CLARIN's central infrastructure

Dieter Van Uytvanck

CLARIN-PLUS Tools & Services Workshop

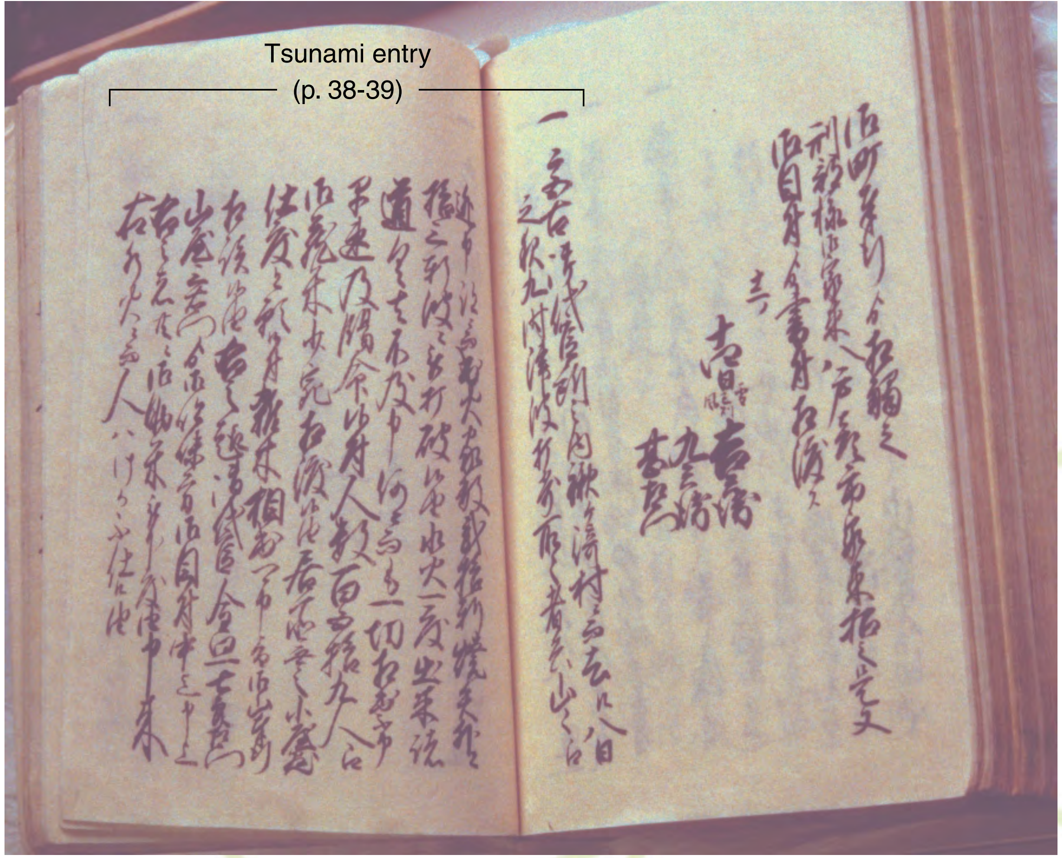
2 June 2016

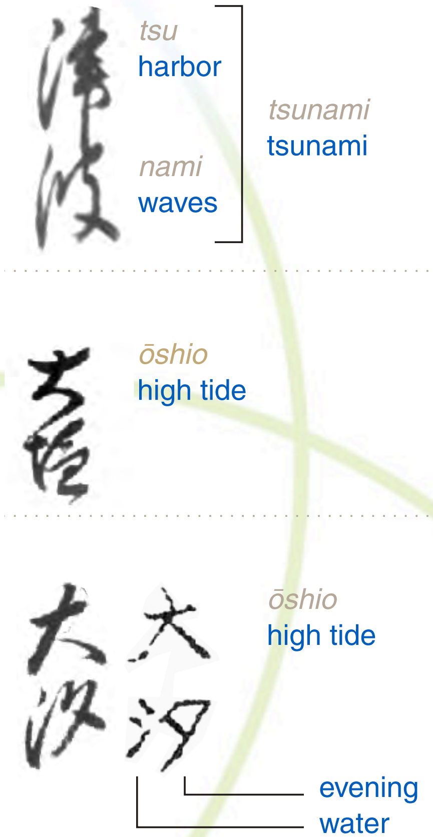
Vienna

CLARIN?

- **C**ommon **La**nguage **R**esources and Technology **In**frastructure
- Research Infrastructure for the **humanities and social sciences**
- Provides easy and sustainable access for scholars
  - to **digital language data** (in written, spoken, video or multimodal form)
  - to **advanced tools** to discover, explore, exploit, annotate, analyse or combine them

Language resources: more than linguistics


Tsunami entry
(p. 38-39)

