



# **CLARIN: Language and Speech Infrastructure for Researchers in the Humanities and the Social Sciences**

---

Ineke Schuurman  
CCL, K.U.Leuven  
Co-ordinator CLARIN-Vlaanderen

## **Survey**

---



1. Objectives
2. Examples
3. CLARIN-EU
4. Research Infrastructure
5. Technological Infrastructure
6. Support Infrastructure
7. Consortium CLARIN-Vlaanderen
8. CLARIN-Vlaanderen
9. CLARIN-NL and CLARIN-Vlaanderen
10. A bright future for HSS researchers

## 1. Objectives

---



In the Humanities and the Social Sciences (HSS) much of the research is language based.

- Linguistics
- Study of literature
- History
- Philosophy
- Theology
- Communication sciences
- Branches of sociology, law and psychology

For the HSS scholar the universe of texts and recorded speech is what the cosmos is for the astronomer and the natural world for the biologist.

## Objectives

---



- Biology has benefitted from the microscope.
- Astronomy has benefitted from the telescope.
- HSS could benefit from the tools and resources which have been developed in the fields of language and speech technology (LST).
- BUT: the LST products are insufficiently available to HSS researchers and HSS researchers are insufficiently prepared to make good use of them.
- ERGO: the great potential of LST for HSS is not yet realized.

## Objectives

---



- The pan-European CLARIN program aims to set up and maintain a persistent **research infrastructure** allowing HSS researchers to use state-of-the-art LST products and resources.
- Special attention for user-friendliness: easy to use, also for people with limited knowledge of and interest in technology.
- Bridging the gap between HSS research demands and LST provisions.
- Opening new perspectives for research in the Humanities and the Social Sciences.

## 2. Examples

---



- Literature
  - Is an ancient text written by one person or by a collective of different persons?
- Psychology
  - What determines our reading strategy of subtitles? Is there a link with the type of document, or with the purpose of subtitling (translation, vs. hard-of-hearing)?
- History
  - What is the recent political history of state reforms in Belgium? Next to written documents, there are many non-transcribed radio and television interviews with politicians. How can the relevant interviews be traced?

### 3. CLARIN-EU

---



- all languages spoken and/or studied in the European countries are involved in the CLARIN-project (+/- 100)
- per country
  1. its official language(s), plus
  2. languages outside Europe (Inuit languages, native-American languages, African languages, extinct languages, ...)
- 156 members, of which 33 actively participate in the European part of the preparatory phase (2008-2010)
- Construction phase (2011-2015)
- Exploitation phase (from 2015)

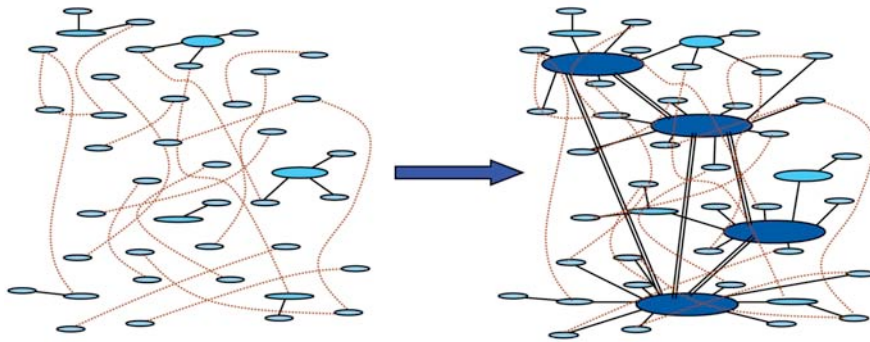
### 4. Research Infrastructure

---



- The CLARIN infrastructure is not so much a piece of hardware or a machine, but rather a repository of LST resources and tools.
- The transformation of current LST repositories into resources and tools for doing HSS research.
- Because of the (language) specificity of the resources and tools we need intensive collaboration across language and state borders.
- From this follows the necessity to adhere to common standards and uniform practices.
- Special effort needed to make the infrastructure easy to use and access.

