Technical infrastructure – an overview

Dieter Van Uytvanck
Max Planck Institute for Psycholinguistics
CLARIN ERIC director
Dieter.VanUytvanck@mpi.nl
CLARIN conference, Sofia
2012-10-26

Blue Gene/P in the Nat. Center for Supercomputing Applications in Sofia

Overview

- Goals: why are we building a research infrastructure?
- Centers: backbone of CLARIN
- Infrastructure pillars
- Conclusion & Outlook

the Vasil Levski National Stadium with a capacity of 43,632 (Sofia)

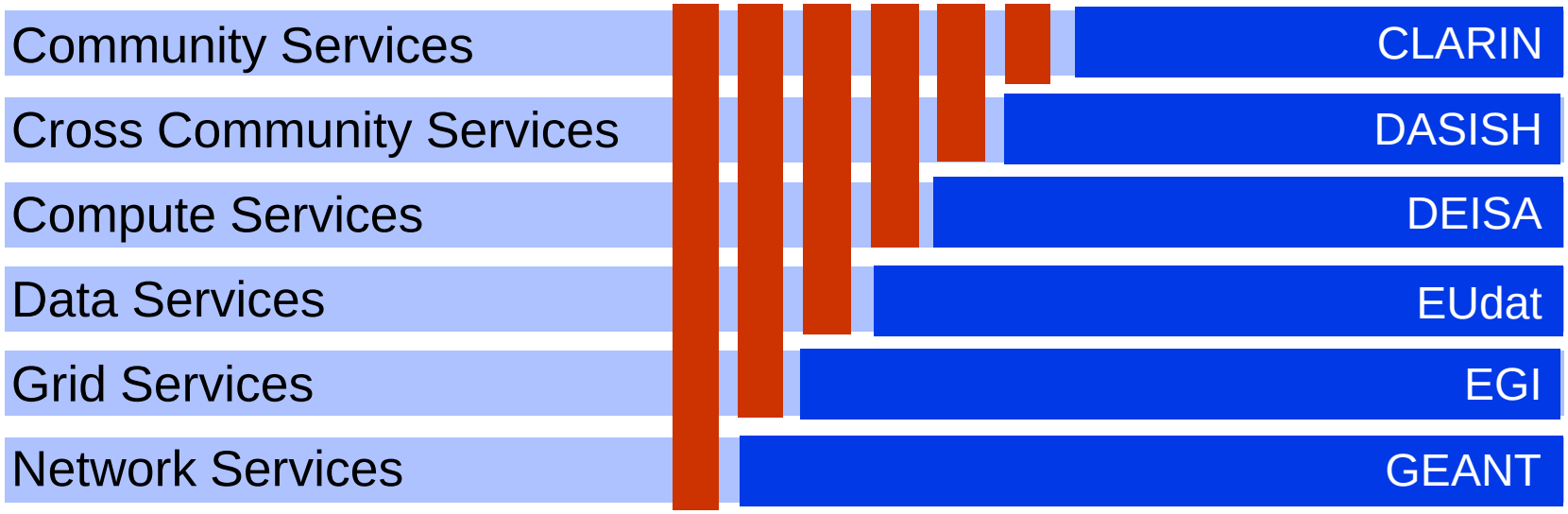
Goals of CLARIN

- Making resources and tools available
  - to a broader audience
  - in an easy-to-use and transparent way
- What and how?
  - Data: corpora, lexica, annotations, …
    - Via your internet browser
    - Or on your local computer
  - Services & software
    - Via a web applications
    - Download and install
    - Via web services (for programmers)

Goals of CLARIN

- Enhance interoperability
  - Standards & protocols
  - Semantic interoperability (~ isoCAT)
- Digital sustainability
  - Long-term availability of:
    - Data
    - Services
  - Openness (easy to find resources and to access / share them)

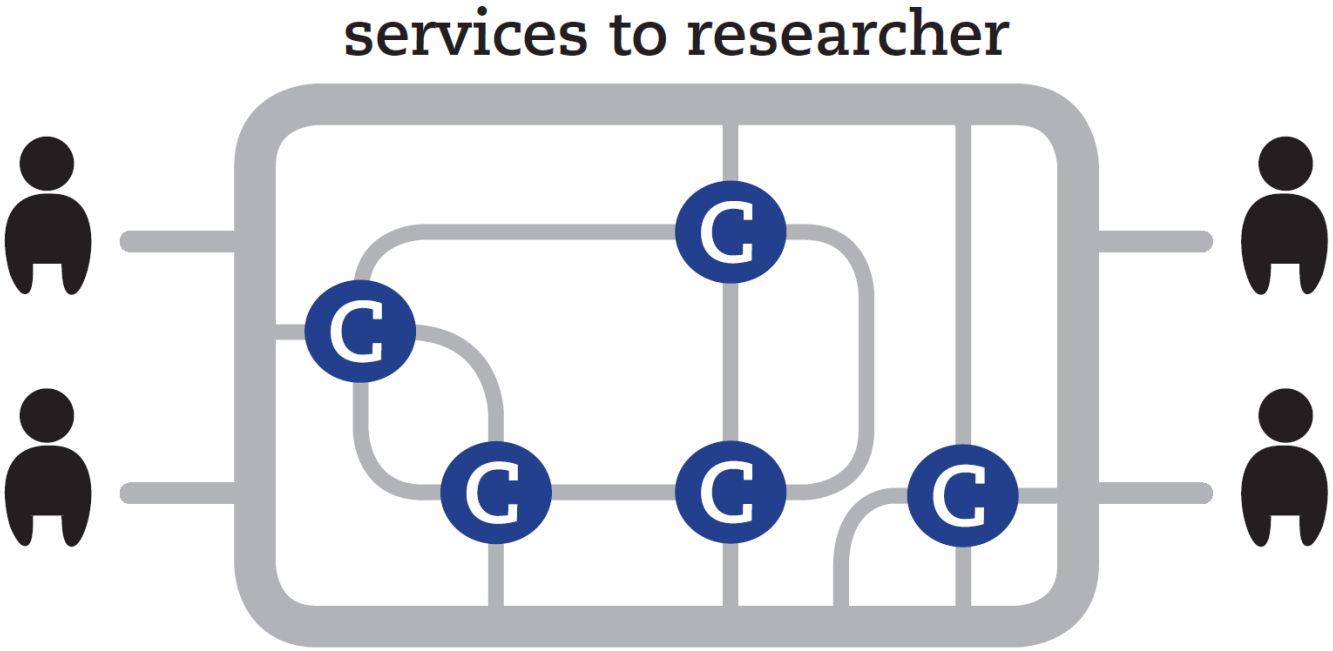
Context: an ecosystem of infrastructures



In the 14th century Tarnovo claimed to be the 3rd Rome based on its cultural influence

Means to get there

- Not one monolithic setup
- But using a well-working paradigm (think of the Internet): distributed architecture



Centers in CLARIN

- Each one with its own specialties
- With a quality assurance/check (cfr. Center Assessment Committee)
- Extendible (with other institutes/countries)
- Not ad-hoc but within a clear vision
- Update center requirements document (on its way):
  - http://www.clarin.eu/page/3542

Current center status

- For the ERIC members: 20 centers (at least one B center for each member)
- Initial list
  - center type objectives (work in progress)
  - will be maintained by the standing committee on technical centers
  - when a center indicates it is ready: assessment by Center Assessment Committee
  - starting point for the center registry