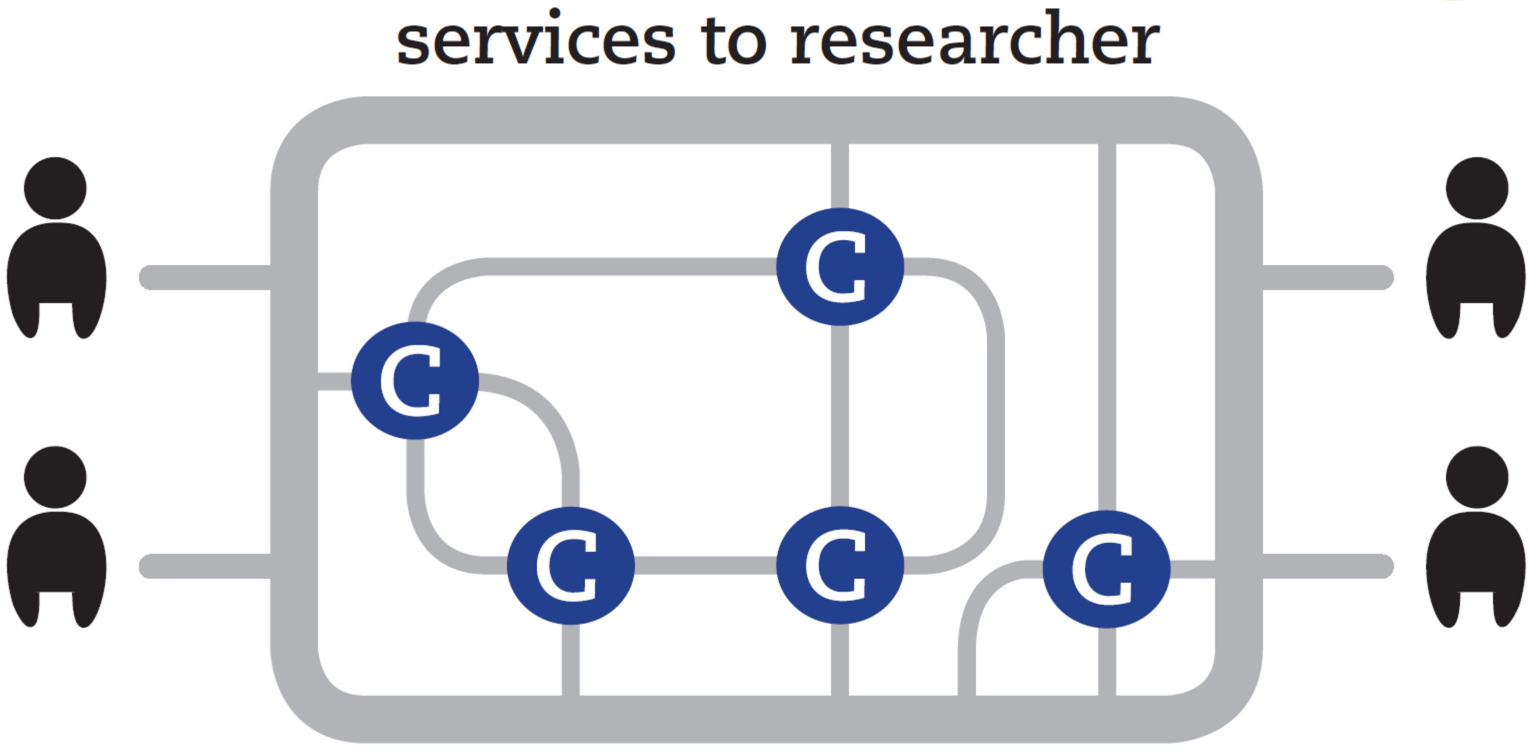


Source:

- Atwater, B.F., Musumi-Rokkaku, S., Satake, K., Tsuji, Y., Ueda, K., and Yamaguchi, D.K., 2015, The orphan tsunami of 1700—Japanese clues to a parent earthquake in North America, 2nd ed.: Seattle, University of Washington Press, U.S. Geological Survey Professional Paper 1707, 135 p.
- doi:10.3133/pp1707

CLARIN centres

- A **distributed architecture**: (http-accessible) files, web applications and web services spread all over Europe
- Nodes in the network: **centres** (http://clarin.eu/centres)



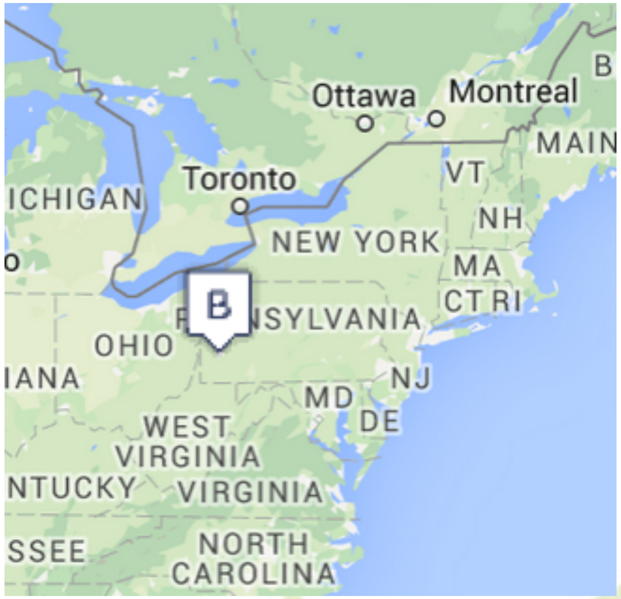
Organisation CLARIN

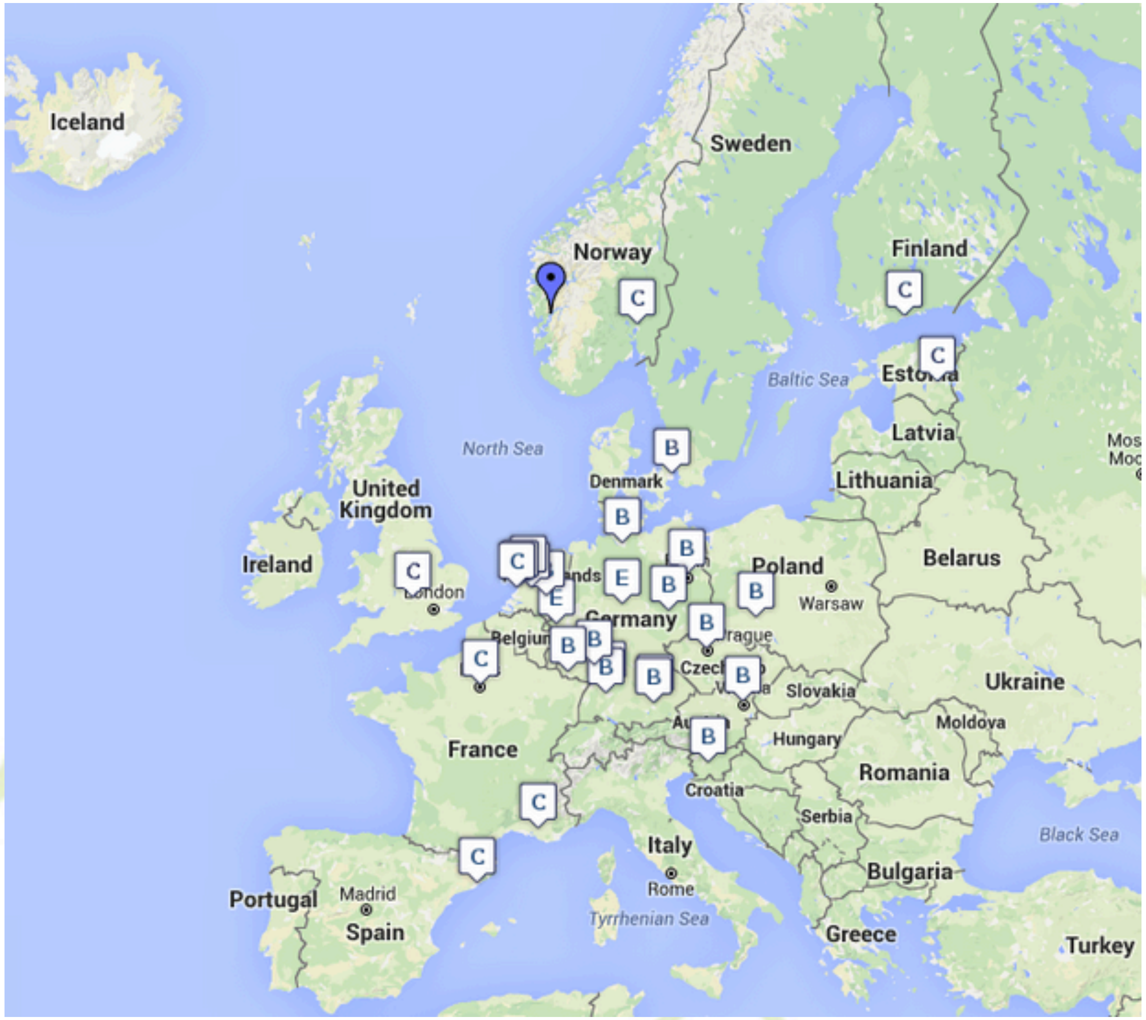
- European (ESFRI) Research Infrastructure
- ERIC since 2012
- Landmark since 2016
- **Members**: Austria • Bulgaria • Czech Republic • Denmark • Dutch Language Union • Estonia • Finland • Germany • Greece • Italy • Lithuania • Netherlands • Norway • Poland • Portugal • Slovenia • Sweden • United Kingdom (observer)

