ÖAW AUSTRIAN ACADEMY OF SCIENCES

ACDH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES

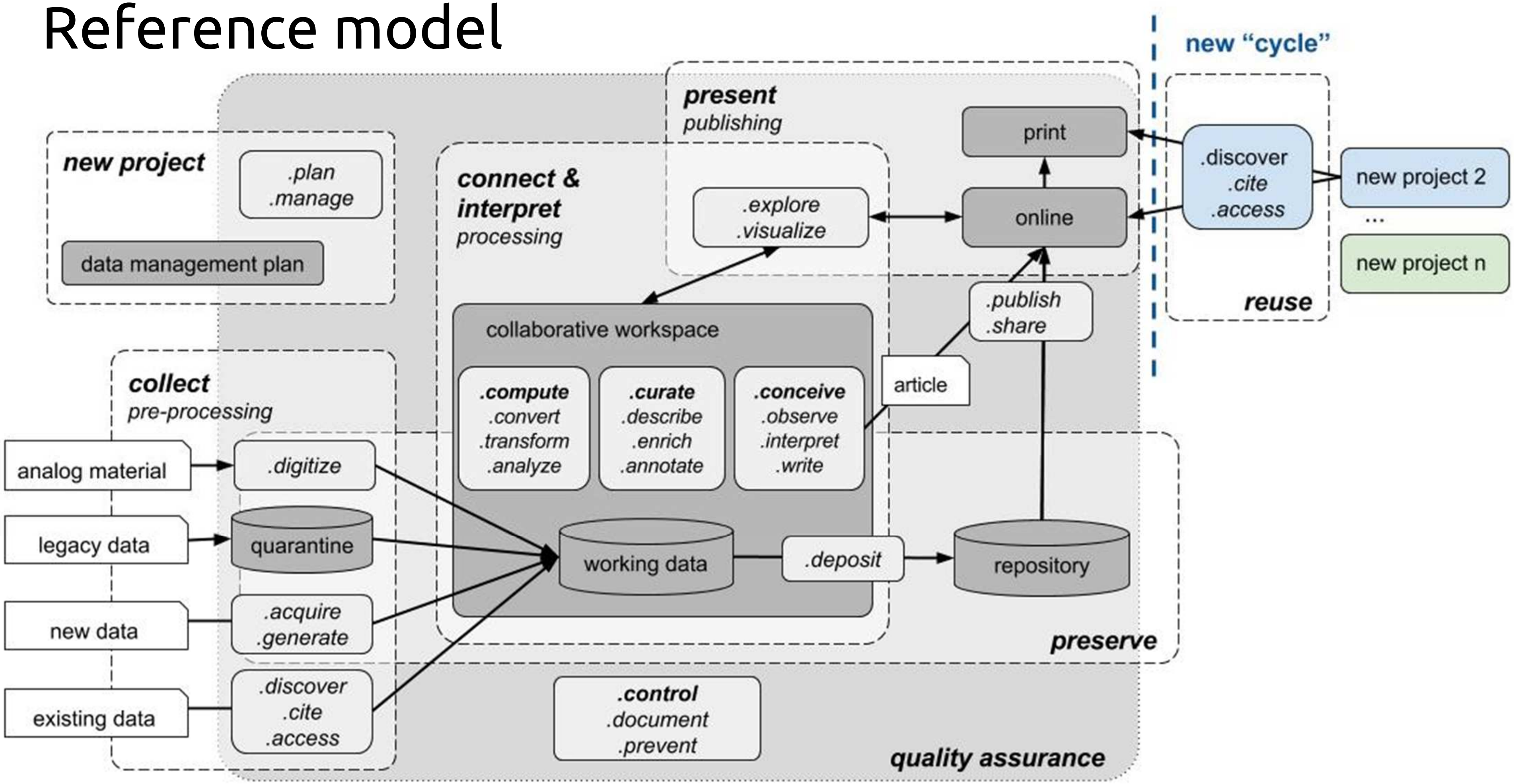
Metadata for the Humanities – a use case from ÖAW

Matej Ďurčo @ e-Infrastructures workshop on Metadata, 2016-06-20

Index

- general considerations
- CMDI
- Use cases - ACDH-OEAW
- Research infrastructures / Aggregators

Reference model



metadata

- “data about data”
- administrative/structural/technical/provenance/descriptive
- Location:
  - separate: XML-files (CMDI, DC), Databases
  - embedded: JPG, TEI, …
- Metadata/data/annotation distinction?

