Driven by Data, Tied by Uncertainty : the Need for Coherent RI Policies

Franciska de Jong

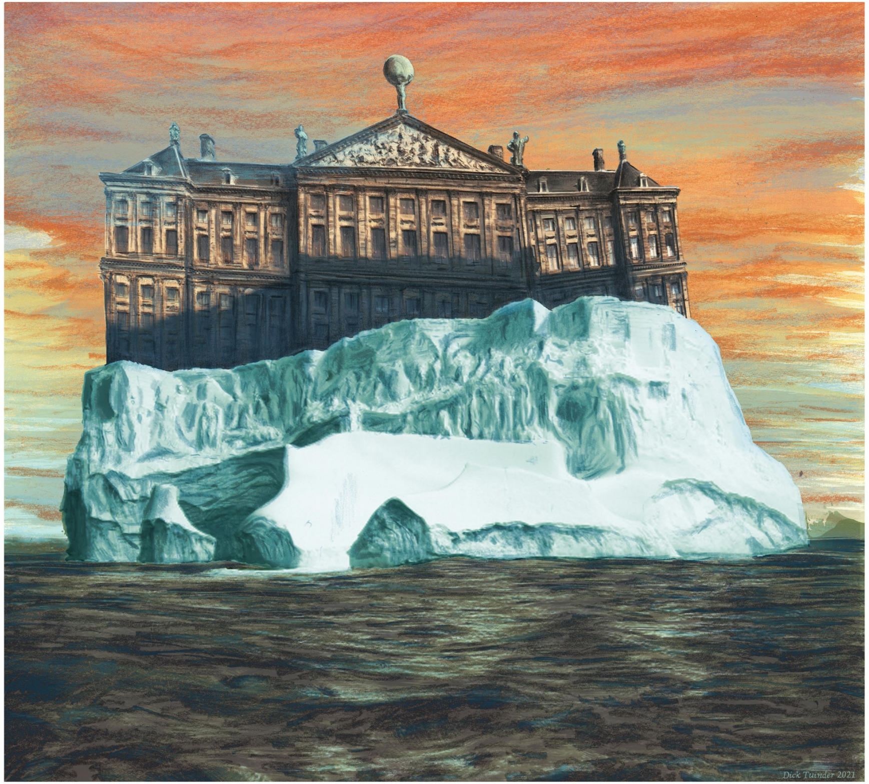
CLARIN ERIC

f.m.g.dejong@uu.nl

e-IRG workshop, Portugal, Session III

Going beyond: How policies can shape the development of research infrastructure facilitating data sovereignty, innovation and cultural change