Benefits for countries

- **Access to the CLARIN Infrastructure**, i.e. to all CLARIN language resources and technology services
- **Access to expertise** via the CLARIN Knowledge Sharing Infrastructure
- **Embedding** in the humanities **research community**, with access to the same data
- **Better visibility of their language**, their research results, their resources and their **cultural heritage**
- **Opportunities**
  - for cross-lingual and -cultural **research**
  - to participate in **research projects** in which CLARIN ERIC participates as a beneficiary

The 33 CLARIN centres



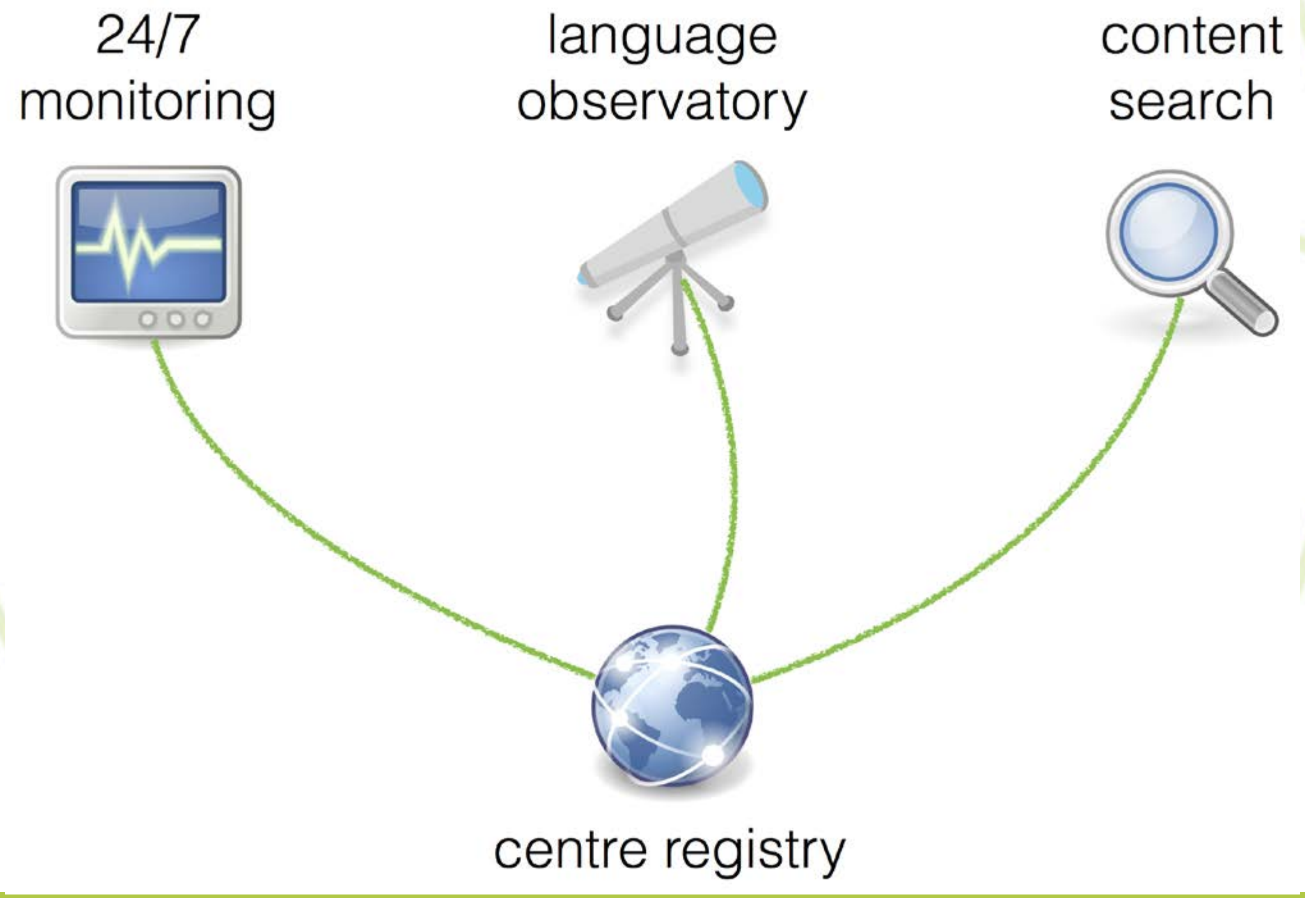


CLARIN technology pillars

- **Federated Identity** - letting users login to protected data and services with their own institutional username and password
- **Persistent Identifiers** - enabling sustainable citations of electronic resources
- **Sustainable repositories** - digital archives where language resources can be stored, accessed and shared
- **Flexible metadata and concept definitions** - to ensure semantic interoperability when describing language resources
- **Well-described and open protocols**, e.g.:
  - **Content search** - offering a search engine for a wide range of language resources
  - **Web service chaining** - giving users the possibility to freely combine language processing services