  Especially with RDF or in relational databases
- Describe structure

  hierarchical and other relations
- Interoperability

  Be able to exchange data across systems (keeping the semantics)
- Single sourced

  Use the most comprehensive format and derive the others
- explicate the model

  DDL, DDT, ODD, XSD

metadata formats

- DC – dublincore (elements/ DCMI terms)
- METS/MODS (LoC)
- ALTO - Analyzed Layout and Text Object (technical metadata for OCR, LoC)
- CMDI – Component Metadata Infrastructure (CLARIN)
- EDM -Europeana Data Model
- DCAT – Data Catalog Vocabulary (w3c)
- ORE – Object Reuse and Exchange
- EAC-CPF, EAD, EAG, ISAD, ISAR, ISDIAH - Archival Holdings
- DDI –Data Documentation Initiative (DDI) statistical and social science data.
- …

x Vocabularies / Classification schemes

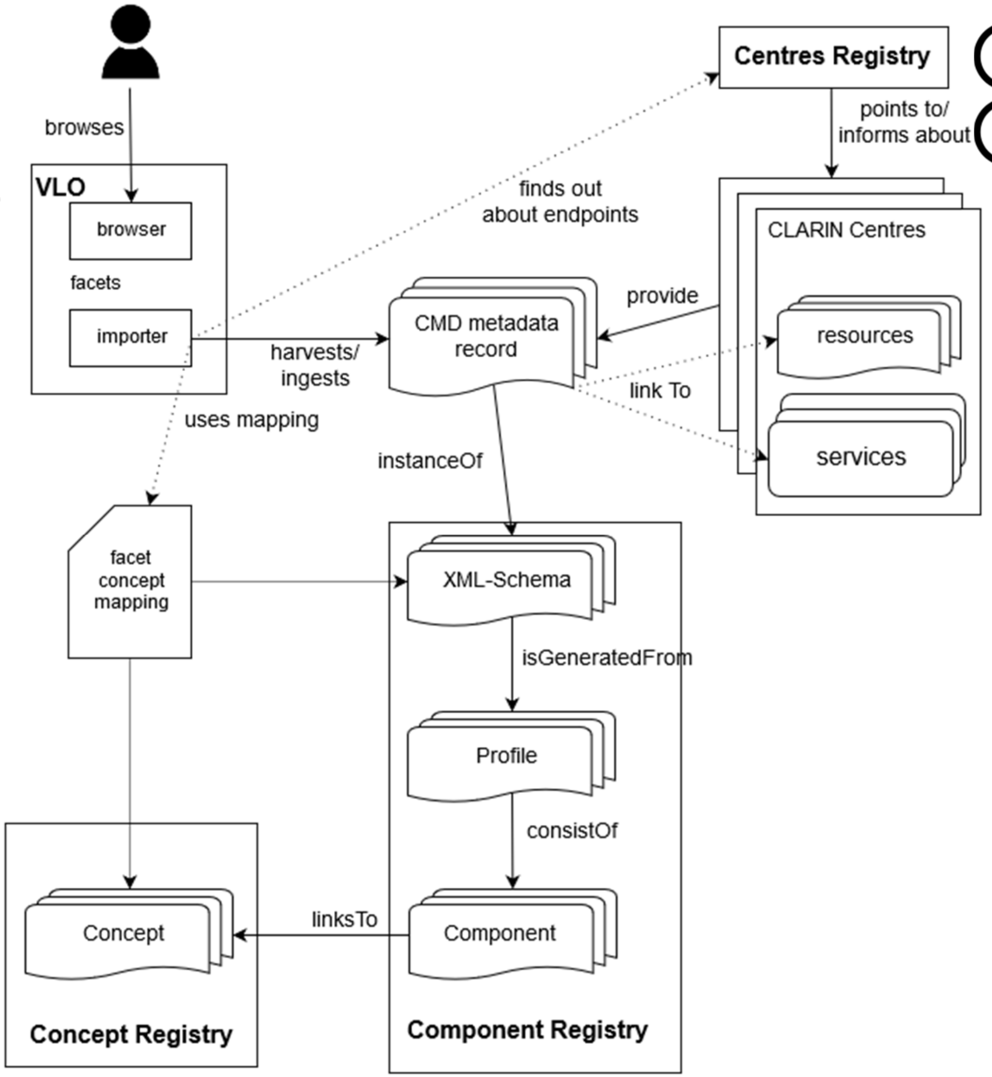
SKOS – Simple Knowledge Organisation System (w3c) as lingua franca;
MADS

