CLARIN-PLUS

www.clarin.eu

Franciska de Jong

f.m.g.dejong@uu.nl

Leuven

19-21 September 2016

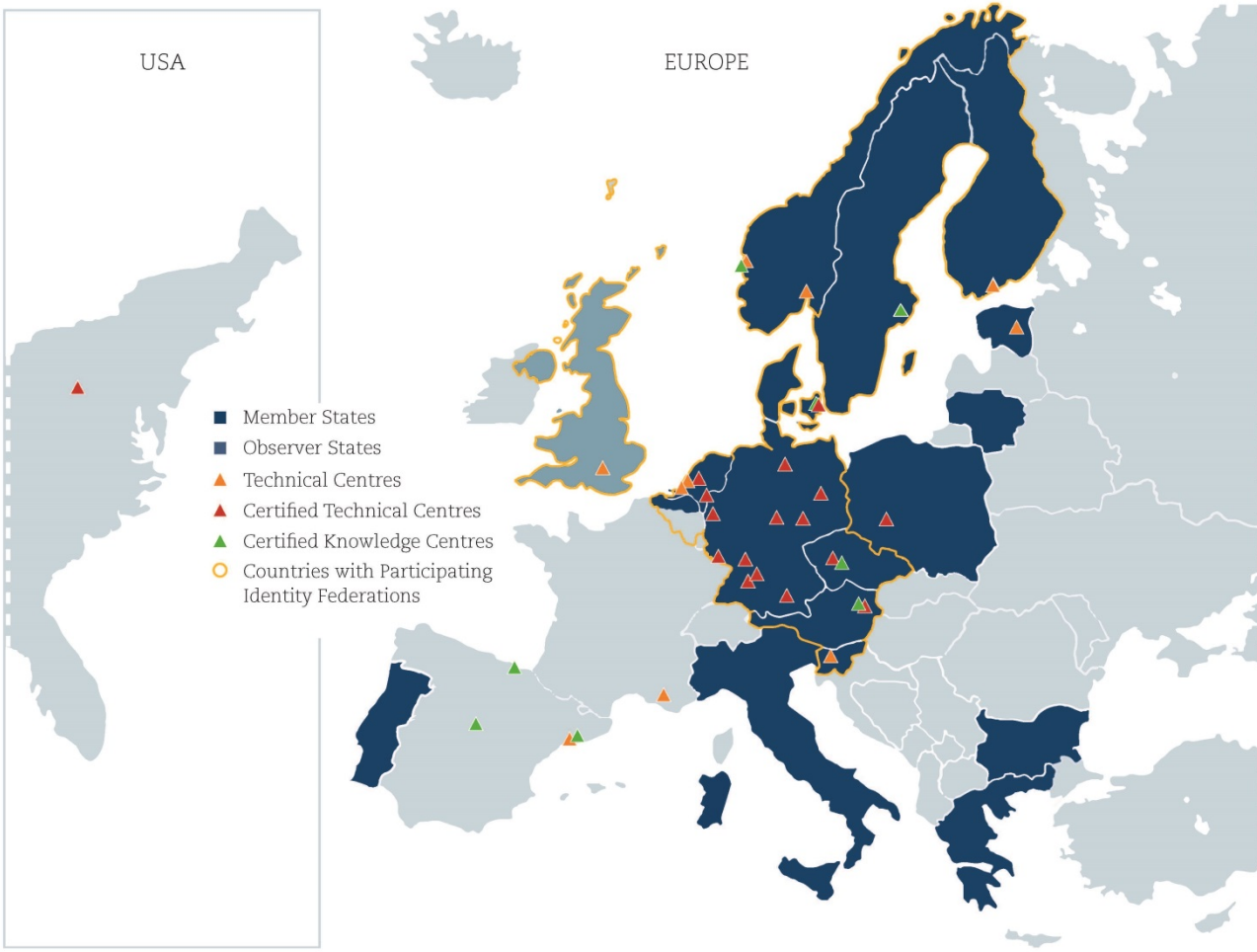
CLARIN in four bullets

- CLARIN is the Common Language Resources and Technology Infrastructure
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form),
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located.

CLARIN ERIC in members and centres

A consortium of countries:

- 19 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, **HU**, IT, LT, **LV**, NL, NO, PL, PT, SE, SI
- 1 observer: UK



CLARIN and data science

- Analytics for text and speech data as a pillar for data science
- Contribution to the development of new methodological frameworks for the integrated processing of multiple datatypes and multidisciplinary research agendas.
- Europe's mulitlinguality as a basis for comparative research of societal phenomena, and in particular those that are reflected in language use:
  - Migration patterns
  - Intellectual history
  - Language variation
  - ….
- Text and speech as data

CLARIN and data science

- Analytics for text and speech data as a pillar for data science
- Contribution to the development of new methodological frameworks for the integrated processing of multiple datatypes and multidisciplinary research agendas.
- Europe's mulitlinguality as a basis for comparative research of societal phenomena, and in particular those that are reflected in language use:
  - Migration patterns
  - Intellectual history
  - Language variation
  - ….
- Text and speech as **social** and **cultural data**

CLARIN in data types

- Parliamentary records
- Literary texts
- Social Media data
- Historical letters
- Oral History data
- Disciplinary libraries
- Institutional archival data
- Broadcast archives
- ….
- Newspaper archives

Prehistory of this workshop

- CLARIN-PLUS: outreach to new users, focus on four specific data types
  - oral history collections
  - newspaper archives
  - parliamentary records
  - social media data
- Collaboration in proposals, research collaboration, etc
  - R&D proposals (FP7)
  - Europeana
  - International projects
  - National initiatives

Newspaper archives as data

- *Aim for this workshop*:
  exploring existing and envisoned approaches for analyzing newspaper archives with the use of CLARIN-compatible standards and processing tools.

- *Long-term vision*:
  The CLARIN infrastructure provides easy access to newspaper archives and services suited for this type data and encourages researchers to develop and address discipline-specific hypotheses and scholarly questions.

User needs: determined by task

Data curation

- OCR
- Metadata creation

Exploration of the data space

- Finding articles on a specific topic, place, person, …
- Finding threads from a specific ttile, period,

Analysis

- Annotation (named entity detection, ….)
- Text mining
- Link generation

Presentation

- Citation of articles
- (Re)combining text and images
- Visualization of patterns

Challenges and multidisciplinary potential

Newspaper archives are considered a rich data type that

- is suited for both *close reading* and *distance reading*
- is often presenting itself as messy or noisy data
- is calling for links with data in other modalities than text

Newspaper data sets have a huge potential for reuse and re-purposing within many fields of study in the humanities and social sciences (and beyond):

Humanities: history, language change, …

Social sciences: social and cultural dynamics, political sciences, economics, …

Lessons learned

- User ambitions tend to be conservative, so ….
  a bit of technology push can be good, but ..
- .. the functionality that tools have to offer should support users in the workflows they know, rather than steer the exploration of data or the application of tools in ways that are not understood, so …
  user needs should be kept in focus.
- Scholarly insights and conlusions without modes for validating/replicating the results have difficulty to gain trust , so …
  black boxes are have little added value
- For collaboration across disciplinary boundaries, communication pitfalls will never stop to exist, so …
  keep talking after this workshop!