CLARIN
Common Language Resources and Technology Infrastructure

CLARIN: Goals and Structure of the Project

Erhard Hinrichs
University of Tübingen
eh@sfs.uni-tuebingen.de

Workshop on Multilinguality in Information Access to Digital Libraries –
User Needs and Evaluation of Multilingual Resources Use

Trento, Italy
2009-09-09

The CLARIN Mission

What:

- Create an infrastructure that makes language resources and technology (LRT), available to scholars of all disciplines, especially social sciences and humanities (SSH)

How:

- Unite existing digital archives into a federation of archives with unified web access
- Provide language and speech technology tools as web services operating on language data in archives

Why a European Infrastructure?

- too much fragmentation
- lack of coordination across countries
- lack of visibility
- lack of interoperability
- lack of sustainability

- expertise exists but not in all countries
- language independent tools can be shared
- language dependent tools can often be ported
- most countries not able to bear the cost

Why now?

- Exponential growth of digital data
- Increasing maturity of language and speech technology:
  - high speed
  - large volumes
  - new research questions

- Growing interest at EU level in research infrastructures (RI)
- ESFRI Roadmap published in 2006
- includes 35 accepted proposals for RIs
- CLARIN is one of them
- all of them will get funding for a 1-3 year preparatory phase

Who we are and where we come from

- The CLARIN consortium currently has 32 partners from 22 EU and associated countries (and more on the waiting list)
- The CLARIN community has 171 members in 32 countries (September 2009)

Overall plan for CLARIN

Preparatory phase:

- 2008-2010
- Put everything in place

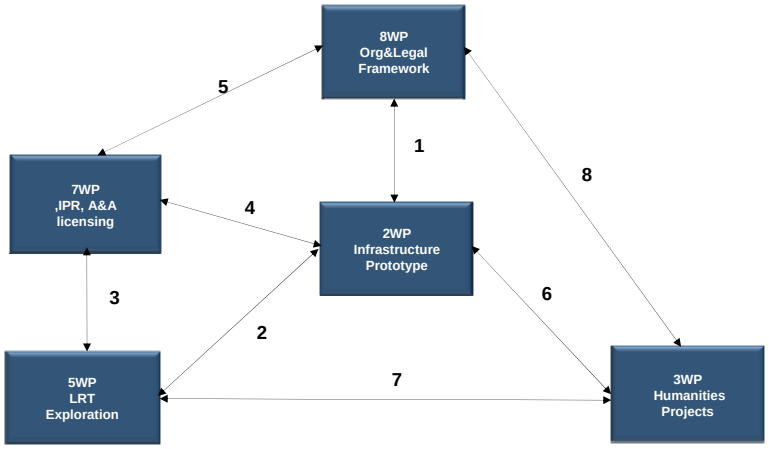
Construction phase:

- 2011-2015
- Build and populate with tools and resources

Exploitation phase:

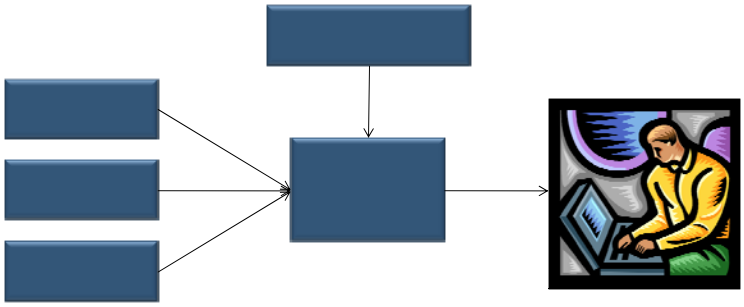
- 2016-….
- CLARIN in full service

How we work



WebLicht: Web-based LRT services for German

- WebLicht: User interface as part of a Service Oriented Architecture (SOA) for Language Resources and Tools (LRT)
- Interacts with the user and combines *distributed services* and metadata from a centralized *repository*