CMDI

Component Metadata Infrastructure

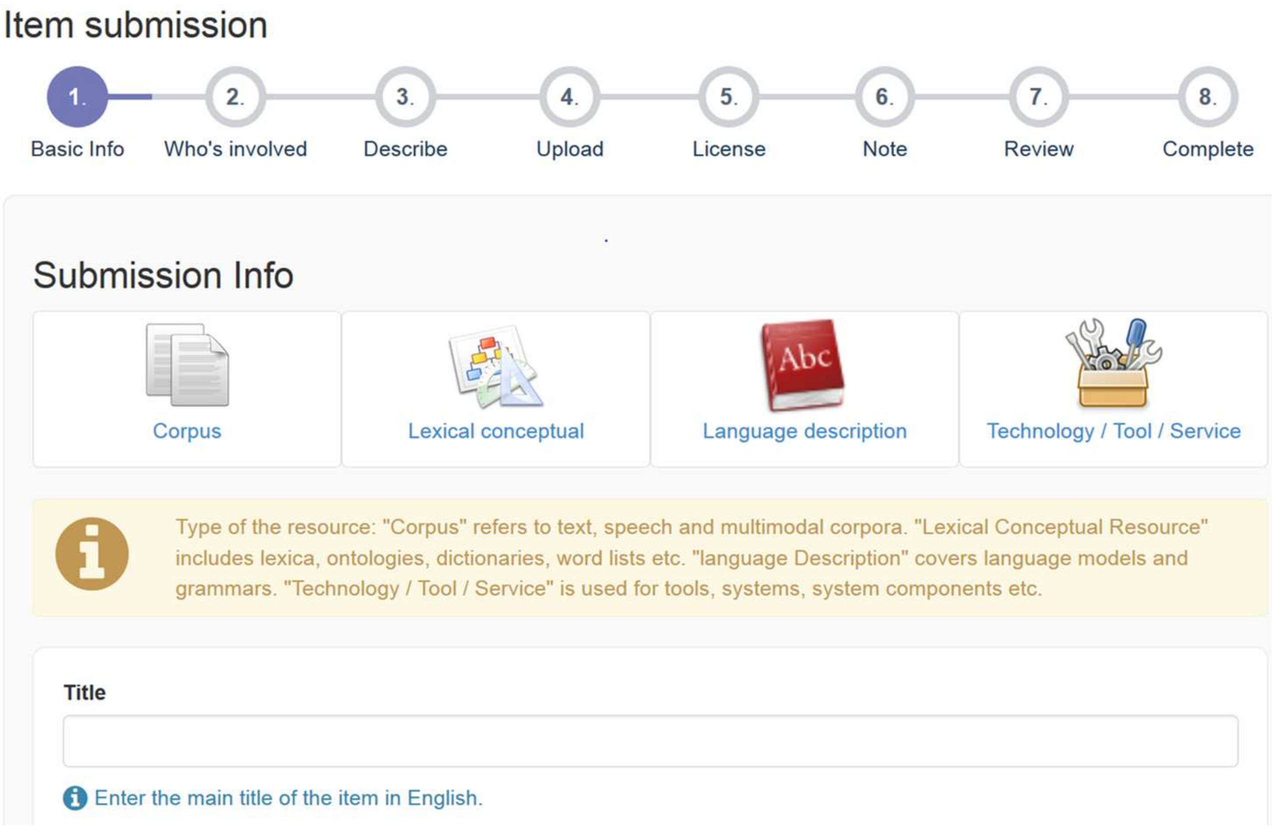
http://clarin.eu/cmdi

- not a single schema, but a metadata framework for creating custom schemas
- reusable components describing individual aspects
- concept-based semantic interoperability
- Component Registry
  ~ 1200 components,
  ~ 200 profiles
- Concept Registry
  > 3.000 concepts
- CLARIN VLO - MD catalogue
  ~ 1 Mio. records