ERIC center candidates

Country/Organisation	Center Name	Type (A and B)
Austria	Clarin Center Vienna (CCV)	A+B
Bulgaria	CLARIN-BG, IICT-BAS	B
Czech Republic	LINDAT-CLARIN	A+B
Denmark	UCPH	B
Dutch Language Union	(to be added)	B
Estonia	Center for Estonian Language Resources (CELR)	B
Germany	IDS	A+B
Germany	UTU	A+B
Germany	BAS	B
Germany	BBAW	B
Germany	HZSK	B
Germany	UdS	B
Germany	ASV	B
Germany	IMS	B
Germany + Netherlands	MPI	A+B
Netherlands	Meertens Instituut	A+B
Netherlands	INL	B
Netherlands	Huygens Instituut	B
Netherlands	DANS	B
Poland	LTC CLARIN-PL	B

Other center candidates

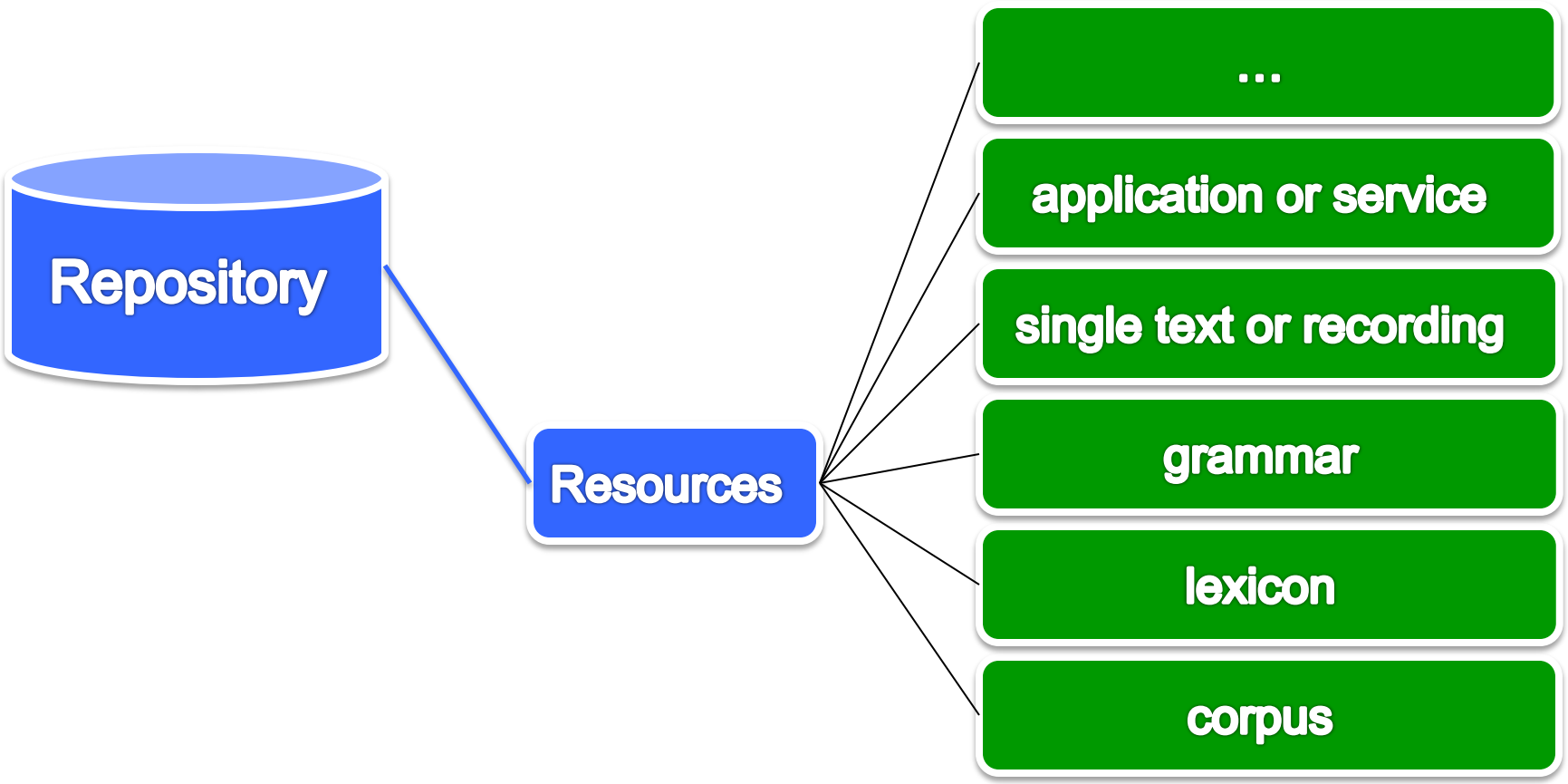
Country/Organisation	Center Name	Type (A and B)
Croatia	Zagreb University Centre for Computer Science	A
Finland	CSC	A
France		2 x A + B
Greece		A or B
Latvia	IMCS	B
Lithuania		B
Norway	National Library & Uninett	A
Norway	others	4 x B
Hungary		
Italy		
Malta		
Romania		
Slovakia		
Slovenia		
Spain		
Sweden		
UK		
EURAC		