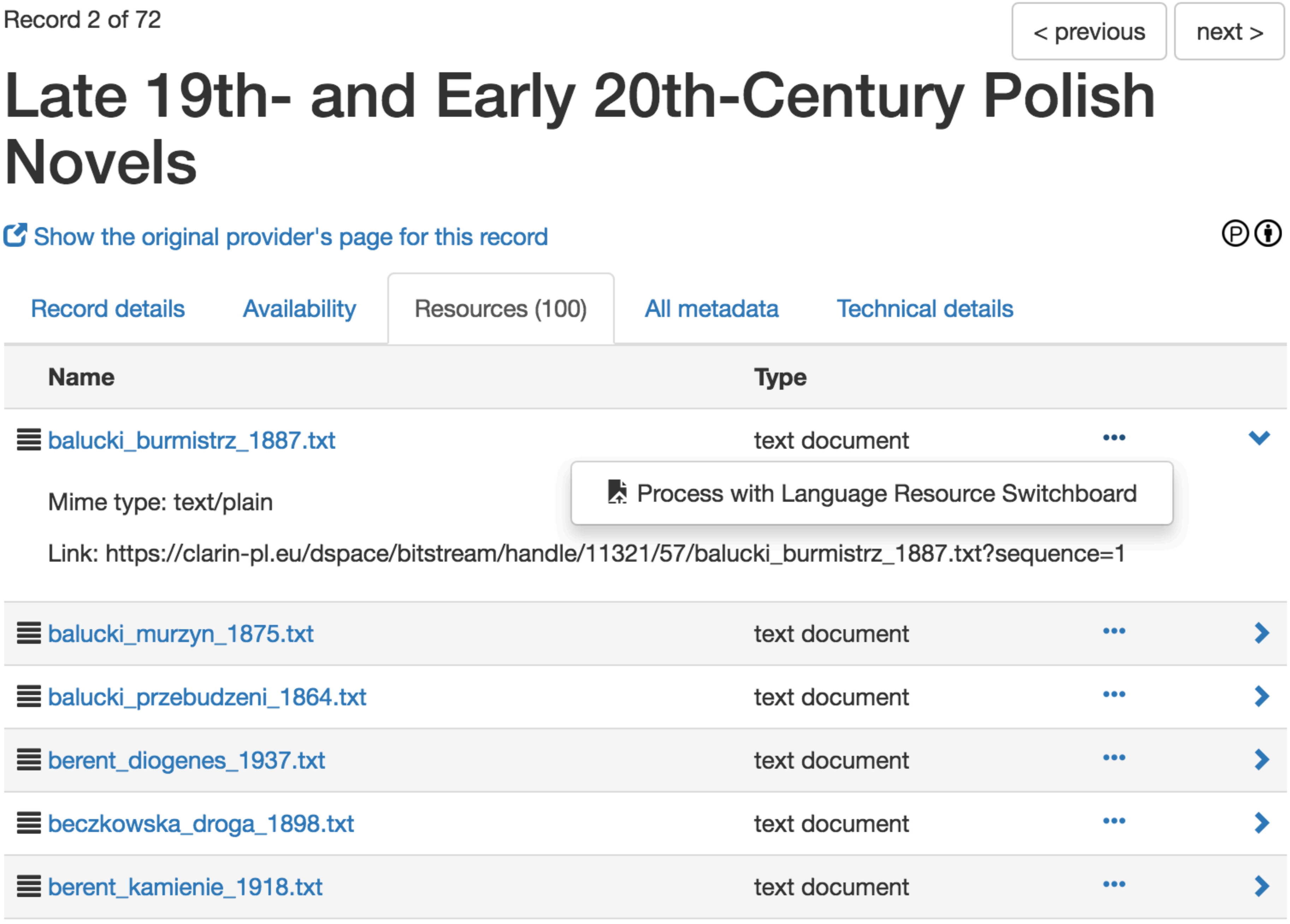
Seamless integration *within CLARIN*

- Centres and Services are not isolated islands but part of a well-integrated setup



Seamless integration *within CLARIN*

- e.g. ELAN with WebLicht (tagging) and WebMaus (phonetic alignment)
- e.g. the Language Resource Switchboard