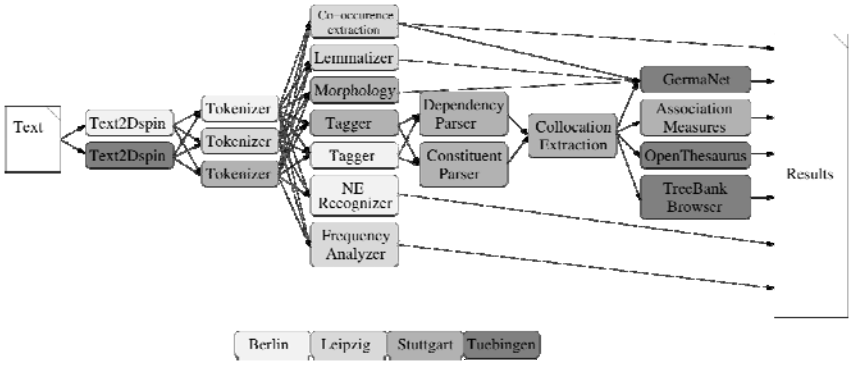


WebLicht: Web-based LRT services for German

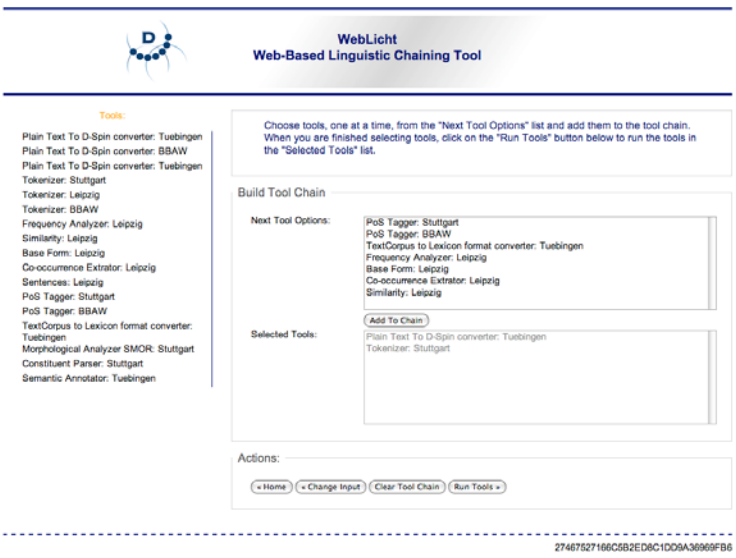
- Services are implemented as REST style webservices
- They are offering a wide range of linguistic tools and aplications:
  - Tokenizer
  - Part-Of-Speech Tagger
  - Parser
  - Semantic Annotators
  - Converters from format into another
  - …

WebLicht: Web-based LRT services for German

Trento, 2009-09-08
www.clarin.eu



WebLicht: Web-based LRT services for German



The D-Spin Dataformats

- Developed by Helmut Schmid (Stuttgart) and Volker Boehlke (Leipzig)
- In general: Stand-off formats, different annotation layers are stored in one file
- Several variations (TextCorpus, Lexicon and Metadata)
- WelbLicht makes use of TextCorpus format (TCF)

The D-Spin Dataformats

- Tries to be compatible with established standards, especially dataformats of the ISO/TC 37-SC4 group:
  - LAF: Linguistic Annotation Framework
  - LMF: Lexical Markup Framework
  - MAF: Morpho-Syntactic Annotation Framework
- At the moment, converters are available for:
  - PAULA
  - Negra
  - TüBa-D/Z

More Info

- CLARIN Website: http://www.clarin.eu
- CLARIN Office: clarin@clarin.eu
- CLARIN Newsletter:
  http://www.clarin.eu/newsletter
- CLARIN Members:
  http://www.clarin.eu/members
- D-Spin Homepage
  http://www.sfs.uni-tuebingen.de/dspin

CLARIN
Common Language Resources and Technology Infrastructure

Thank you for your attention

CLARIN has received funding from
the European Community's Seventh Framework Programme
under grant agreement n° 212230

The Problem

- Much data in digital archives language based
- Only known to insiders
- Archives mostly unconnected
- Every archive has its own standards for storage and access
- Normally only simple retrieval of files (text, audio or video documents) Social sciences and humanities researchers are not language or speech technologists
- 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-