Central Sofia Market Hall, Naum Torbov, 1911

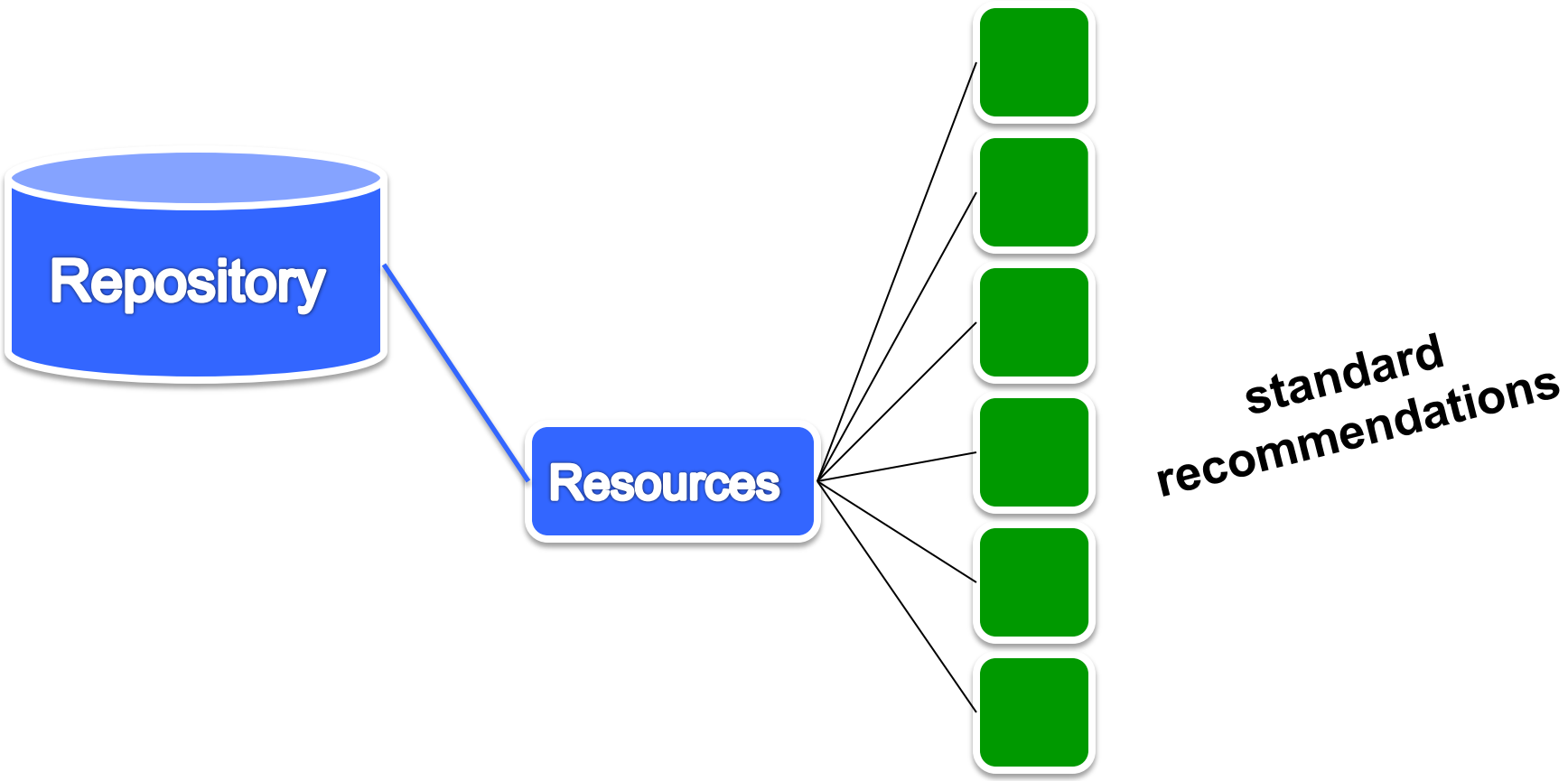
Infrastructure architecture

Repository = sustainable store at a CLARIN center that can be accessed via the internet