metadata authoring

- relational database
- generic XML editors
  - oXygen
- specialized tools
  - ARBIL, COMEDI
- submission form
  - PHAIDRA
  - LINDAT



Item submission

Submission Info

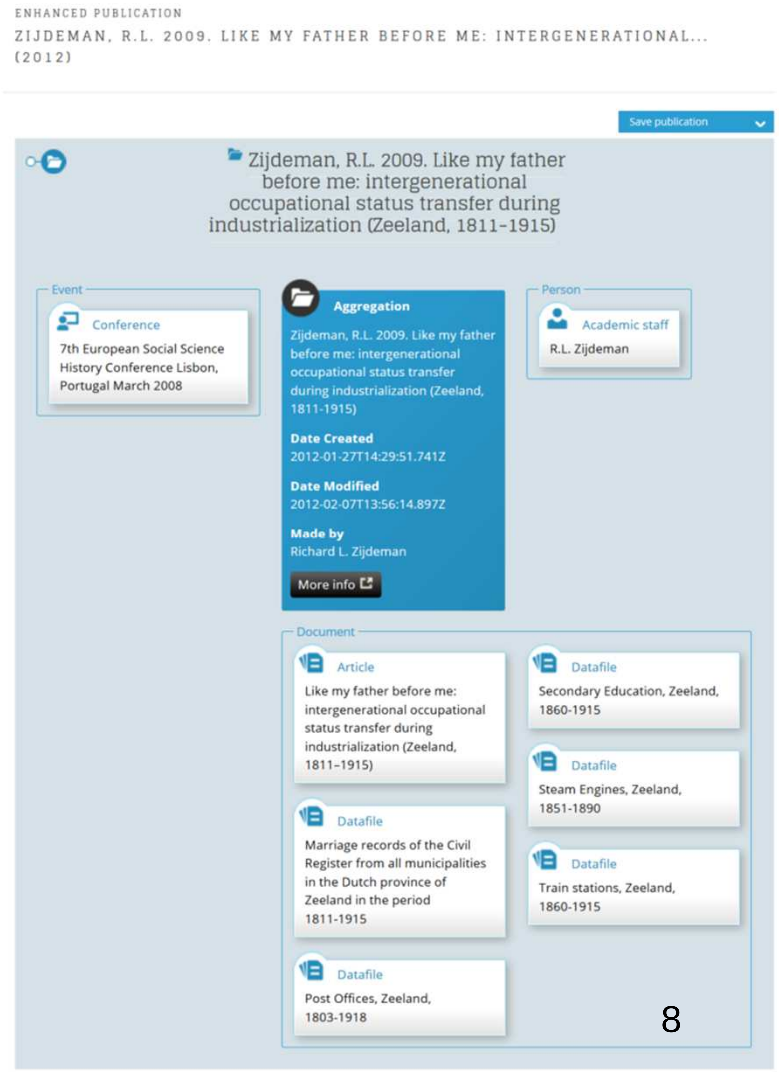
Type of the resource: "Corpus" refers to text, speech and multimodal corpora. "Lexical Conceptual Resource" includes lexica, ontologies, dictionaries, word lists etc. "language Description" covers language models and grammars. "Technology / Tool / Service" is used for tools, systems, system components etc.

Title

Enter the main title of the item in English.