Record 2 of 72

Late 19th- and Early 20th-Century Polish Novels

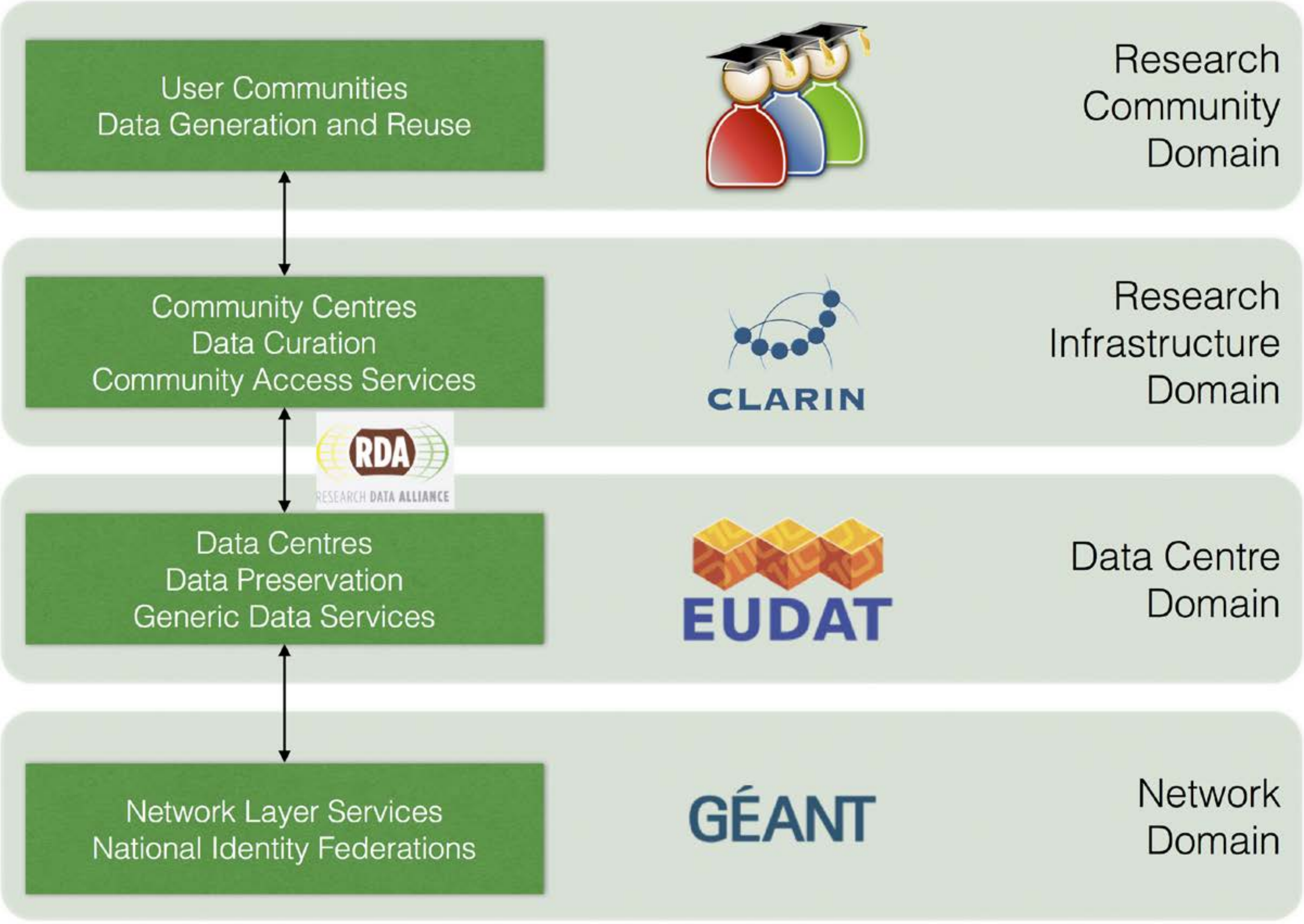
Show the original provider's page for this record

Name	Type
balucki_burmistrz_1887.txt	text document
balucki_murzyn_1875.txt	text document
balucki_przebudzeni_1864.txt	text document
berent_diogenes_1937.txt	text document
beczkowska_droga_1898.txt	text document
berent_kamienie_1918.txt	text document

Mime type: text/plain

Link: https://clarin-pl.eu/dspace/bitstream/handle/11321/57/balucki_burmistrz_1887.txt?sequence=1

Seamless integration *in the infrastructure landscape*



User Communities
Data Generation and Reuse

Research Community Domain

Community Centres
Data Curation
Community Access Services

Research Infrastructure Domain

Data Centres
Data Preservation
Generic Data Services

Data Centre Domain

Network Layer Services
National Identity Federations

Network Domain

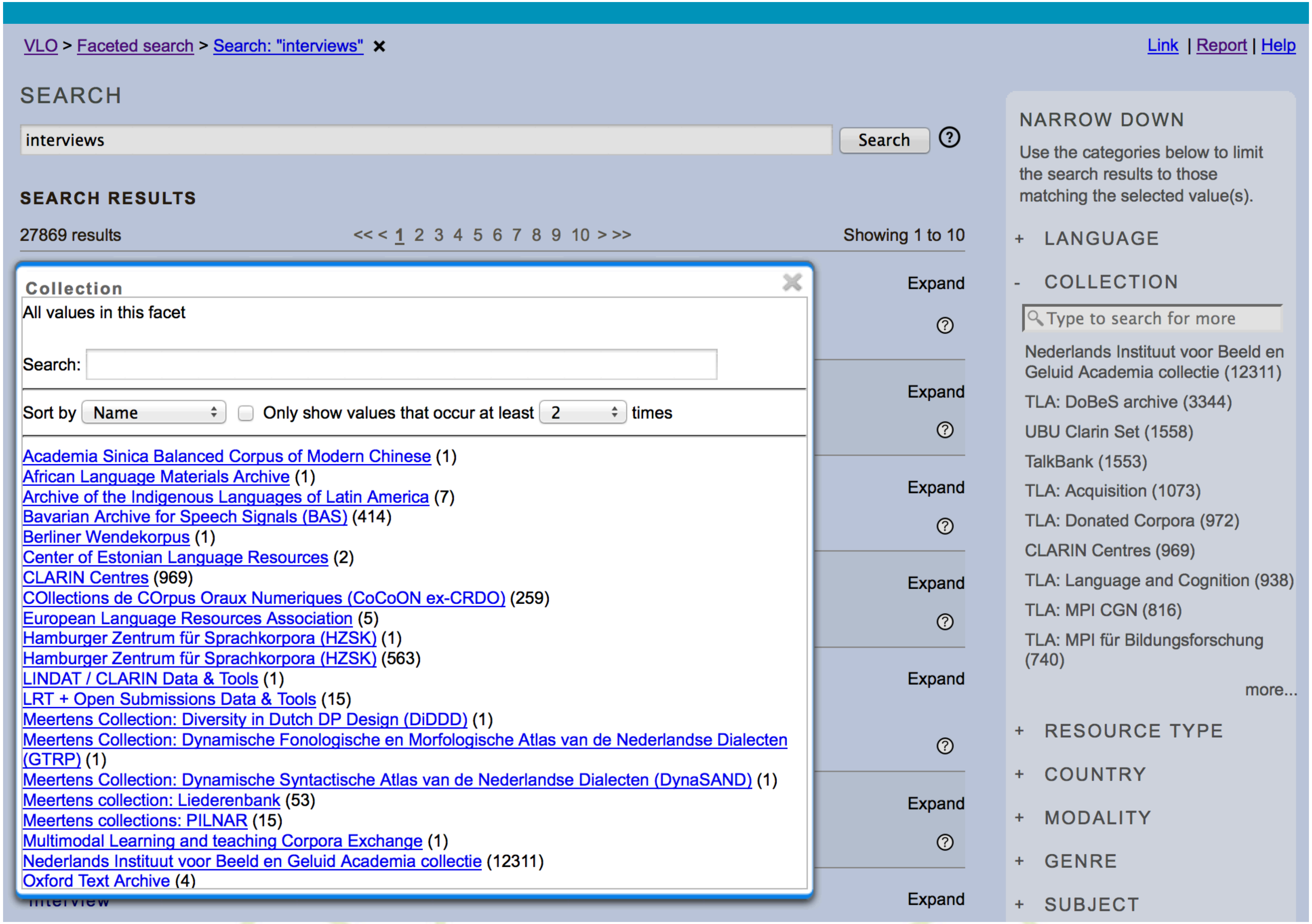
Infrastructure Overview (1)