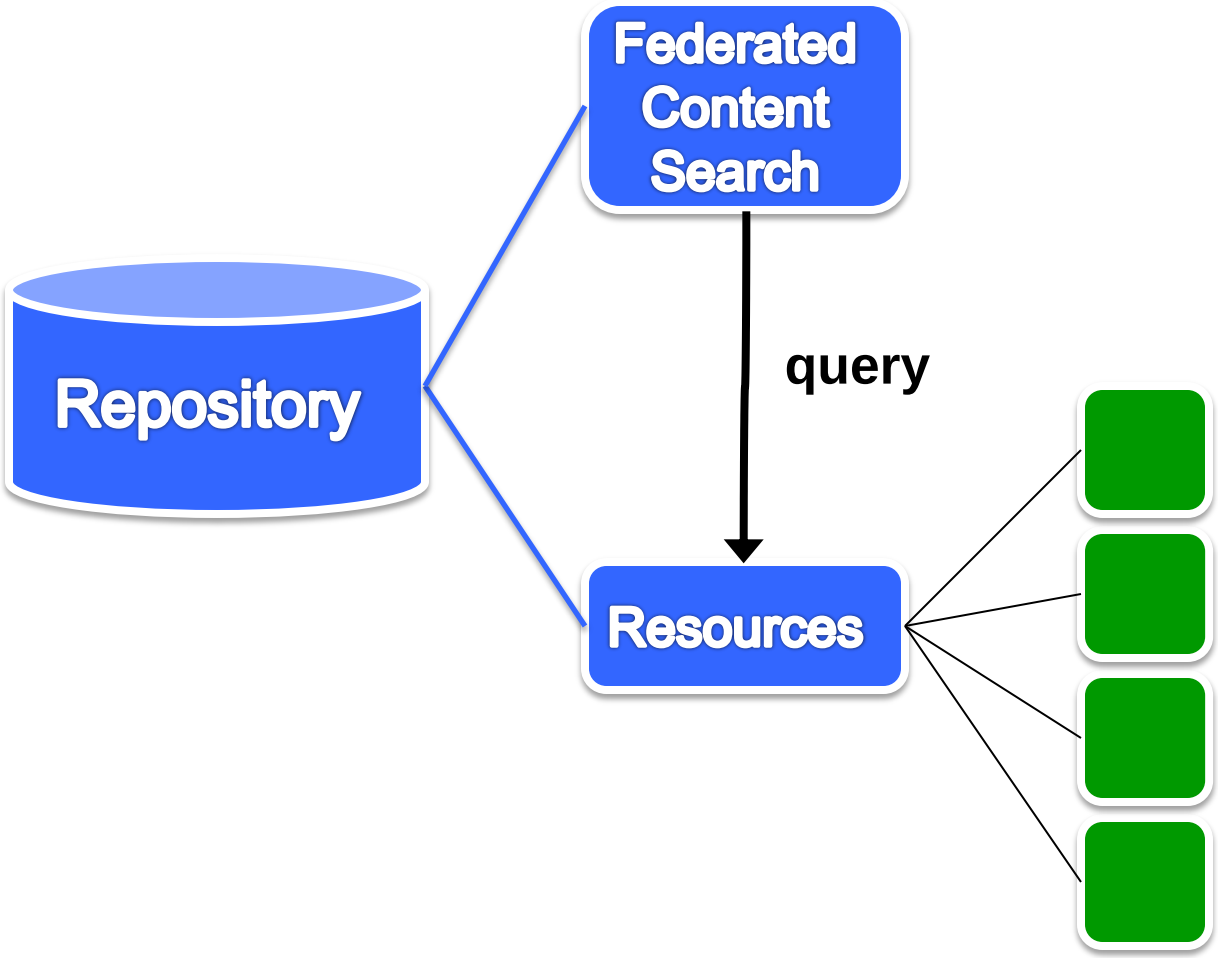
Infrastructure architecture



Infrastructure architecture

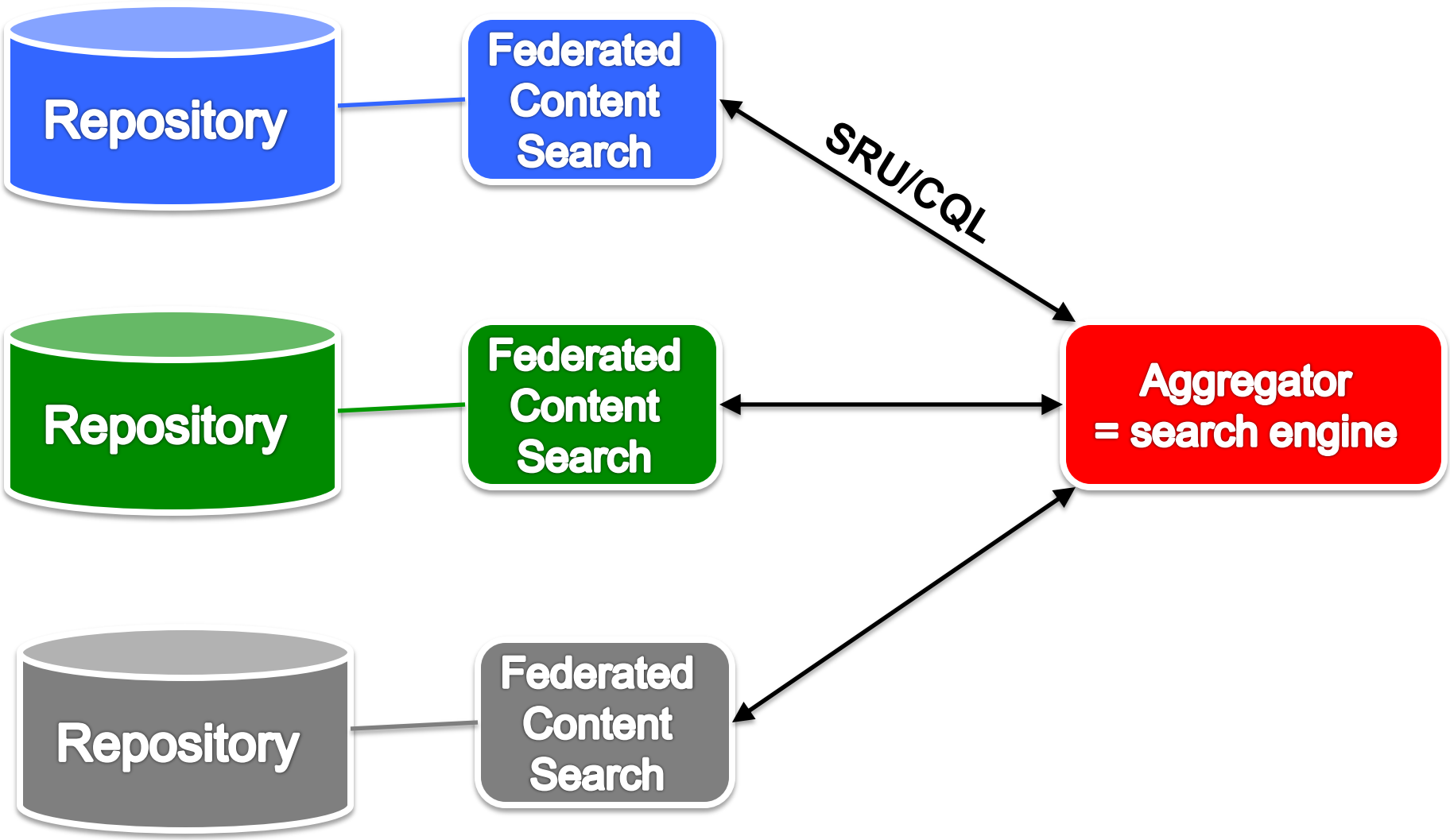


Infrastructure architecture

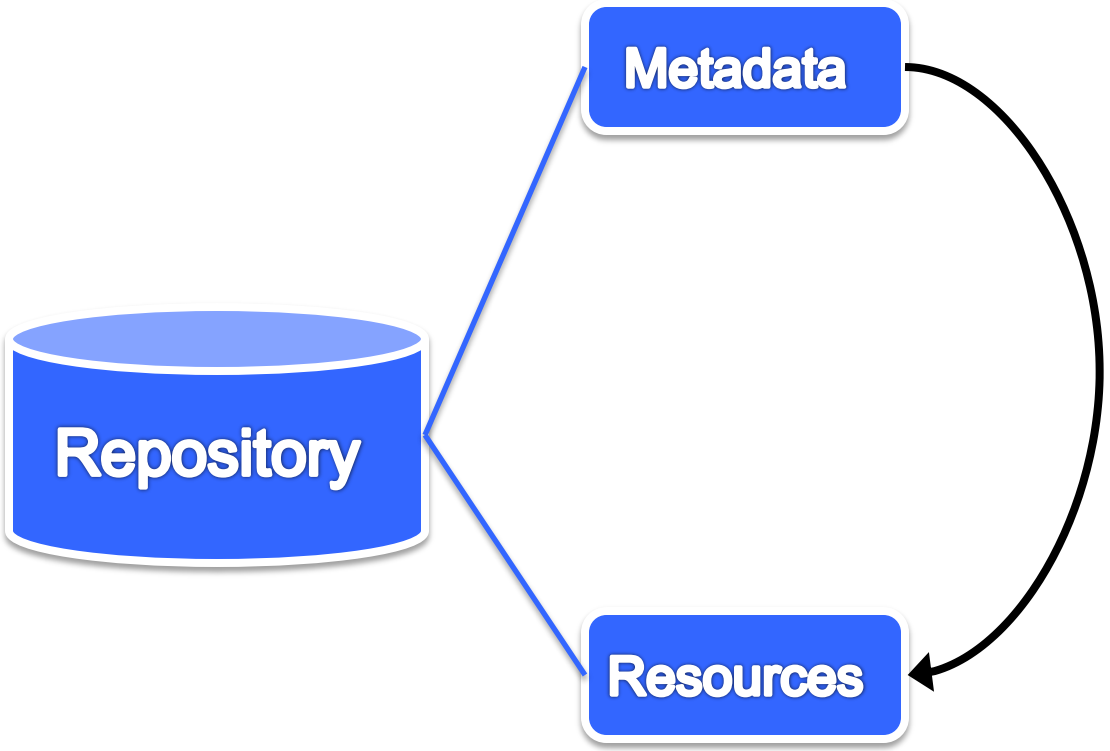


Note: all resources stay at repository!