.publish | .discover

- disseminate metadata over many channels linking back to data in the repository
- Metadata catalogues / Aggregators
  - CLARIN VLO (~ 1 Mio. Language resources)
  - Europeana (52 Mio. Objects ?)
  - recherche-isidore.fr
  - OpenAIRE (13 Mio. pubs, 17.000 datasets, ~ 6.500 repos)
  - OLAC - Open Language Archives Community
  - JSTOR - http://www.jstor.org/
  - narcis.nl@DANS – (1,23 Mio. publications, ~150.000 datasets, 1710 **enhanced publications**)
- OAI-PMH – protocol for metadata harvesting
  - provider exposes metadata via endpoint
  - harvester regularily fetches metadata
  - one Registry of data providers



ENHANCED PUBLICATION

ZIJDEMAN, R.L. 2009. LIKE MY FATHER BEFORE ME: INTERGENERATIONAL... (2012)

Zijdeman, R.L. 2009. Like my father before me: intergenerational occupational status transfer during industrialization (Zeeland, 1811-1915)

Event
Conference
7th European Social Science History Conference Lisbon, Portugal March 2008

Aggregation
Zijdeman, R.L. 2009. Like my father before me: intergenerational occupational status transfer during industrialization (Zeeland, 1811-1915)
Date Created
2012-01-27T14:29:51.741Z
Date Modified
2012-02-07T13:56:14.897Z
Made by
Richard L. Zijdeman

Person
Academic staff
R.L. Zijdeman

Document

Article
Like my father before me: intergenerational occupational status transfer during industrialization (Zeeland, 1811–1915)

Datafile
Marriage records of the Civil Register from all municipalities in the Dutch province of Zeeland in the period 1811-1915

Datafile
Post Offices, Zeeland, 1803-1918

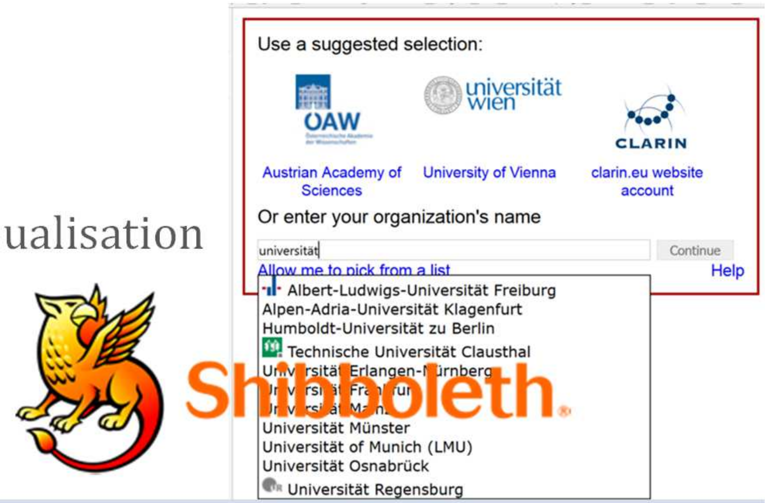
Datafile
Secondary Education, Zeeland, 1860-1915

Datafile
Steam Engines, Zeeland, 1851-1890

Datafile
Train stations, Zeeland, 1860-1915

.access

- raw data vs. search endpoint vs. application vs. visualisation
- landing page
- Restrictions (License and availability)
  - CLARIN License categories (VLO)



Use a suggested selection:

