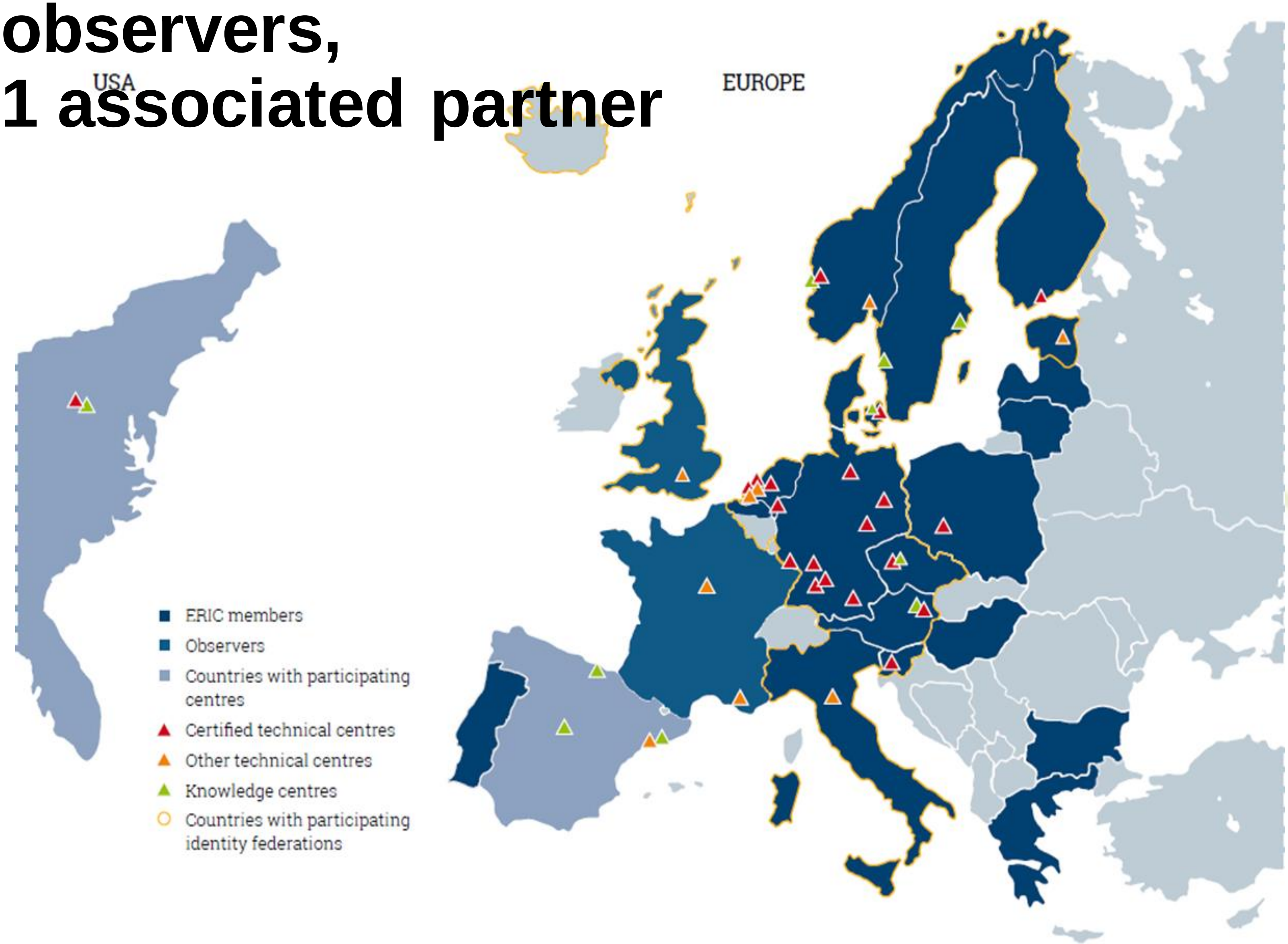
Franciska de Jong
executive director CLARIN ERIC
f.m.g.dejong@uu.nl

Sofia, Bulgaria
27-29 March 2017

CLARIN in five 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
- through a **single sign-on** online environment.

CLARIN ERIC: 19 members, 2 observers, 1 associated partner


USA
EUROPE
■ ERIC members
■ Observers
■ Countries with participating centres
▲ Certified technical centres
▲ Other technical centres
▲ Knowledge centres
○ Countries with participating identity federations

CLARIN in data types

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

CLARIN and data science (1)

- Analytics for text and speech data: **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 **multilinguality** as a basis for **comparative research** of societal and cultural phenomena, and in particular those that are reflected in language use

CLARIN and data science (2)

- Examples of research topics
  - Intellectual history
  - Migration patterns
  - Language variation across period and region
  - Dynamics in mental health conditions
  - Parliamentary discourse
- Text and speech as **social** and **cultural data**
- From tools for the study of **lexical units** …

…to analysis tools for **big data**

Prehistory of this workshop

- H2020 project CLARIN-PLUS:
  outreach to new users, focus on four specific data types
  - oral history collections
  - newspaper archives
  - parliamentary records
  - social media data
- Joint proposals and research collaboration
  - national initiatives
  - R&D proposals (FP7, H2020, COST)
  - international projects

Parliamentary records as data

Long-term vision:

The CLARIN infrastructure provides easy access to paliamentary materials, services suited for this type data can easily be found and employed, and it encourages researchers to develop and address discipline-specific hypotheses and scholarly questions.

Aims for this workshop:

- exploring existing and envisoned approaches for analyzing parliamentary records (tekst, speech) with the use of CLARIN-compatible standards and processing tools
- generation of an overview of relevant resources to stimulate synergy and cross-country collaboration
- creation of an action plan

Challenges and multidisciplinary potential

Parliamentary data sets 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 tekst and speech
- is created under specific circumstances that need to be well understood before strong conclusions can be drawn

Parliamentary 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, discourse analysis, …

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 and/or replicating the results have difficulty to gain trust , so …
  black boxes have little added value
- For collaboration across disciplinary boundaries, communication pitfalls will never stop to exist, so …
  keep talking after this workshop!

CLARIN:

Infrastructural support

for the study and use of

language as social and cultural data