Infrastructure architecture

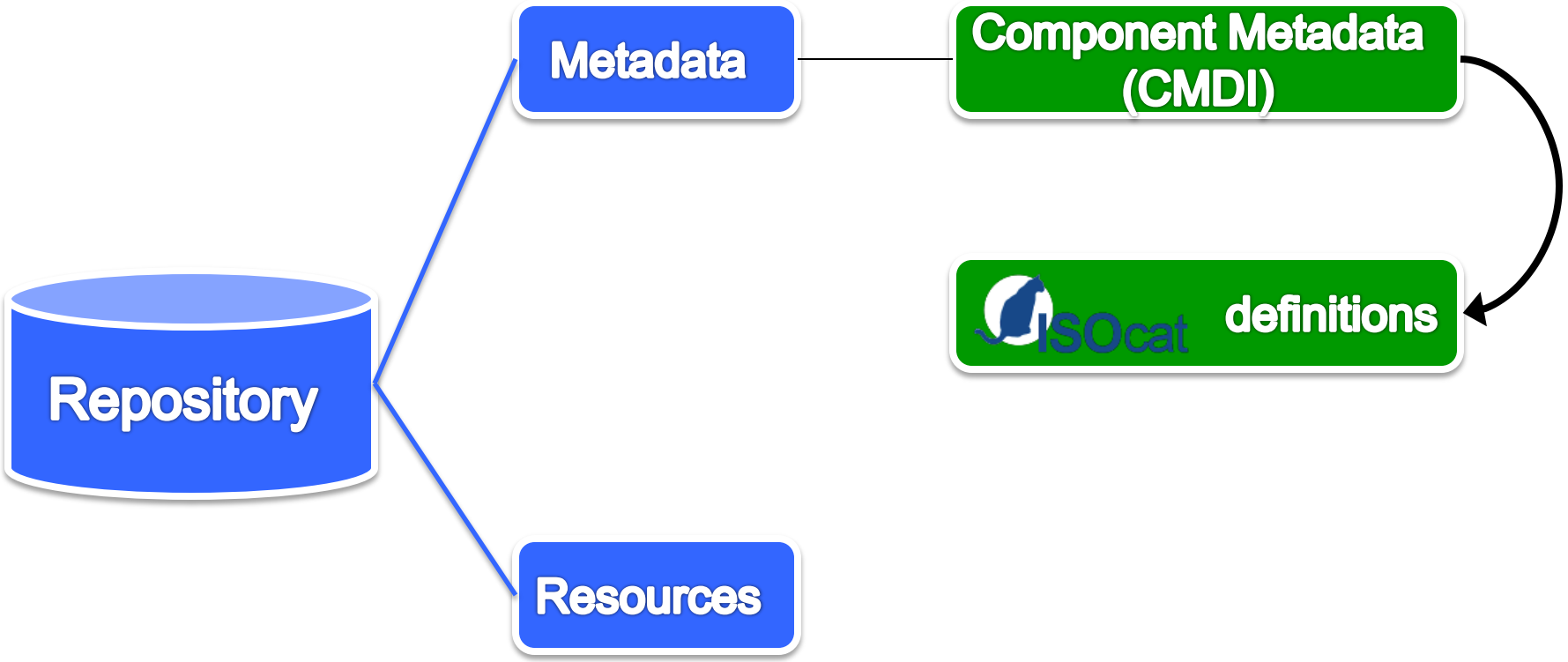


Infrastructure architecture

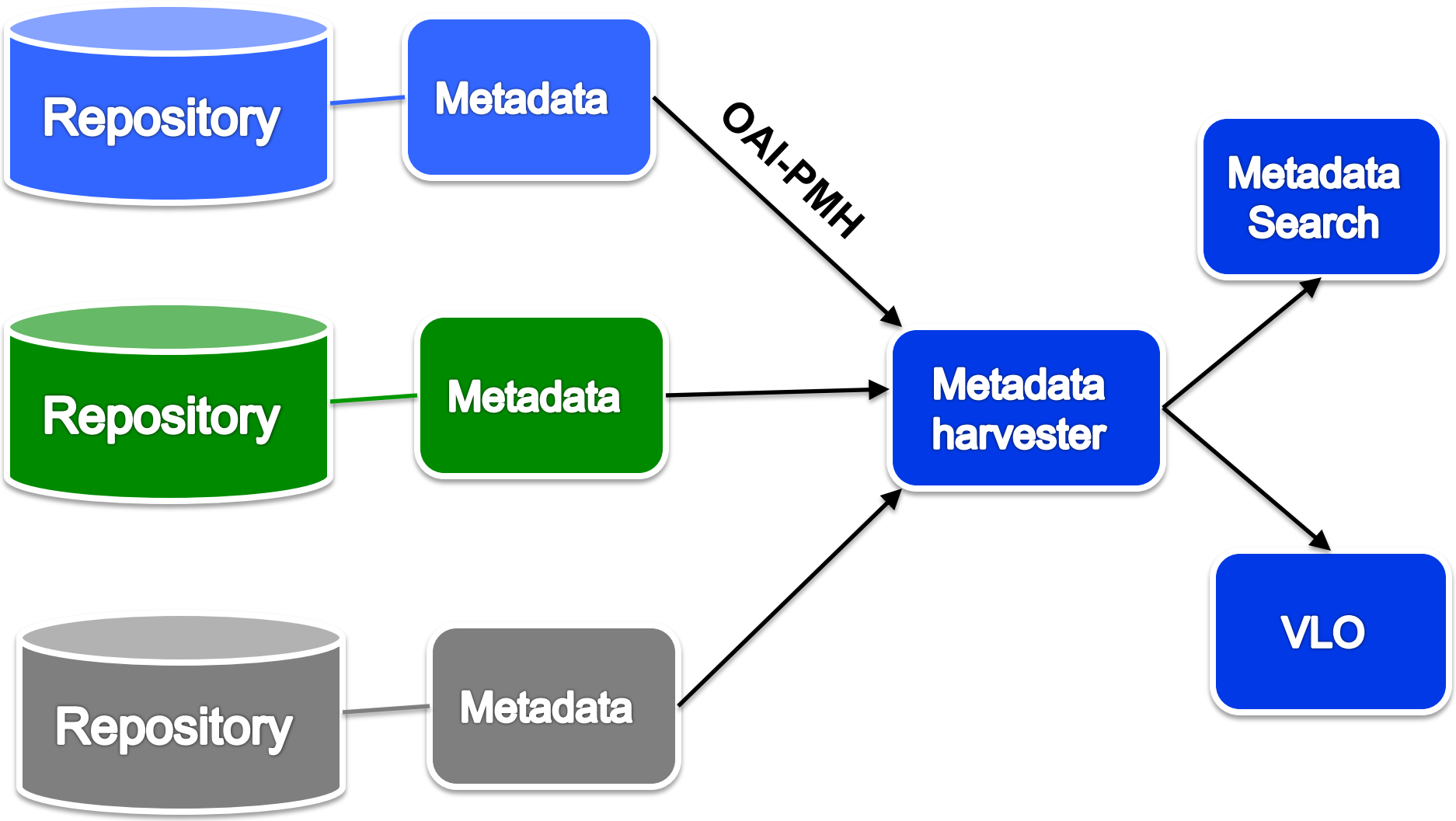


Human- and computer-readable information

Infrastructure architecture



Infrastructure architecture

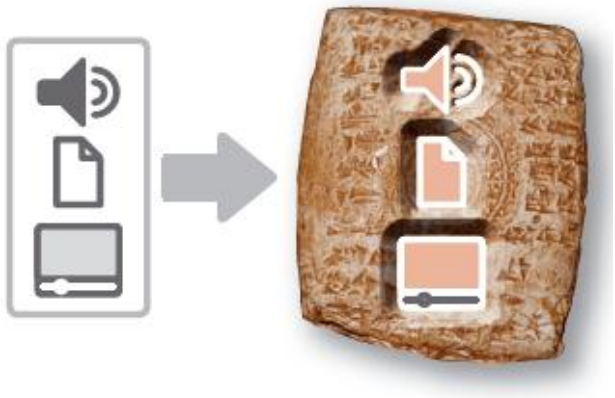


Infrastructure Pillars

Pobiti Kamani is a rock phenomenon located in Varna Province, Bulgaria

Pillars: repositories

- Each A+B/B/C center needs one



- Available Services:
  - OAI-PMH metadata harvester
  - Center registry:
    - technical information
    - contact information
  - Metadata exploration:
    - Virtual Language Observatory
    - others are being implemented

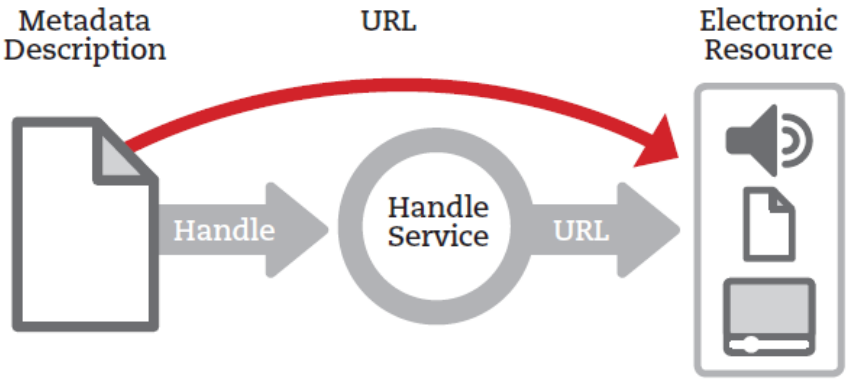
Pillars: repositories

- Metadata creation tools
- documentation and experiences

- Issues:
  - granularity + modeling in general > need for information exchange among centers
  - if interest exists: workshop on how repositories are being implemented in CLARIN

Pillars: Persistent Identifiers

- Sustainable citations
- with **persistent identifiers**
- Requirement for B centers



- Available Services:
  - European Persistent Identifier Consortium (EPIC) service
  - relying on the well-established handle protocol
    - high availability (mirrors) + scalability
  - part identifiers
    - new feature: address part of resources