Austrian Academy of Sciences

University of Vienna

clarin.eu website account

Or enter your organization's name

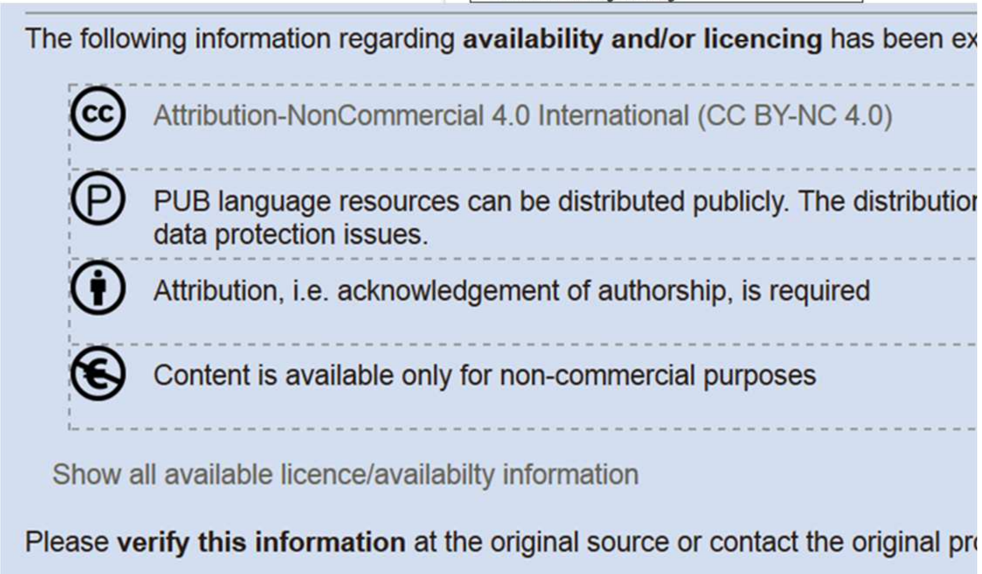
universität

Continue

Allow me to pick from a list

Help

Albert-Ludwigs-Universität Freiburg
Alpen-Adria-Universität Klagenfurt
Humboldt-Universität zu Berlin
Technische Universität Clausthal
Universität Erlangen-Nürnberg
Universität Frankfurt
Universität Mainz
Universität Münster
Universität of Munich (LMU)
Universität Osnabrück
Universität Regensburg



The following information regarding **availability and/or licencing** has been ex

Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

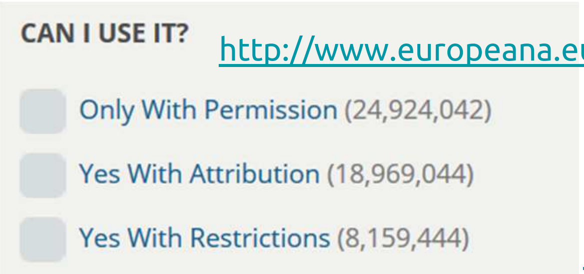
PUB language resources can be distributed publicly. The distributio data protection issues.

Attribution, i.e. acknowledgement of authorship, is required

Content is available only for non-commercial purposes

Show all available licence/availabilty information

Please **verify this information** at the original source or contact the original pr



CAN I USE IT?

http://www.europeana.eu/portal/search

□ Only With Permission (24,924,042)

□ Yes With Attribution (18,969,044)

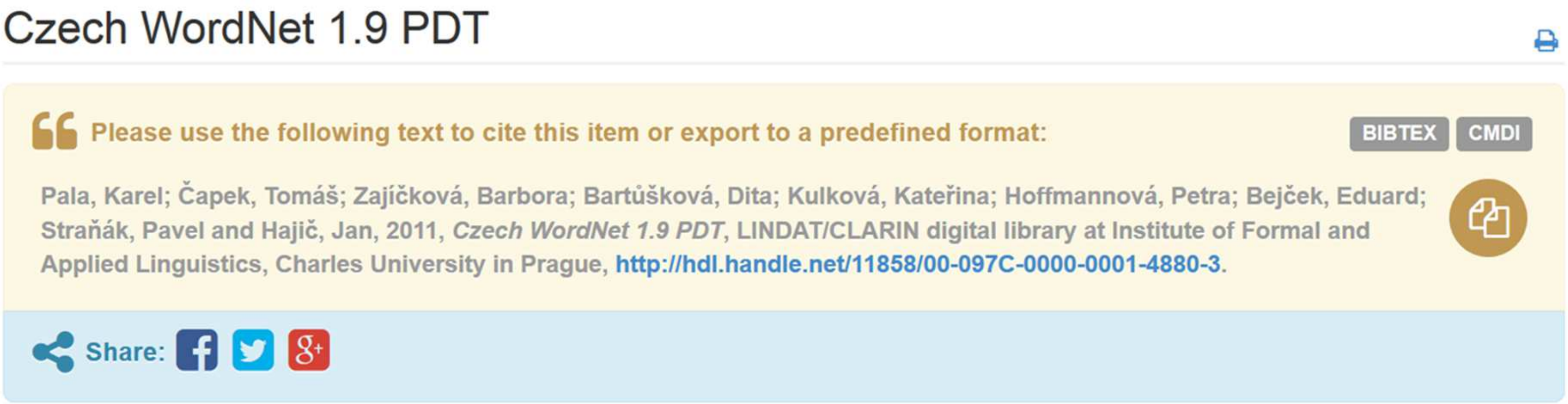
□ Yes With Restrictions (8,159,444)