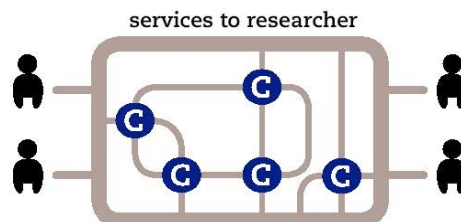
## 5. Technological infrastructure



## HSS Researcher



User is only confronted with the 'outside', the internal organisation of the RI will be hidden

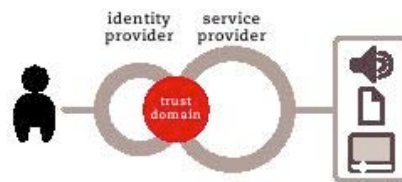


## CLARIN: trust domain



- Every user will have one 'identity', to be used again and again (e.g. provided by the university)
- Rights and duties of a user are laid down between identity provider and service provider
- User may compose virtual collections (details stored automatically upon request)

→ User is not hampered by administrative details

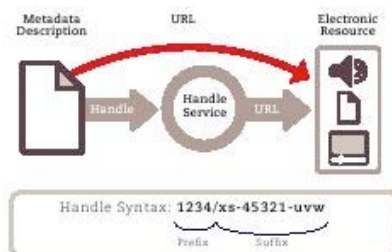


## CLARIN: Persistent Identifier service



- Essential: stable references
  - For user (in paper)
  - For CLARIN (internal use)
- Classical URLs not stable
- CLARIN: user refers to PID. This PID is used by the PID-service (e.g. handle-service) to establish a link.

- At the intermediate level of PID-service changes of locations are stored



## CLARIN: concept registry service

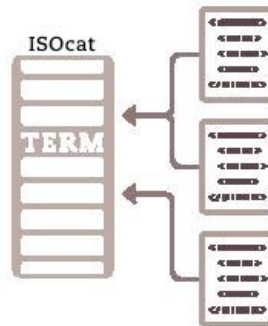


Issue: everybody uses (linguistic) concepts based on mother tongue, discipline, theoretical background

In CLARIN, these concepts are linked to standardised ones

Per user, preferred readings will be stored

CRS: based on ISOcat, TC3/SC4, ISO 12620

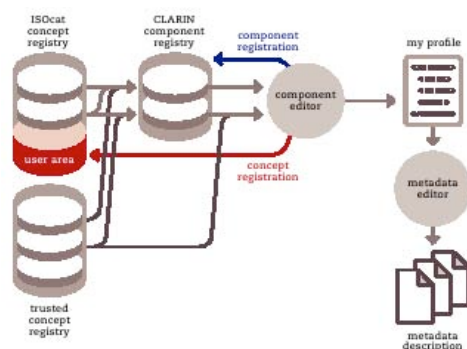


## CLARIN: Component Metadata



- Tools and resources need to be described such that both humans and machines can locate them: metadata

- Upon request, a user profile can be stored for future use

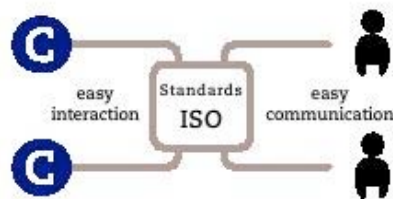


## CLARIN: standards, best practice



CLARIN uses standards and 'best practice' in all domains

→ guiding principle for new resources and technology or adaptation of old ones

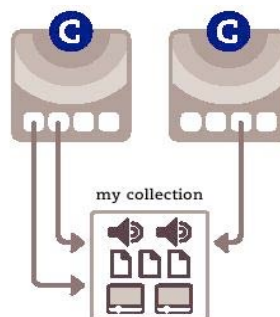


New standards will come up in the course of time: adaptation of resources while keeping older versions

## CLARIN: virtual collection



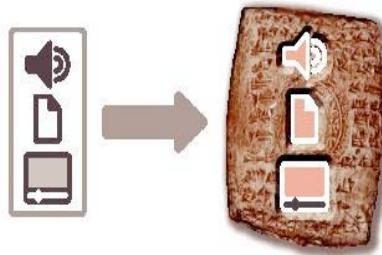
- 'virtual collections' (with components originating from several locations) can be consolidated, providing them with an identity (metadata, pid)
- This can be done for an individual user, or a group of users (like the students of professor X)



