanks to ESFRI + EC

- W. Krull (Gen. Sec. VolkswagenFoundation) described the pre-conditions for a creative culture,
  - new paths require long-time scales
  - funders' involvement should be based on trust and long term commitment
  - if we can establish trust, we (the funders) can be successful

- for the research infrastructures we have in mind we know that we will not do the construction job in 3 years – not to speak about long-term availability

Therefore, we need to thank those bright minds behind ESFRI who figured out a few years ago that we need to build new infrastructures and we need to thank the EC to be willing to get this from the ground!
We need long-term national funding schemes.

End

Thanks for your attention.

some URLs:

CLARIN:	http://www.clarin.eu
Grid Project:	http://www.dam-lr.eu
ISO TC37/SC4:	http://www.tc37sc4.org
Standards Project:	http://lirics.loria.fr/

Collaboration

- the LRT community is not limited to Europe – it's international
- we want to collaborate with all interested and committed experts
- need to understand requirements from different languages
- need your contributions to overviews, standards, etc
- need to integrate non-European centers and their LRT

Concrete

1. Would like to invite your experts to participate in working groups of interest
2. Would like to invite you to participate in the establishment of a network of strong service centers
3. Would like to invite you to join the LRT Federation

Need to work on a funding scheme supported by the EC and the national governments (ours and yours).

Robert-Jan Smits (EC) stated clearly: collaboration is wanted!

CLARIN Work Packages