Panelsessie: “Het aanmaken, beheren en delen van taalgegevens: bestaande praktijken en uitdagingen”

Moderator: Steven Krauwer (Universiteit Utrecht, CLARIN)

Vragen voor de panelsessie vrijdag 11 juni 11:45 – 12:15 (https://lr-coordination.eu/nl/node/366)

(1) What is the frame for sharing language data in NL?

- Policy level
  - □ AI strategy?
  - □ Language policy?
  - □ Other important developments on policy-level?
- Legal frame
  - □ What are the most relevant laws/regulations for data sharing?
  - □ Include also latest developments with regard to open data in your country
  - □ Where relevant cover legal from for both public sector information and data from private organisations
- Institutional level
  - □ Any specifics about panelists' institutional processes for data sharing

(2) What are the main challenges for sharing language data in NL?

- □ Please let each panelist name from his/her perspective!
- □ Examples that may pop-up: legal issues (specify), issues regarding responsibilities and management of data (specify), lack of training (specify) etc.

(3) What, from your own perspective, would need to be done to improve the situation and facilitate the sharing of language data in NL?

- □ Please let each panelist explain from his/her own perspective!

Struktuur panelsessie vrijdag 11 juni 11:45 – 12:15

We doen het in 4 rondes:

1. elke panelist stelt zichzelf en zijn organisatie voor (elk 1 slide met naam, functie, beschrijving organisatie, totaal 4x 2 minuten)
2. elke panelist gaat in op het frame (vraag 1: policy, legal, institutional), steeds vanuit het eigen perspectief (elk 2 slides, 4x 2 minuten)
3. elke panelist gaat in op de challenges (vraag 2) en wat er (weer vanuit het eigen perspectief) gedaan moet worden (vraag 3) (elk 2 slides, 4x 2 minuten)
4. vragen uit het publiek aan de panelists (max 6 minuten in theorie, waarschijnlijk veel minder)

Panelists hebben hun slides woensdag 9 juni ingeleverd, zodat ze tevoren op de juiste plek in de moederpresentatie kunnen worden gestoken, waardoor er geen wisselingen nodig zullen zijn. Die kosten te veel tijd - en dat hebben we niet!

Ronde 1: sprekers en organisaties stellen zich voor

Ted van der Togt, Koninklijke Bibliotheek Research

Hans Overbeek, KOOP (Kennis- en Exploitatiecentrum Officiële Overheidspublicaties)

Vincent Vandeghinste, ELG (European Language Grid)

Franciska de Jong, CLARIN (Common Language resources and Technology Infrastructure)

KB nationale bibliotheek

KB, the national Library of the Netherlands, maintains the national collection of everything published in and about the Netherlands (books, newspapers, magazines and other publications).

Together with Nationaal Archief we work on getting our collections more usable.

NA maintains the largest collection of archives in the Netherlands. The collection consists of important documents of the Government, archives of the province Zuid-Holland and parivate archives

Ted van der Togt, KB Research

KOOP

Kennis- en Exploitatiecentrum
Officiële Overheidspublicaties

KOOP is the Dutch publications office. It serves as the official publisher of the central and local government of the Netherlands.

officielebekendmakingen.nl

- Tractatenblad (treaties)
- Staatsblad and Staatscourant
- Provinciale, waterschaps-, gemeente- and gemeenschappelijkeregelingbladen

overheid.nl

- Consolidated legislation and regulation
- Parliamentary information (“kamerstukken”)