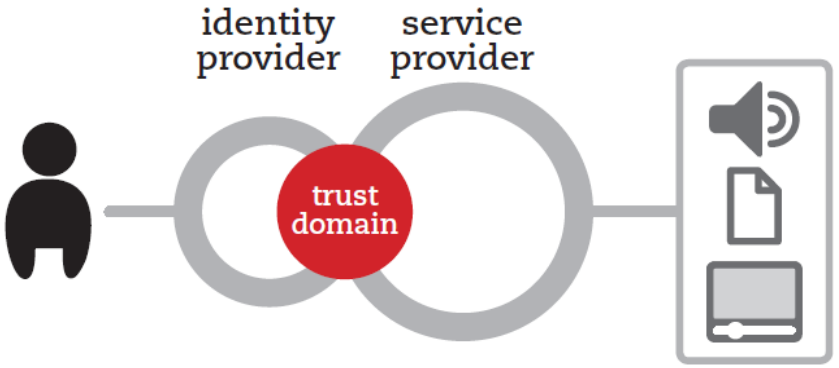
Pillars: Persistent Identifiers

- Expected: EPIC API version 2 (beta available)
  - manage own prefix
  - batch requests

- Issues: delay of of this new version – seems to be resolved by now

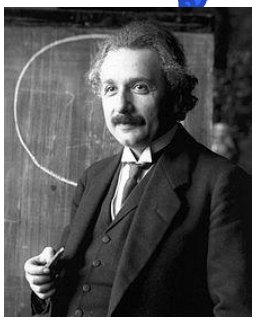
Pillars: Federated Identity

- Use your own (institutional) login to access password-protected resources
- Requirement for B centers



- Available Services:
  - Service Provider Federation
  - CLARIN Identity Provider
  - Easy-to-use discovery service

I would like to use a CLARIN service…



1. wants to access

7. uses

6. redirects to resource for authorization check

Service Provider

Identity Provider

5. User enters credentials

3. User selects IdP

2. redirects to

4. redirects to

DiscoJuice version 1.0

Discovery Service

Pillars: Federated Identity

- CLARIN's Service Provider Federation (SPF)
  - agreement: CLARIN ERIC becomes responsible legal entity
  - further extension of the SPF
    - Connect more CLARIN services to it (currently 8)
    - Connect to more national identity federations (currently 6)
- Steps you can already take:
  - Setup a Service Provider at your center
  - Join your national Identity Federation
  - Connect to the CLARIN Identity Provider

Pillars: Federated Identity

- At the same time we are exploring the options to make use of eduGAIN
- Pro:
  - you only need to join your national Identity Federation
  - no need for SPF (less administration)
- Con:
  - Opt-in: only institutions that agreed to use eduGAIN can use Services
  - E.g. Czech Republic: of all 21 IdPs only 3 have joined eduGAIN, so we would need to convince 18 organisations to do so

Pillars: Federated Identity

- Next steps (issues / solutions)
  - See if we can smoothen the opt-in process by signing a Code of Conduct
  - Look at the experiences of CZ and NL with the opt-in process
  - Continue to extend the Service Provider Federation as it currently seems to be the most certain way to reach our users

Pillars: web services

- Possibility to
  - create chains of language resource processing webservices on the fly
  - call web services from local applications
- Available Services:
  - WebLicht (resource processing)
  - Weblicht associated applications
  - CMDI web service core model
  - Several web service plugins for local applications, e.g.:
    - WebMAUS
    - AVAteCH
- Outstanding issue: combining AAI and web services (trust delegation)

Pillars: federated content search

- Federated Content Search (in development)
  - www.clarin.eu/fcs
- Available services:
  - validity checker
  - aggregator
- Details in tomorrow's session

Pillars: registries

- ISOcat Data Category Registry
  - http://www.isocat.org/
- Recently shibbolized: you can now login with your existing account
- About 700 entries for the metadata profile
- CLARIN-NL‘s ISOcat coordinator (Ineke Schuurman) has done a lot of curation / quality check
  - please contact her before ingesting a lot of new data categories
- Forum available for ISOcat-related questions
- Still there are outstanding issues
  - idea to organize a special session/workshop on this topic

Pillars: registries

- Relation Registry: alpha-Version available
  - http://lux13.mpi.nl/isocat/relcat/site/index.html
- Schema Registry: alpha-Version available
  - http://lux13.mpi.nl/isocat/schemacat/site/index.html
- Vocabulary registry
  - OpenSkos / CLAVAS
  - SMC - Semantic Mapping Component - is a module within the CMDI framework of the CLARIN infrastructure. It provides mappings between fields in heterogeneous metadata descriptions combining information from Component Registry, Data Category Registries (ISOcat, dublincore) and Relation Registry.