26 May 2021



Artist impression of the future of the Royal Palace in Amsterdam

Dirk Tuinder 2021

Data in transition

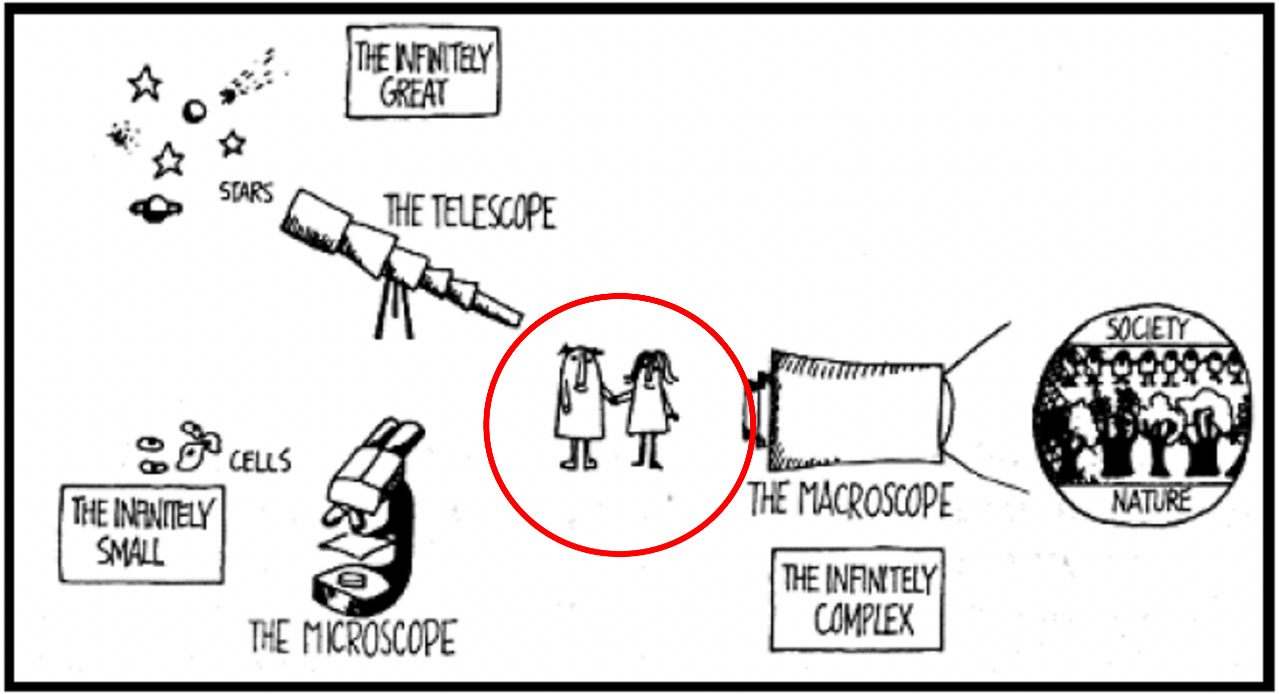
From

- something to be cherised and protected
- something to be explored based on a research question
- data collection as step 1
- something to be collected, closely studied and then stored away

To

- something to be shared
- something that can suggest research questions
- reuse of pre-existing data
- something that is deposited as an machine-actionable object, ideally preserved together with links to other objects that can contextualize the object of study

Full data interoperability: the macroscopic potential



Fonte: (ROSNAY, 1979)

CLARIN ERIC – a distributed research infrastructure for language resources

After 10 years: a consortium of

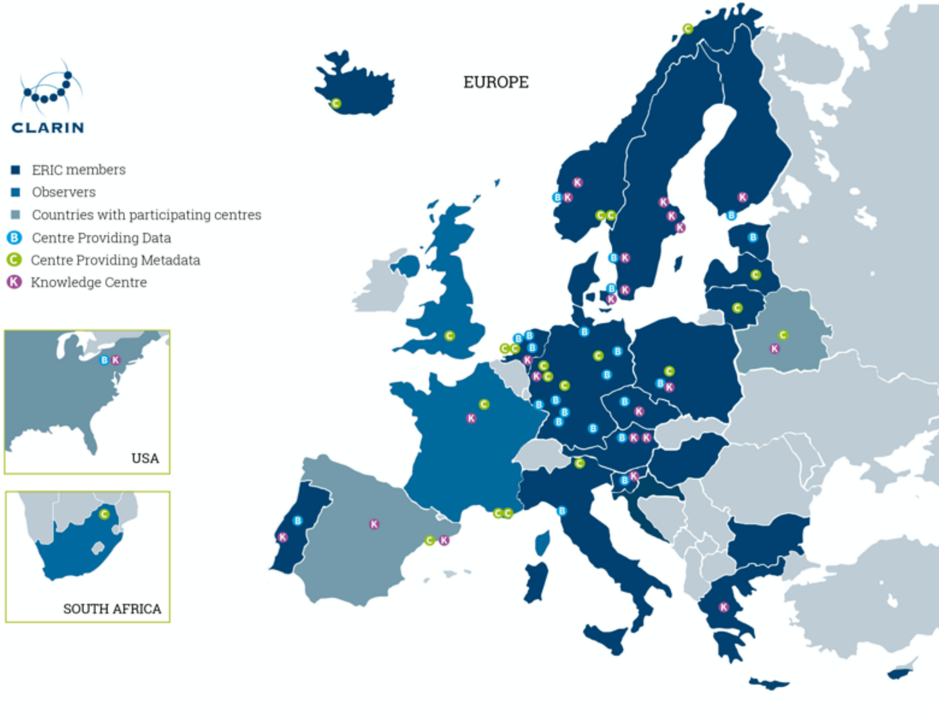
- 21 members
- 3 observers
- 1 linked party