data.overheid.nl

- Data register of the Dutch government

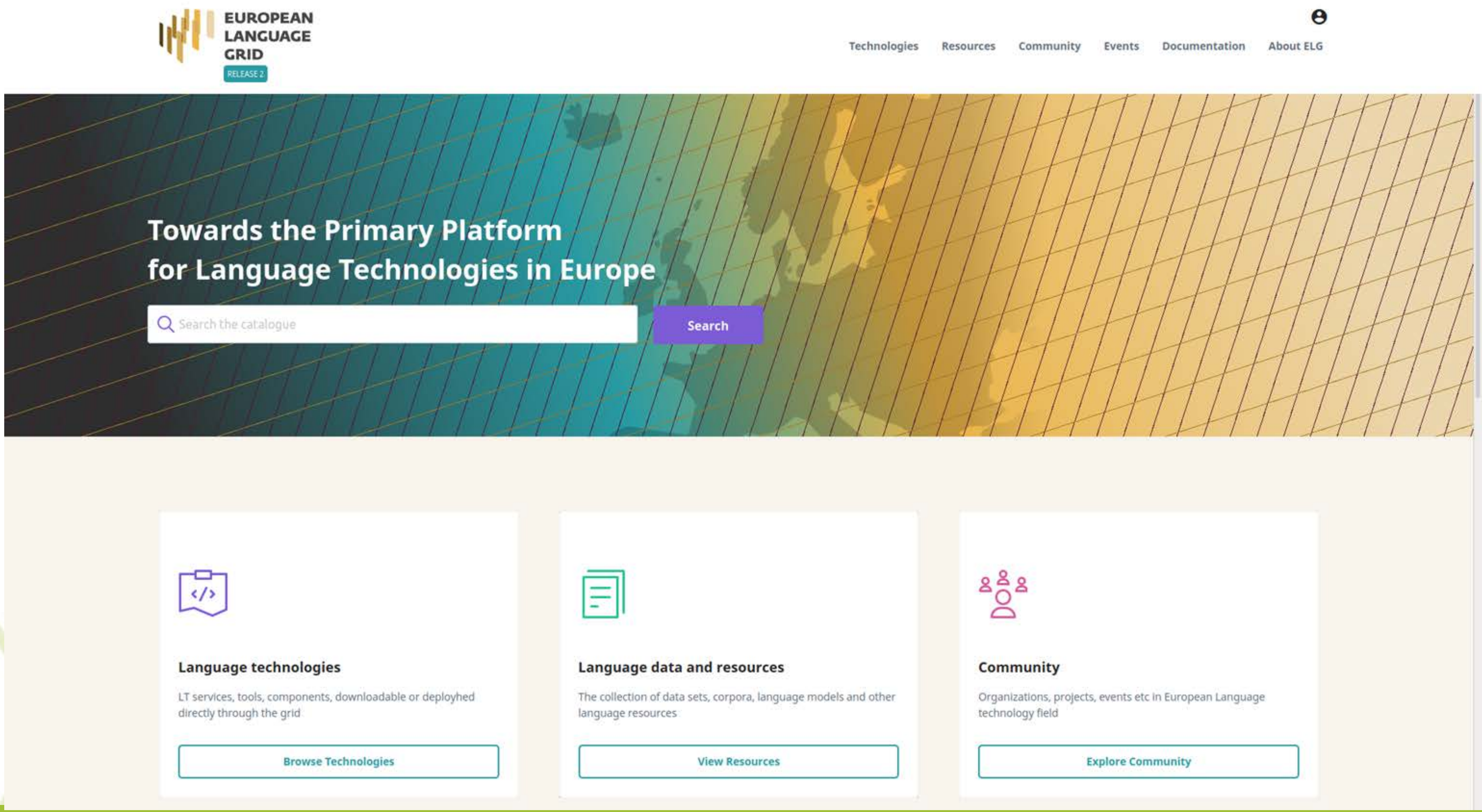
Hans Overbeek - Advisor content standards

Need Dutch content? Just let me know! hans.overbeek@koop.overheid.nl

Yellow pages of Language Resources in Europe

Vincent Vandeghinste - ***NCC lead the Netherlands***

https://www.european-language-grid.eu/



EUROPEAN LANGUAGE GRID

Towards the Primary Platform for Language Technologies in Europe

Language technologies

LT services, tools, components, downloadable or deployhed directly through the grid

Language data and resources

The collection of data sets, corpora, language models and other language resources