- https://trac.clarin.eu/wiki/InfrastructureOverview
- No trac account? See http://clarin.eu/svn

- 1. Technology
  - 1.1. Gateway applications
    - 1.1.1. Virtual Language Observatory
    - 1.1.2. Federated Content Search engine
    - 1.1.3. Virtual Collection Registry

Central services for researchers

Virtual Language Observatory



VLO > Faceted search > Search: "interviews" ✖

Link | Report | Help

SEARCH

interviews

SEARCH RESULTS

27869 results

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Showing 1 to 10

Collection

All values in this facet

Search:

Sort by Name ☐ Only show values that occur at least 2 times

Academia Sinica Balanced Corpus of Modern Chinese (1)
African Language Materials Archive (1)
Archive of the Indigenous Languages of Latin America (7)
Bavarian Archive for Speech Signals (BAS) (414)
Berliner Wendekorpus (1)
Center of Estonian Language Resources (2)
CLARIN Centres (969)
COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO) (259)
European Language Resources Association (5)
Hamburger Zentrum für Sprachkorpora (HZSK) (1)
Hamburger Zentrum für Sprachkorpora (HZSK) (563)
LINDAT / CLARIN Data & Tools (1)
LRT + Open Submissions Data & Tools (15)
Meertens Collection: Diversity in Dutch DP Design (DiDDD) (1)
Meertens Collection: Dynamische Fonologische en Morfologische Atlas van de Nederlandse Dialecten (GTRP) (1)
Meertens Collection: Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND) (1)
Meertens collection: Liederenbank (53)
Meertens collections: PILNAR (15)
Multimodal Learning and teaching Corpora Exchange (1)
Nederlands Instituut voor Beeld en Geluid Academia collectie (12311)
Oxford Text Archive (4)

interview

NARROW DOWN

Use the categories below to limit the search results to those matching the selected value(s).

+ LANGUAGE

- COLLECTION

Type to search for more

Nederlands Instituut voor Beeld en Geluid Academia collectie (12311)
TLA: DoBeS archive (3344)
UBU Clarin Set (1558)
TalkBank (1553)
TLA: Acquisition (1073)
TLA: Donated Corpora (972)
CLARIN Centres (969)
TLA: Language and Cognition (938)
TLA: MPI CGN (816)
TLA: MPI für Bildungsforschung (740)

more...