http://beta-vlo.clarin.eu

.cite

- Persistent Identifiers (PID)
  - Handle.net, DOI, ARK
    http://hdl.handle.net/11858/00-1734-0000-0009-FEA1-D
- Activities
  - DataCite – DOI(PID) for datasets
  - Thor - integration between articles, data, and researchers across the research lifecycle
  - RDA Working Group on Data Citation Dynamic Data Citation

Cite-helper LINDAT:



Czech WordNet 1.9 PDT

Please use the following text to cite this item or export to a predefined format:

Pala, Karel; Čapek, Tomáš; Zajíčková, Barbora; Bartůšková, Dita; Kulková, Kateřina; Hoffmannová, Petra; Bejček, Eduard; Straňák, Pavel and Hajič, Jan, 2011, *Czech WordNet 1.9 PDT*, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0001-4880-3.

repositories

- Software:
  Fedora, DSpace, CKAN, …

- Services:
  (institutional / domain-specific / infrastructural)
  - GAMS
  - PHAIDRA
  - epub.oeaw
  - ads – archaeology data service
  - CLARIN Depositing Services
  - Datahub by Open Knowledge Foundation – “give your data a home” (10.695 datasets)
  - Figshare – “credit for **all** your research”
  - Zenodo

  - Re3data –Registry of research data repositories
  - COAR – Confederation of Open Access Repositories

Use Cases - Language Resources

- amc - austrian media corpus - corpora
  http://hdl.handle.net/11022/0000-0000-478D-2 [TextCorpusProfile]

- DictGate - lexicographic resources
  http://hdl.handle.net/11022/0000-0000-001E-F [LexicalResourceProfile]