Community

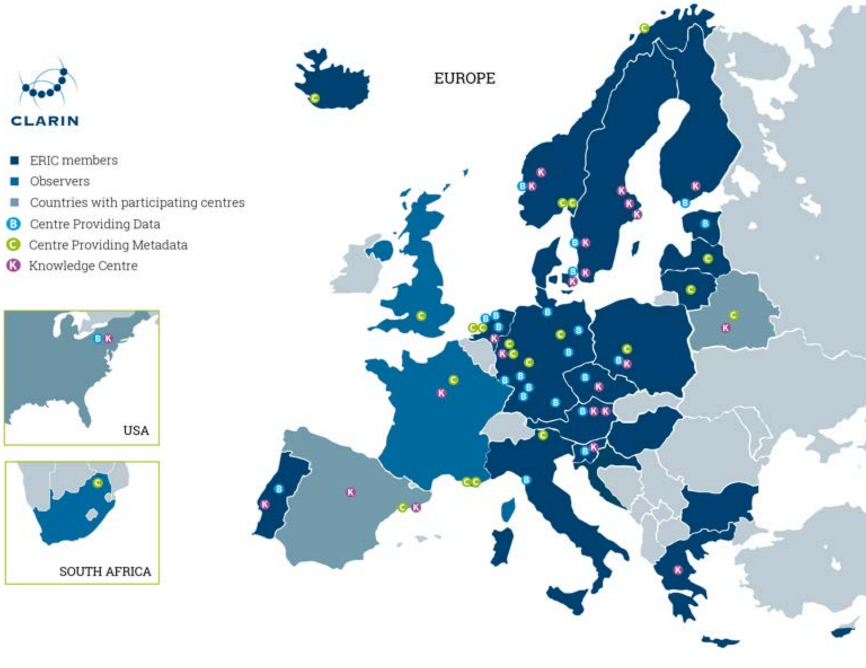
Organizations, projects, events etc in European Language technology field

Franciska de Jong - *director of CLARIN ERIC*- www.clarin.eu

Research Infrastructure CLARIN
(= Common Language Resources and Technology Infrastructure)

A consortium of type ERIC; after 10 years:

- 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 discove
- chained services:
- language data - in written, spoken, video or multimodal form
- advanced tools - to discover, explore, exploit, annotate, analyse or combine data sets, *wherever they are located*

Ronde 2: sprekers reageren vanuit het perspectief van hun organisatie op de eerste vraag

(1) What is the frame for sharing language data in NL?

- Policy level
  - □ AI strategy?
  - □ Language policy?
  - □ Other important developments on policy-level?
- Legal frame
  - □ What are the most relevant laws/regulations for data sharing?
  - □ Include also latest developments with regard to open data in your country
  - □ Where relevant cover legal from for both public sector information and data from private organisations
- Institutional level
  - □ Any specifics about panelists' institutional processes for data sharing

KB nationale bibliotheek

- Ambition to bring all printed (public domain) publications online (via services like Delpher & DBNL)
- Dataservices and Linked Open Data (data.bibliotheken.nl)
- Negotiations with rights holders about TDM, for out of commerce and copyrighted material.
- National and International Cooperation (Clariah, Future Library Lab, European Digital Reading Lab, Cultural AI Lab)
- Example project “Web Publications for digitized content” together with the Nationaal Archief, TU Delft and Bureau van Leeuwen & van Leeuwen
- KB LAB https://lab.kb.nl/
- *NA Datalab* https://www.nationaalarchief.nl/over-het-na/datalab-nationaal-archief

(1a) Sharing language data in NL

Current situation

- Need for translated national legislation is underestimated
- Many translations by 3rd parties already available, but hidden

Regulations

- “Wet open overheid” (I found no translation ;-) → active disclosure of “everything” via PLOOI (PLatform Open Overheid Informatie by KOOP)
- "Wet Elektronische Publicatie" → ALL legislation and regulation available online

(1b) Sharing language data in NL

Institutional level

- Data.overheid.nl:
    - catalogue of all data sets (open&closed)
    - data broker to help finding and disclosing hidden data sets
- Entity extraction (e.g. legal references)
- Semantic web technology in metadata and reference data

Framework

Language Technologies (LT) are vital to maintain an inclusive Digital Single Market

- 24 official languages and many more additional languages
- fragmentation of LT business environment in Europe
  - high number of specialised companies,
  - prevents advanced LT research transfer into industry and commerce.
- ELG is a scalable platform
  - delivering easy access to hundreds of commercial and non-commercial LT for all European languages
  - aiming to be the key platform for LT in Europe
  - running instruments and service
  - data sets and resources
- aiming to improve the Multilingual Digital Single Market and create new jobs

What is the frame for sharing language data in NL?