A distributed network of >60 centres

25 CTS certified data centres,

strong focus on FAIRness & interoperability

- *federated login:*
- *central metadata harvesting for easy discovery:*
- *chained services.*



CLARIN data types and communities of use in SSH and beyond

- Newspaper archives
- Parliamentary records
- Literary texts
- Historical letters
- Broadcast archives
- Oral History data
- Social Media data
- L-2 Learner Resources
- Survey data
- Patient recordings
- Excavation reports

- Digital humanities
- Linguistics and Philology
- Data Science /AI
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture, Folklore, Anthropology
- Speech therapy
- General Public

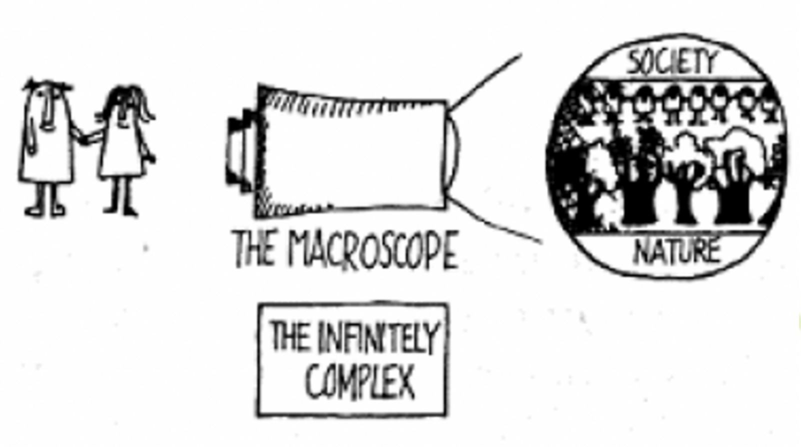
Text and speech as lens for SSH phenomena

- Language variation and multilinguality potentially provide the basis for comparative research of societal and cultural phenomena that are reflected in language use.
- Text and speech data can be used as social and cultural data
- Aligned RIs strong incentive for multidisciplinary initiatives
- Some examples:
  - Migration patterns
  - Intellectual history
  - Language variation across period and region
  - Dynamics in mental health conditions
  - Parliamentary discourse

Macroscopic perspective, SSH and the role of language

Inherent complexity of the objects of study in SSH

- Context is all
- Dynamics at a range of speeds
- Comparative perspective is key
  - synchronic
  - diachronic
- Subject and object overlap
- …



Language as multiplier for complexity

- Multilinguality
- Layered semantics
- Change over time
- …

CLARIN and Open Science 1/2

- Promotion for the sharing and re-use of language data through sustainable data registries:
  - language data studied from a comparative perspective
  - multidisciplinary collaboration
- Adherence to FAIR data principles: **F**indable, **A**ccessible, ***Interoperable***, **R**e-usable
- Enhancement and deployment of the interoperability of language data and services through, a.o.
  - a common metadata framework, ensuring resources are machine-actionable
  - networked certified data repositories, federated service offer
- Promotion for
  - responsible data science - to ensure humans stay in the loop and can take responsibility for results
  - replicability of research