Pillars: registries

- Center registry: beta version available

CLARIN

Center Register

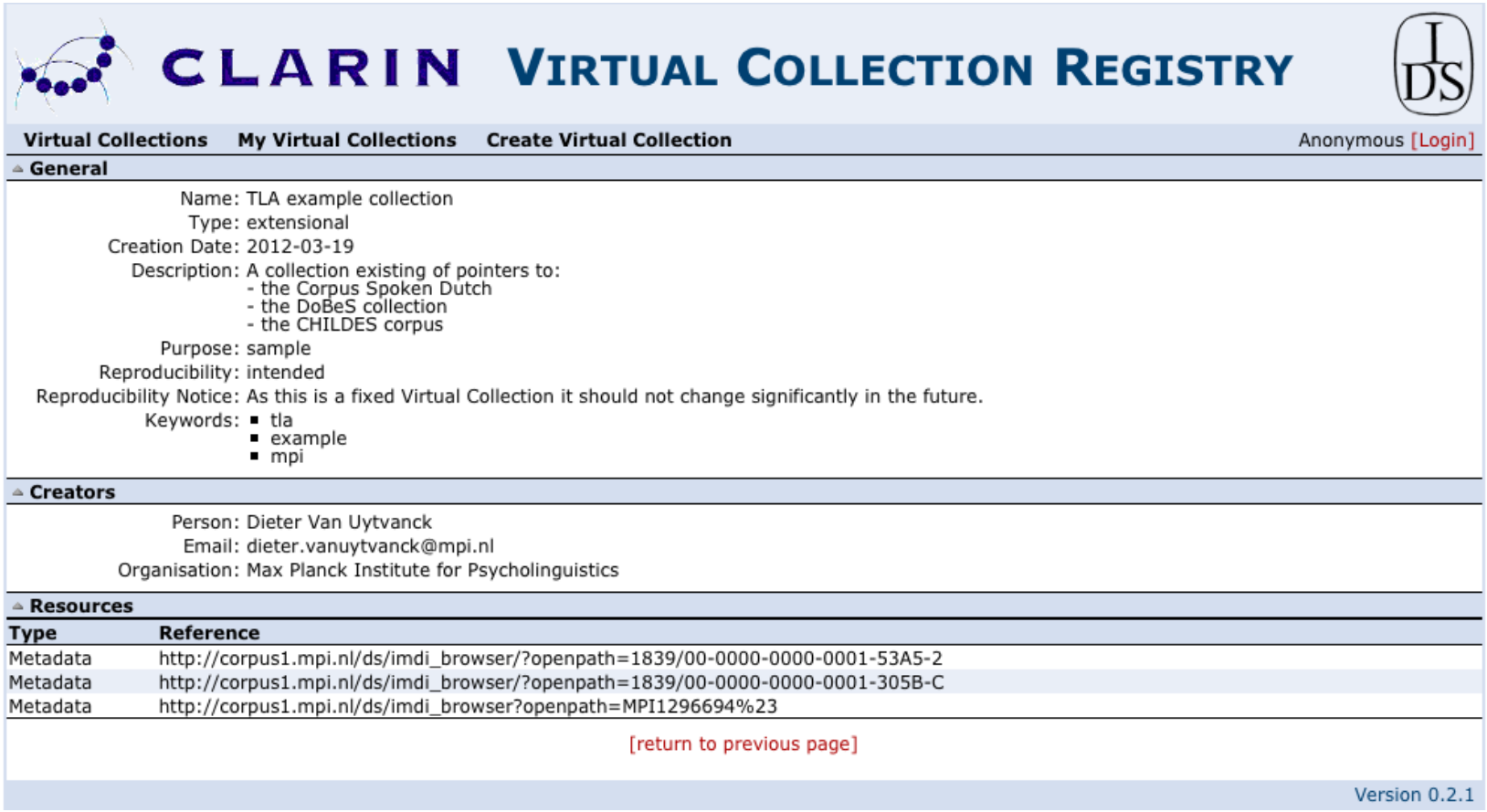
Center	Repository	CenterStatus	Country	web services and Metadata	AAI	PID	assessment
Eberhard Karls Universität Tübingen	Fedora Commons	Aiming for B	DE	CMDI	One Service Provider available	handle via EPIC	Planning on implementing Data Seal of Approval
ASV Leipzig	Fedora Commons	Aiming for B	DE	CMDI , OLAC	One service provider available	handle via EPIC	Planning on implementing Data Seal of Approval
Bayerisches Archiv für Sprachsignale	own system	Aiming for B	DE	CMDI	Currently no Service Providers available	handle via EPIC	none
Institut für Deutsche Sprache	COSMAS II, OWID, Subversion	Aiming for A	DE		IDP and SPs available	handle (own server)	currently none
MPI for Psycholinguistics	IMDI + Lamus	Aiming for A	NL	CMDI	Several Service Providers available	handle (own server)	Data Seal of Approval granted: https://assessment.datasealofapproval.org /assessment_48/seal/html
Universität des Saarlandes	Fedora Commons	Aiming for B	DE	CMDI , OLAC	Not available yet	handle via EPIC	
Berlin-Brandenburg Academy of Sciences and Humanities	Fedora Commons	Aiming for B	DE	OLAC	Service Provider available	urn:nbn	
Hamburger Zentrum für Sprachkorpora (HZSK)	Fedora Commons	Aiming for B	DE		No Service Providers available	handle via EPIC	none
IMS, Universität Stuttgart	Fedora Commons	Aiming for B	DE		Currently no Service Providers available	handle via EPIC	Planning on implementing Data Seal of Approval

up

©The Center Registry has been developed by CLARIN-D

Pillars: registries

- Virtual Collection registry: alpha version available
- Need for use cases and testing



CLARIN VIRTUAL COLLECTION REGISTRY

Virtual Collections | My Virtual Collections | Create Virtual Collection | Anonymous [Login]

General

Name: TLA example collection
Type: extensional
Creation Date: 2012-03-19
Description: A collection existing of pointers to:
- the Corpus Spoken Dutch
- the DoBeS collection
- the CHILDES corpus

Purpose: sample
Reproducibility: intended
Reproducibility Notice: As this is a fixed Virtual Collection it should not change significantly in the future.
Keywords:
- tla
- example
- mpi

Creators

Person: Dieter Van Uytvanck
Email: dieter.vanuytvanck@mpi.nl
Organisation: Max Planck Institute for Psycholinguistics

Resources

Type	Reference
Metadata	http://corpus1.mpi.nl/ds/imdi_browser/?openpath=1839/00-0000-0000-0001-53A5-2
Metadata	http://corpus1.mpi.nl/ds/imdi_browser/?openpath=1839/00-0000-0000-0001-305B-C
Metadata	http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI1296694%23

[return to previous page]

Version 0.2.1

Pillars: monitoring

- Keep an (automated) eye on the state of the infrastructure components
- Some (nagios/icinga) plugins already available
  - SRU/CQL
  - OAI-PMH
  - valid SPF metadata
- Have monitoring in place for all CLARIN centers

Maslen nos on the Black Sea coast

Conclusions

- Setting up a distributed infrastructure as CLARIN means a lot of hard work…
- … but at the same time it is a unique chance (2nd ERIC!)
- Luckily we are together in this:
  - exchange experiences
  - share services
- No need to start from scratch:
  - CLARIN preparatory phase foundations
  - know-how gathered in various national CLARIN projects
- Let's start integrating!

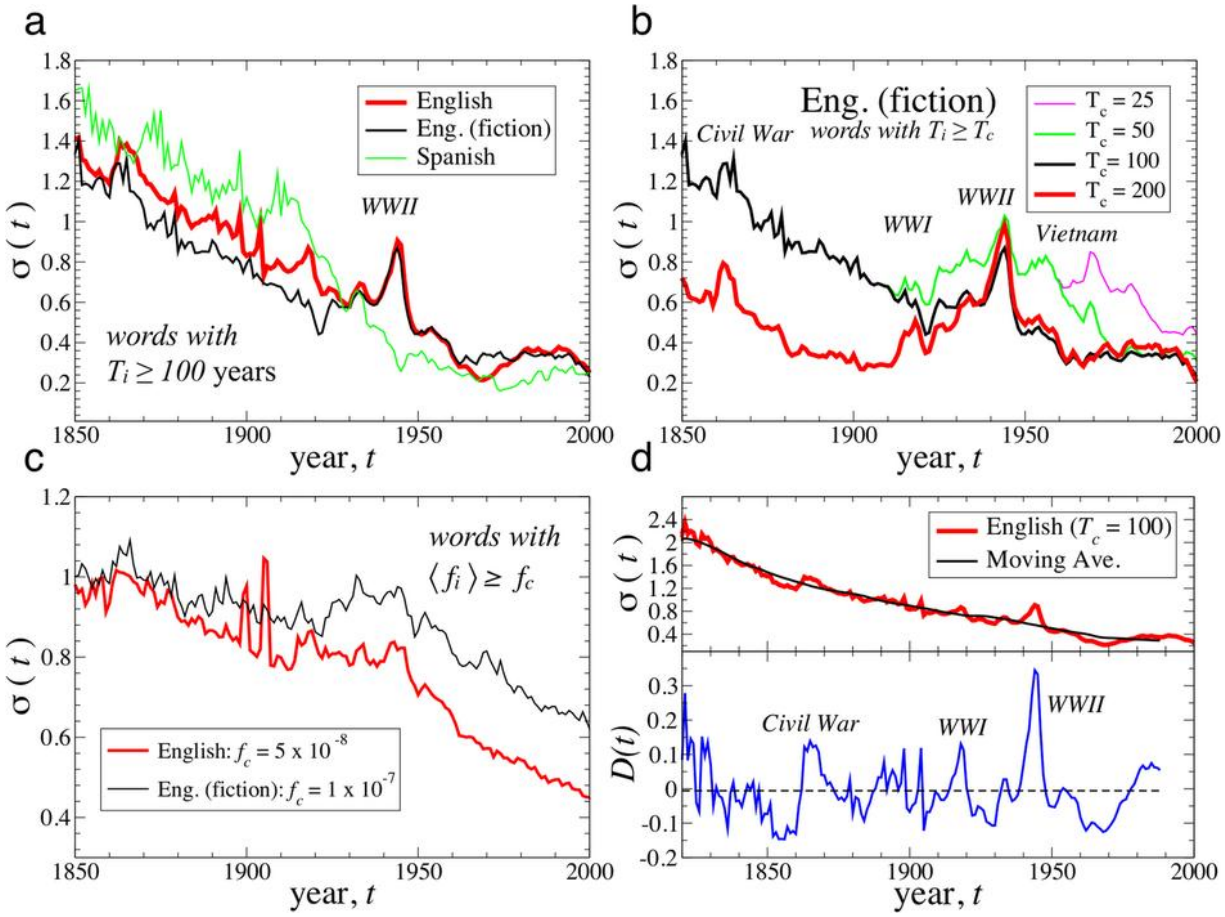
More information

- FAQs: http://www.clarin.eu/faq
- Short Guides: http://www.clarin.eu/node/1474
- Reference Manual: http://www.clarin.eu/node/3484
- Handbook CLARIN-D (in the pipeline)

Bulgarian folk dance *horo*.