Policy level

- Investments in a national AI programme have recently been granted: a budget is allocated to the NL AI Coalition in the context of the so-called Groeifonds funding.
- The Netherlands is actively involved in the shaping of the European Open Science Cloud. Open Data and adherence to the FAIR principles is an important criterion for eligibility for public funding for data infrastructures.
- NWO funding for CLARIAH-NL as large-scale research infrastructure.

Legal frame

- Research data is more an more expected to become available in line with the Open Science agenda: openly available when possible, protected if necessary.
- The Netherlands has always been a strong advocate of Open Science. Currently a national initiative known under the name NPOS (https://www.openscience.nl/) is actively promoting data interoperability, alignment of data access policies and the establishment of a network of skilled data stewards.

Institutional level

- CLARIN/leverages the investment of national/institutional processes for language data sharing.
- If a data centre adheres to common standards of metadata interoperability, CLARIN can harvest the resources and make them discoverable through the central platform known as Virtual Language Observatory.
- CLARIN is contributing to EOSC federation of services with so-called thematic services for language data, which bring even wider visibility of harvested data.

Ronde 3: sprekers reageren op de volgende vragen

(2) What are the main challenges for sharing language data in NL?

- □ Please let each panelist name from his/her perspective!
- □ Examples that may pop-up: legal issues (specify), issues regarding responsibilities and management of data (specify), lack of training (specify) etc.

(3) What, from your own perspective, would need to be done to improve the situation and facilitate the sharing of language data in NL?

- □ Please let each panelist explain from his/her own perspective!

KB nationale bibliotheek

- IT infrastructure big change (New Digital Storage)
- Standards are essential (but could only be decided together with network partners, National & International)
- Description of data quality (FAIR)
- Many OCR data sets need to be curated before becoming really useful (for AI but also other purposes).
- Challenges:
  - Who does the curation?
  - Where do we store this curated data?
  - What standards do we use for this curated data?

(2) Challenges for sharing language data in NL

- Switch from passive disclosure (Wob) to active disclosure (Woo)

(3) How to improve the sharing of language data in NL?

- Coordinated regime (stelsel) of data catalogs
- Standardisation of metadata (DCAT)
- Identify more base registers. Apply or complain policy for uniform use of reference data: use URIs, NOT labels, to identify organisations, publications, datasets, themes (e.g. EUROVOC, NL TOP-lijst) and other resources

No AI without IA *)

*) AI: Artificial Intelligence
IA: Information Architecture

Challenges

- Collecting information on available resources
  - what is available
  - licenses
    - research only
    - commercial use
    - online service versus download
- Keeping information up to date
  - online services should keep functioning
  - yellow pages should be kept up to date
- Ethical and legal issues
  - GDPR
    - anonymization / pseudonymization is not always possible
    - Informed consent for video data (sign language)
  - IPR
    - value of text for publishers
  - How to convince law departments?

Potential solutions?

- Build automated checks to keep information up to date?
- Keep regular contact with yellow page contact persons at each resource provider
- Make prototype contracts for different situations available to everyone
- Make prototype informed consent forms available
- Provide examples of large organisation that make their data / tools available
- Provide tools for automated anonymization / pseudonymization

What are the main challenges for sharing language data in NL?