CLARIN and Open Science 2/2

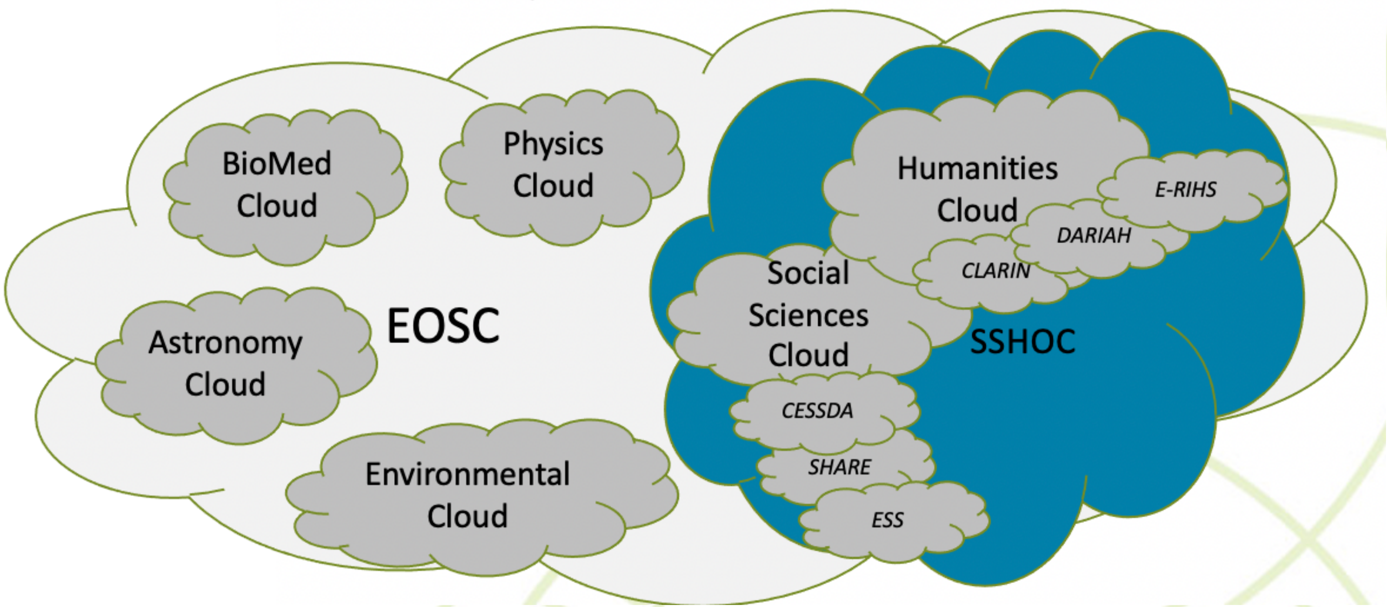
- Support for linguistic diversity
  - Data covering more than 1500 languages
  - Tools for many languages
  - Language resources in all modalities
- Promotion of collaboration between academia and the GLAM sector (GLAM = Galleries, Libraries, Archives, Museums)
- Exploring models for collaboration with industry: AI, machine translation, data science
- Integration in European Open Science Cloud
  - Services registered in EOSC Portal
  - SSH Open Cloud project to realize the SSH Open Marketplace (H2020-INFRAEOSC)

EUROPEAN OPEN SCIENCE CLOUD

SSHOC social sciences & humanities open cloud

EOSC as a Cloud of Disciplinary Clouds

- a distributed infrastructure, with shareable resources, optimized access for researchers and cost efficiency (aka: system of systems)
- offering options from the current research infrastructure landscape where much has been created already: thematic and generic services, platforms, collaboration at the level of clusters.



Parliamentary data as focus for networked SSH nodes

Relevance of parliamentary data

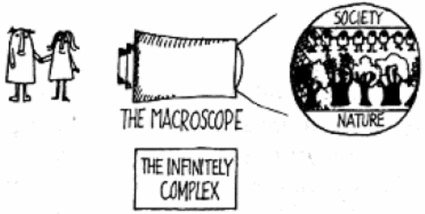
Parliamentary data directly corresponds to events and dynamics with global impact on human health, social and cultural life, and economics. In most contexts records available in open access. Basis for comparative study of public debate. Linked to many other open data sources.