Some use cases

- Mass text analysis (Petersen et al., 2012): doi:10.1038/srep00313

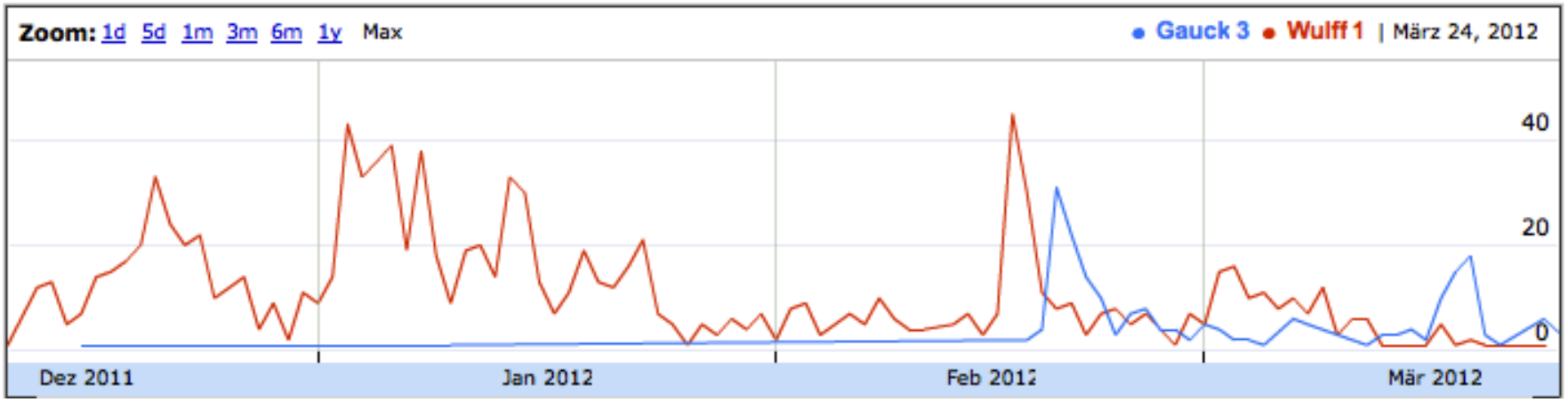


Some use cases

- Mass text analysis: Dynamic Corpus Analyzer

Name	Number
Gauck	212
Wulff	1131

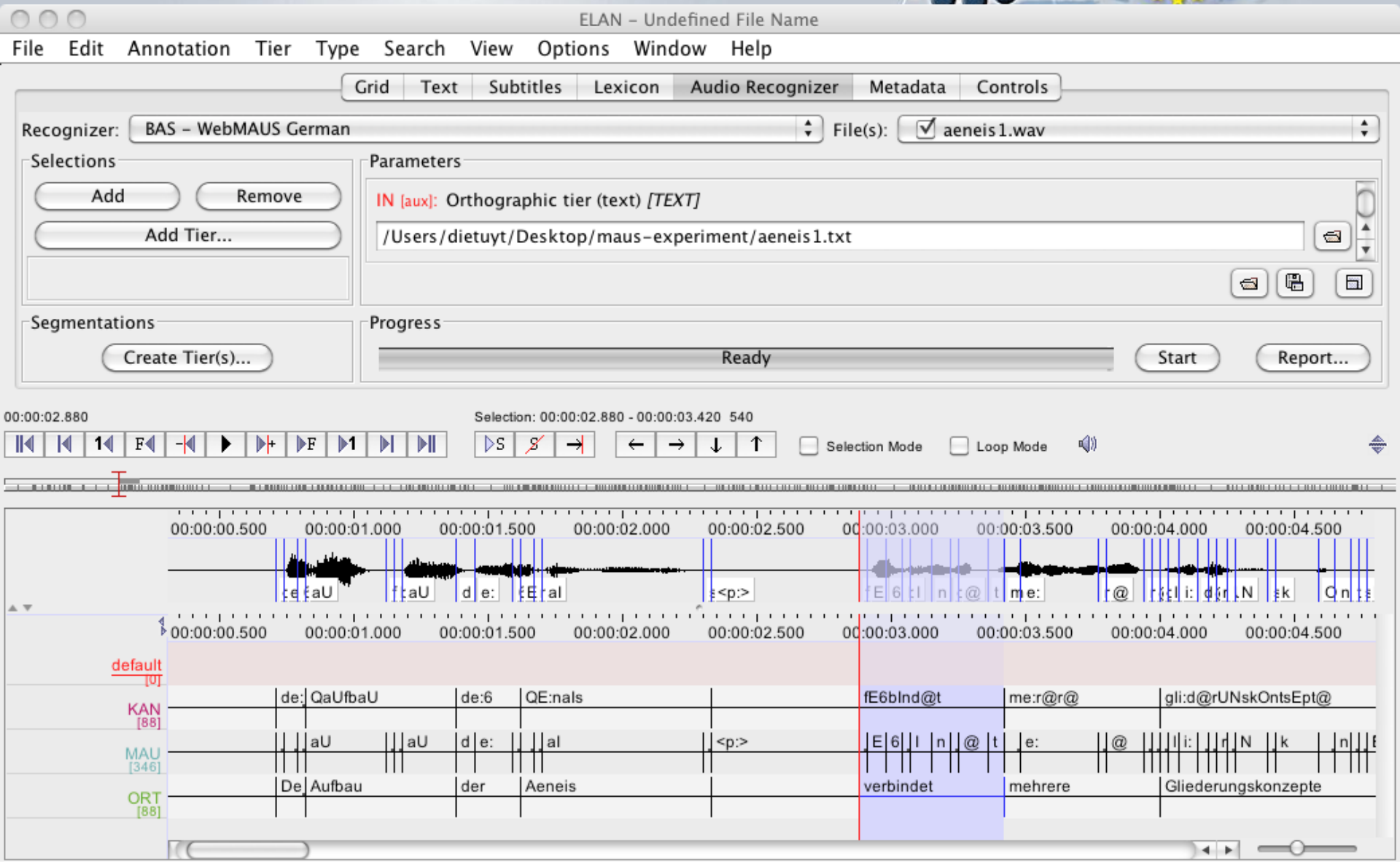
Export



Some use cases

- Automatic creation of phonetic transcriptions + alignment with webMAUS (BAS) from ELAN

Some use cases



Some use cases

- Enhanced publications: mpi.nl/trobriand