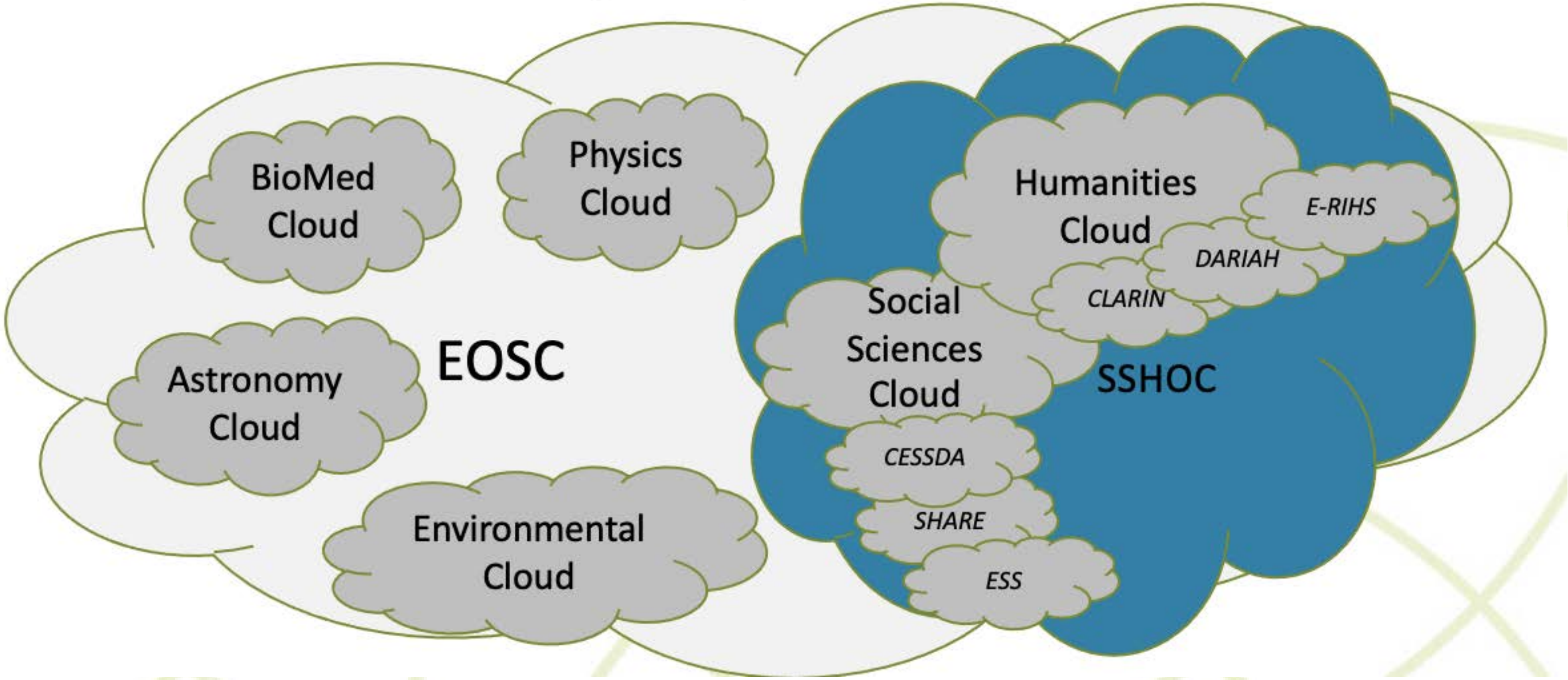
- Unclarity and uncertainty on how to comply with the GDPR framework
  - Disruption and contradictions, in the case of e.g. interview data, survey data
  - For RIs impact is expected from collaboration with industry, but open data can often only be shared for non-commercial purposes
- In research contexts: no clear guidelines for where to deposit data to comply with rules for the management of research data
- Limited institutional capacity for supporting researchers
- Focus on in-house development; not always aligned with emerging standards rooted in developments elsewhere.

What would need to be done to improve the situation and facilitate the sharing of language data in NL?

- Clarity on roles and responsibilities at the various relevant levels:
  - National organisations (e.g. OCW, NWO, VSNU, etc.)
  - Individual universities and academic institutes
  - Faculties and departments
  - Individual researchers and disciplinary communities
- Incentives in the assessment and reward system for academic researchers generating data sets
- Acceptance of disciplinary repositories that comply with international standards as adequate route for Research Data Management
- Better understanding of the potential of EOSC for thematic services

Added value of EOSC:
a Cloud of Disciplinary Clouds

- a distributed infrastructure, with shareable resources, optimized access for researchers and cost efficiency
- 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.



Laatste ronde (als de tijd het toelaat): het woord is aan het publiek