+ RESOURCE TYPE

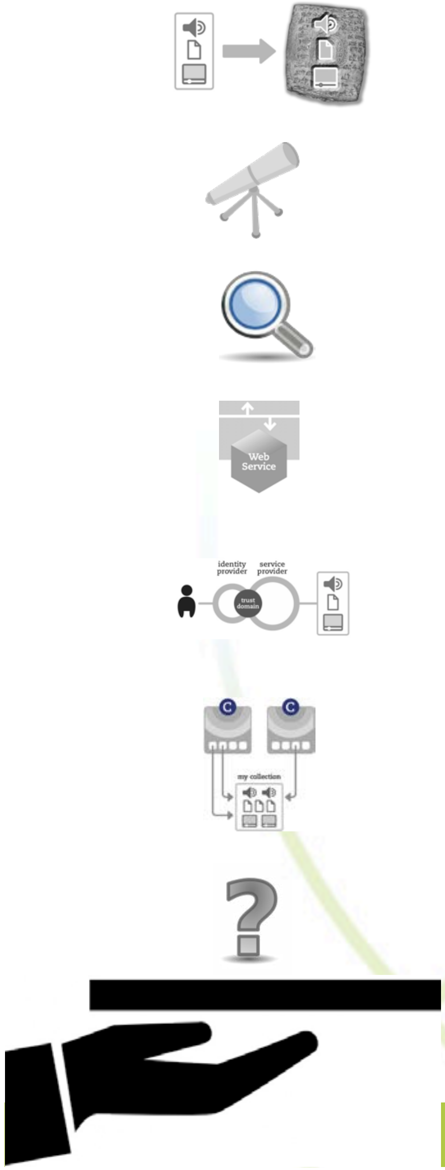
+ COUNTRY

+ MODALITY

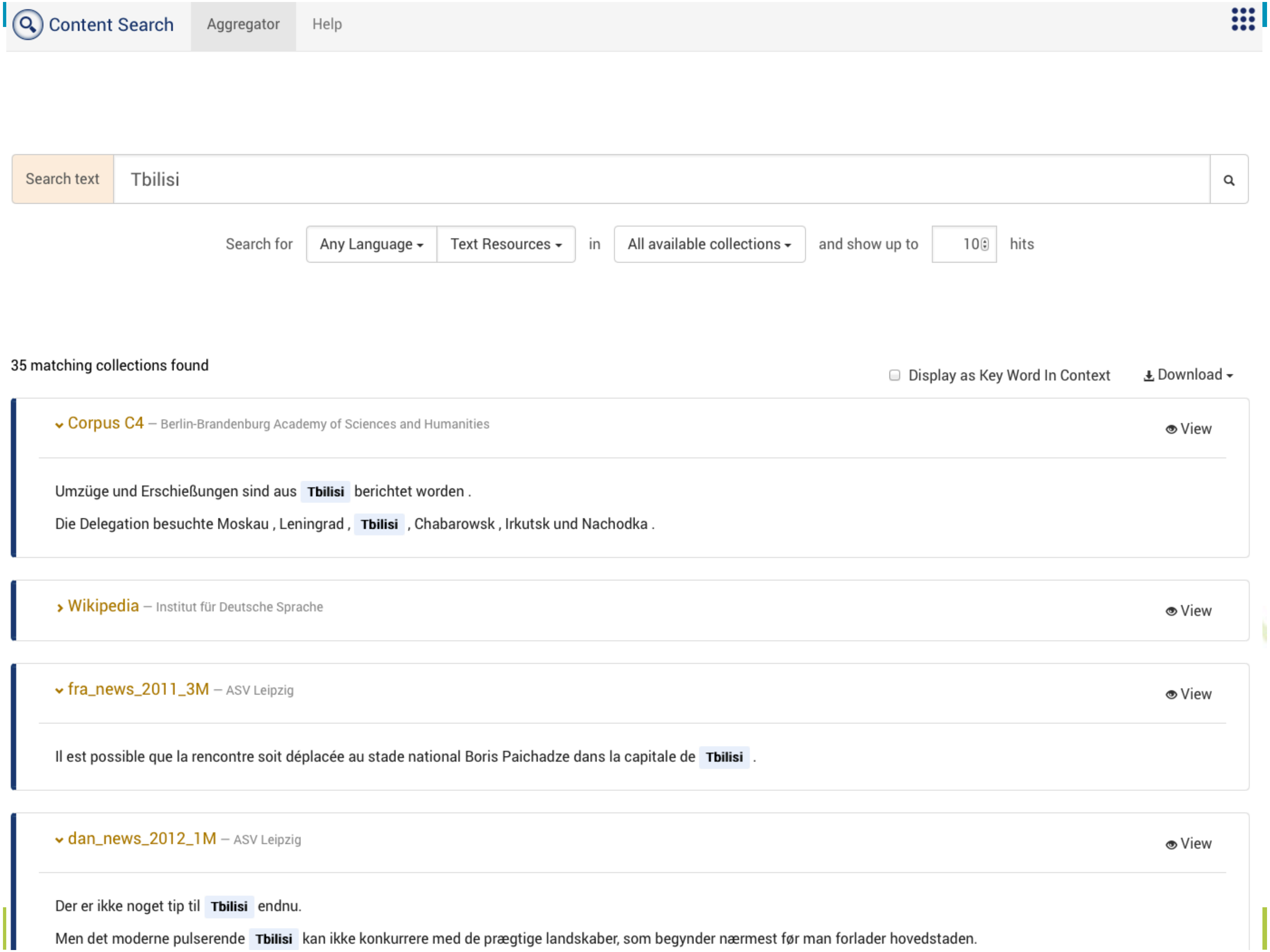
+ GENRE

+ SUBJECT

Central services for researchers



Federated Content Search



Search text: Tbilisi

Search for Any Language ▾ Text Resources ▾ in All available collections ▾ and show up to 10 hits

35 matching collections found

□ Display as Key Word In Context

Corpus C4 — Berlin-Brandenburg Academy of Sciences and Humanities

Umzüge und Erschießungen sind aus **Tbilisi** berichtet worden .

Die Delegation besuchte Moskau , Leningrad , **Tbilisi** , Chabarowsk , Irkutsk und Nachodka .

Wikipedia — Institut für Deutsche Sprache

fra_news_2011_3M — ASV Leipzig

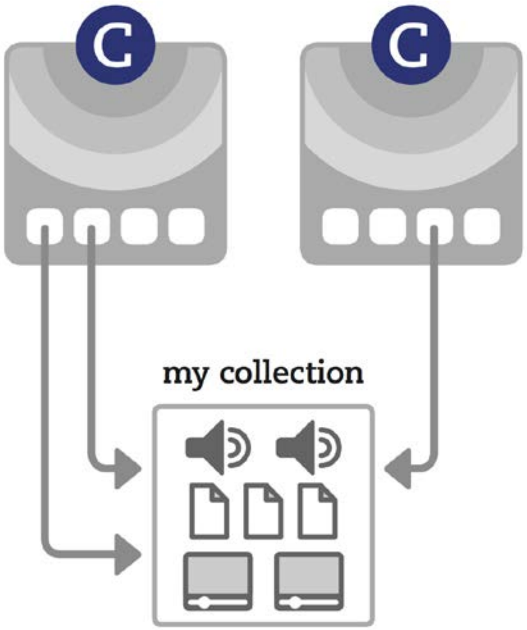
Il est possible que la rencontre soit déplacée au stade national Boris Paichadze dans la capitale de **Tbilisi** .