Initiatives (2016-2021)

- Parliamentary data selected as primary resource family, to stimulate harmonisation of metadata, improve interoperability and discoverability and support comparative studies; >25 sources available.
- Development of TEI-based encoding standard (link), basis for a workshop series (2017-2020)
- ParlaMint project (2020-2021); link
  - Enable comparative studies of public debate on all topics related to the COVID-19 dynamics.
  - Resources and tools for focused observations on trends, opinions, decisions on lockdowns and restrictive measures as well as on the consequences with respect to health, medical care systems, employment, education, culture, etc.

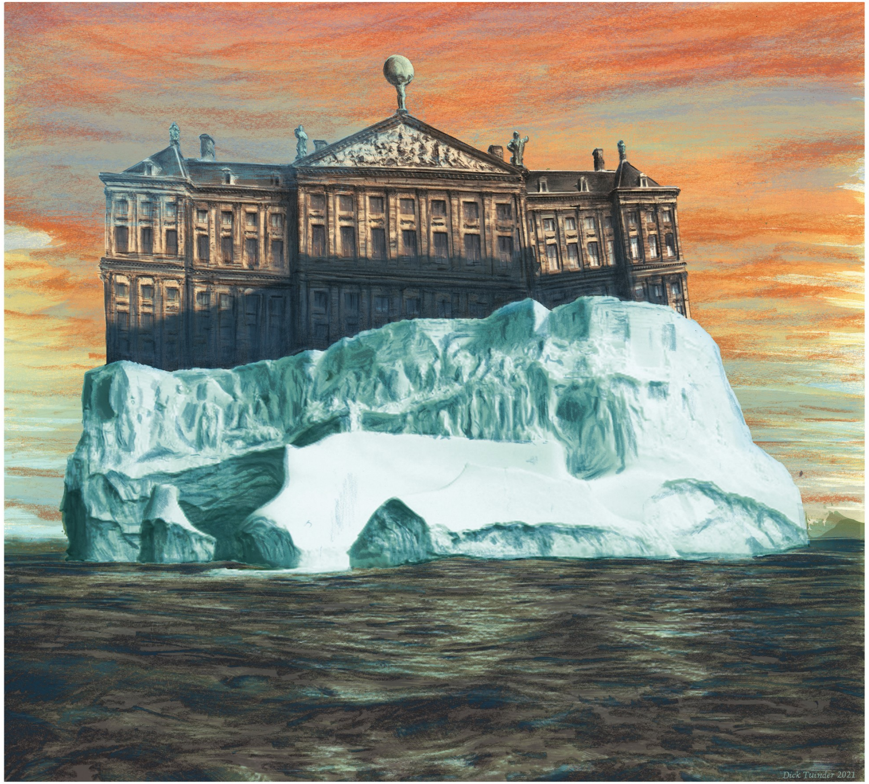
High ambition calls for more coherent policies

- Multifaceted landscape in need of proper balancing of priorities and resources
  - Disciplinary dynamics, emerging cluster collaboration
  - Local priorities (institutional, national)
  - European perspectives, international collaboration



- Policies to align mission driven initiatives (UN SDGs, HE) with RI development
- Future of HE funding
  - Support for individual Ris?
  - Funding for cluster activities?
- International perspectives (formal models for RI collaboration beyond Europe?)
- How to bridge Open Science policy and potential for uptake by industry
- Stakeholder priorities:
  - multiple voices from EC
  - national investments *versus* support for domain-driven dynamics
  - responsibility for pan-European priorities and potential societal impact

Driven by uncertainty



Artist impression of the future of the Royal Palace in Amsterdam

Dirk Tuinder 2021