## CLARIN: long term preservation



- Tools: limited lifespan
  - Resources: should remain available:
    - Preservation of cultures and languages
    - ‘reproducibility’ of research
- Two aspects:
- content (*Eindhoven corpus!*)
  - ‘format’ (storage media)
- Essential aspect of CLARIN



## 6. Support Infrastructure



- Help desk for personal and/or project dependent advice
- A wiki site addressing more general questions
- Hands-on training sessions for HSS researchers
- Course material for HSS students
- Demonstrators to be presented to the HSS community

Largely tasks for the national groups!

## 7. CLARIN-Vlaanderen - consortium

---



All members of CLIF:

- Katholieke Universiteit Leuven
  - Centre for Computational Linguistics (coordinator, partner in CLARIN-EU)
  - ESAT-PSI : image and speech processing
  - Interdisciplinary research on Technology, Education and Communication
  - Language Intelligence & Information Retrieval
- Universiteit Antwerpen
  - Computational Linguistics & Psycholinguistics
- Vrije Universiteit Brussel
  - ETRO-DSSP: digital speech and audio processing
- Universiteit Gent
  - ELIS: digital speech processing
- Hogeschool Gent
  - Language and Translation Technology Team

## 8. CLARIN-Vlaanderen

---



- Financed by EWI (Departement Economie, Wetenschap en Innovatie)

European level: partner (rather abstract level: how to make tools and resources CLARIN-compatible all over Europe)

Flemish level:

- Arousing awareness amongst Flemish HSS-researchers
- Inventarisation of their LST needs
- Finding ways to help them

## CLARIN-Vlaanderen

---



More specific wrt last point :

- Adapting the architecture of the Flemish LST repositories to fit into the CLARIN federation
- Adapting the current tools and resources contained in the repositories to make them CLARIN compliant and suitable for HSS research
- Adding new tools and resources to meet the needs of HSS in terms of coverage (for instance, lemmatizers or taggers for older variants of the language)
- Special attention for Dutch and for African languages

## 9. CLARIN-NL & CLARIN-Vlaanderen

---



- CLARIN-Vlaanderen and CLARIN-NL closely collaborate in the preparatory phase (2008-2010) in order to avoid duplication of efforts.
- Flemish-Dutch pilot infrastructure (2009-2011):  
*TST-Tools voor het Nederlands als Webservices in een Workflow* (proposal submitted in both Flanders (EWI) and the Netherlands (CLARIN-NL))
- Point of departure: real needs of HSS researchers (Sagalassos, Huygens Instituut; KADOC, M2P, KDC, Aletta)
- Goal: to create the environment they need for their research, largely based on existing LST resources

## CLARIN-NL & CLARIN-Vlaanderen



- TTNWW covers both language and speech
- In both collaboration of Flemish and Dutch researchers

### Furthermore

- 10 Dutch projects (demonstrators etc)
- 3 Flemish projects, dealing with new tools and resources for HSS researchers

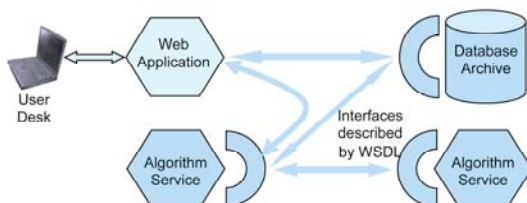
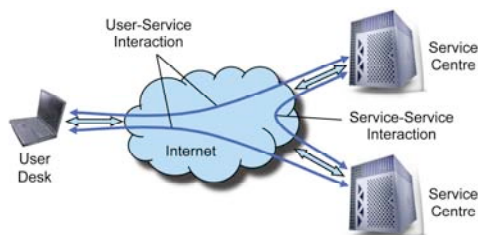
All these projects should start in 2010.

## 10. A bright future for HSS researchers



### 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



**THANK YOU !**

**Questions ?**

<http://www.ccl.kuleuven.be/CLARIN>

**ineke @ccl.kuleuven.be**