dan_news_2012_1M — ASV Leipzig

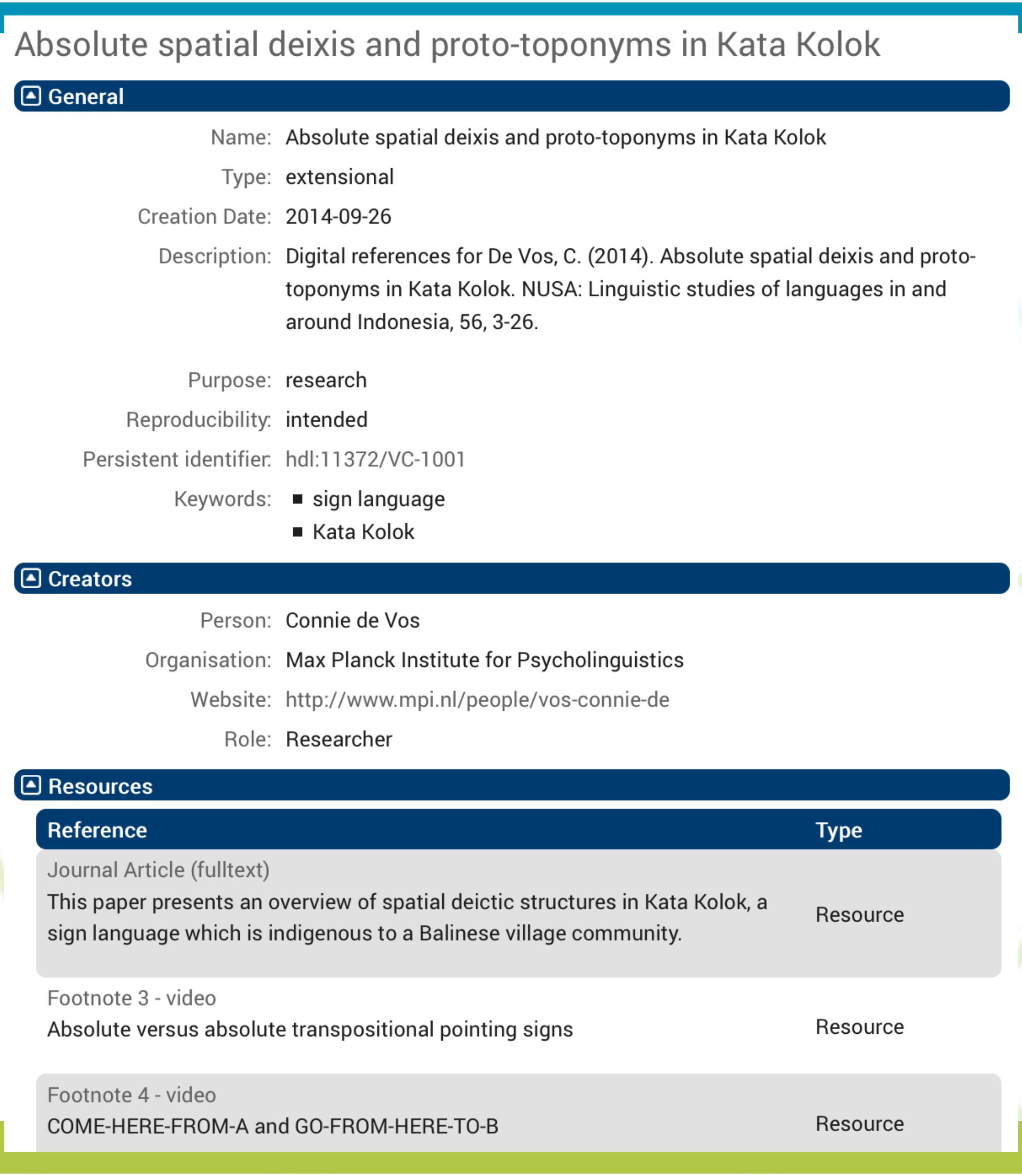
Der er ikke noget tip til **Tbilisi** endnu.

Men det moderne pulserende **Tbilisi** kan ikke konkurrere med de prægtige landskaber, som begynder nærmest før man forlader hovedstaden.

Central services for researchers



Virtual Collections



Absolute spatial deixis and proto-toponyms in Kata Kolok

General

Name:	Absolute spatial deixis and proto-toponyms in Kata Kolok
Type:	extensional
Creation Date:	2014-09-26
Description:	Digital references for De Vos, C. (2014). Absolute spatial deixis and proto-toponyms in Kata Kolok. NUSA: Linguistic studies of languages in and around Indonesia, 56, 3-26.
Purpose:	research
Reproducibility:	intended
Persistent identifier:	hdl:11372/VC-1001
Keywords:	■ sign language ■ Kata Kolok

Creators

Person:	Connie de Vos
Organisation:	Max Planck Institute for Psycholinguistics
Website:	http://www.mpi.nl/people/vos-connie-de
Role:	Researcher

Resources

Reference	Type
Journal Article (fulltext) This paper presents an overview of spatial deictic structures in Kata Kolok, a sign language which is indigenous to a Balinese village community.	Resource
Footnote 3 - video Absolute versus absolute transpositional pointing signs	Resource
Footnote 4 - video COME-HERE-FROM-A and GO-FROM-HERE-TO-B	Resource

Infrastructure Overview (2)

- 1.2. Infrastructure applications
  - 1.2.1. Centre Registry
  - 1.2.2. Metadata
    - 1.2.2.1. Metadata Harvester
    - 1.2.2.2. Component Registry
    - 1.2.2.3. Concept Registry
    - 1.2.2.4. Curation module
  - 1.2.3. Federated Identity
    - 1.2.3.1. SAML metadata aggregation: PyFF
    - 1.2.3.2. CLARIN Identity Provider
    - 1.2.3.3. Discovery Service
    - 1.2.3.4. Authorisation Service
  - 1.2.4. Piwik
  - 1.2.5. Validators
    - 1.2.5.1. OAI-PMH (metadata) validator
    - 1.2.5.2. CMDI (metadata) validator
    - 1.2.5.3. SRU-CQL (FCS) validator
  - 1.2.6. Language Resource Switchboard

Infrastructure Overview (3)

- 1.3. Software development
  - 1.3.1. Subversion repository
  - 1.3.2. Trac
  - 1.3.3. Nexus maven repository
  - 1.3.4. Docker repository
  - 1.3.5. Third-party infrastructure (hosted in the cloud)
    - 1.3.5.1. GitHub
    - 1.3.5.2. Travis
    - 1.3.5.3. Jenkins

- 1.4. System management
  - 1.4.1. Icinga
  - 1.4.2. Grafana

- 1.5. Office systems [...]

Infrastructure Overview (4)

- 1.6. Communication systems
  - 1.6.1. Website
  - **1.6.2. Mailing Lists**
  - 1.6.3. MailChimp
  - 1.6.4. Basecamp
  - **1.6.5. Slack**
  - 1.6.6. Social media

- 1.7. Third-party infrastructure applications & services
  - **1.7.1. B2DROP [EUDAT]**
  - 1.7.2. B2SHARE [EUDAT]
  - **1.7.3. B2SAFE [EUDAT]**
  - 1.7.4. B2ACCESS [EUDAT]
  - 1.7.5. Potential A-services
  - **EPIC: PID services**

Take-home message about tools and services

- think about integration, e.g.
  - Federated Content Search
  - Language Resource Switchboard
- avoid closed shops
  - use federated login, if authentication is needed
- think about usability
  - CLARIN human interface guidelines
- consider re-using existing applications (e.g. corpus query engines)
- CLARIN ERIC is at your service with
  - outreach (featured showcases/resources)
  - mobility grants
  - advice

Thank you for your attention!