- DTA - Deutsches Textarchiv
  (TEI -> CMDI)
  2630 records in VLO

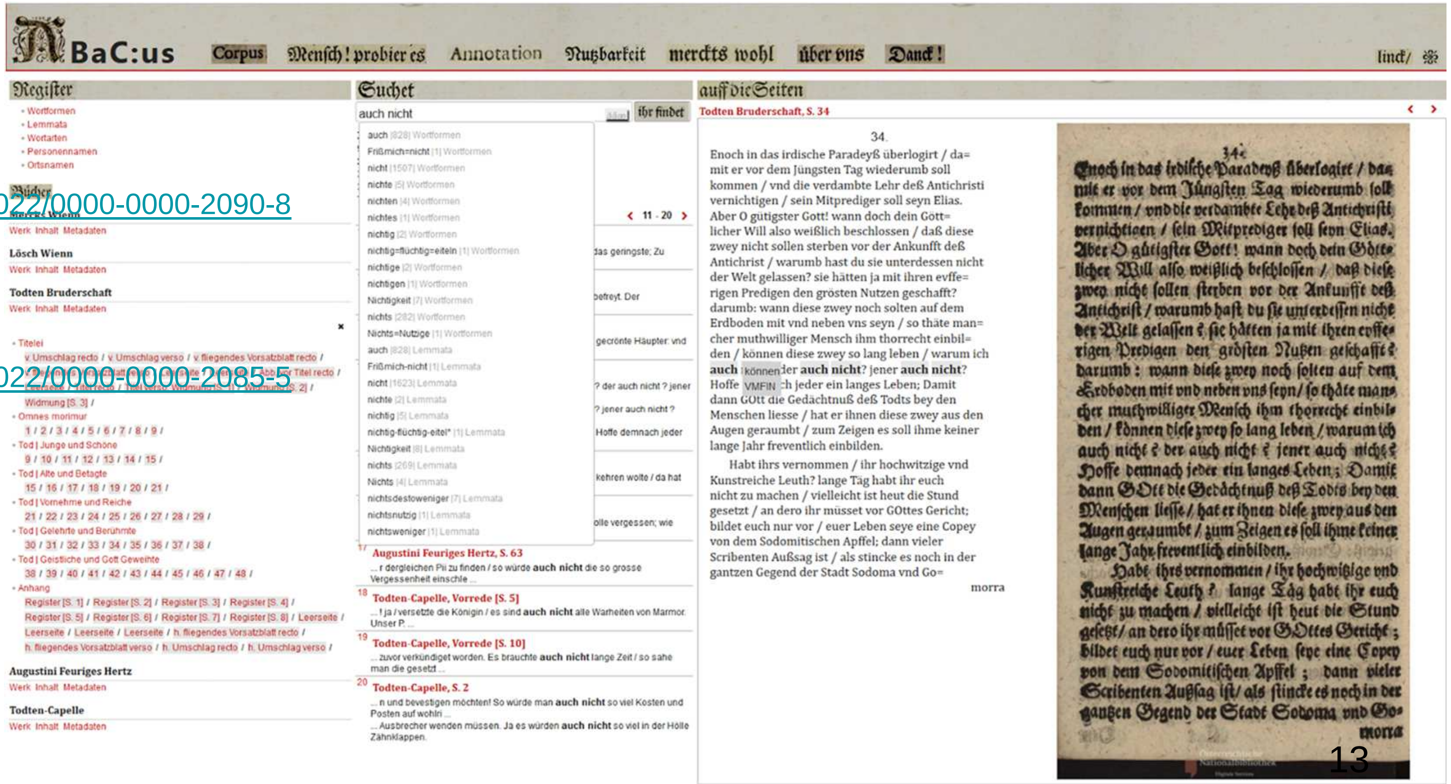
Use Cases - Digital Editions - ABaC:us

- collection of historic texts from 17th century

CMDI record for

- collection
  http://hdl.handle.net/11022/0000-0000-2090-8
- texts
  [teiHeader]
  http://hdl.handle.net/11022/0000-0000-2085-5

...



Regiſter

- Wortformen
- Lemmata
- Wortarten
- Personennamen
- Ortsnamen

Bücher

Mercks Wienn
Werk Inhalt Metadaten

Lösch Wienn
Werk Inhalt Metadaten

Todten Bruderschaft
Werk Inhalt Metadaten

×

- Titelei
  v. Umschlag recto / v. Umschlag verso / v. fliegendes Vorsatzblatt recto / v. fliegendes Vorsatzblatt verso / Leerseite / Leerseite / Abb. vor Titel recto / Leerseite / Titel recto / Titel verso / Widmung [S. 1] / Widmung [S. 2] / Widmung [S. 3] /
- Omnes morimur
  1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 /
- Tod | Junge und Schöne
  9 / 10 / 11 / 12 / 13 / 14 / 15 /
- Tod | Alte und Betagte
  15 / 16 / 17 / 18 / 19 / 20 / 21 /
- Tod | Vornehme und Reiche
  21 / 22 / 23 / 24 / 25 / 26 / 27 / 28 / 29 /
- Tod | Gelehrte und Berühmte
  30 / 31 / 32 / 33 / 34 / 35 / 36 / 37 / 38 /
- Tod | Geistliche und Gott Geweihte
  38 / 39 / 40 / 41 / 42 / 43 / 44 / 45 / 46 / 47 / 48 /
- Anhang
  Register [S. 1] / Register [S. 2] / Register [S. 3] / Register [S. 4] / Register [S. 5] / Register [S. 6] / Register [S. 7] / Register [S. 8] / Leerseite / Leerseite / Leerseite / Leerseite / h. fliegendes Vorsatzblatt recto / h. fliegendes Vorsatzblatt verso / h. Umschlag recto / h. Umschlag verso /

Augustini Feuriges Hertz
Werk Inhalt Metadaten

Todten-Capelle
Werk Inhalt Metadaten

Suchet

auch nicht | ihr findet

auch |828| Wortformen
Frißmich=nicht |1| Wortformen
nicht |1507| Wortformen
nichte |5| Wortformen
nichten |4| Wortformen
nichtes |1| Wortformen
nichtig |2| Wortformen
nichtig=flüchtig=eiteln |1| Wortformen
nichtige |2| Wortformen
nichtigen |1| Wortformen
Nichtigkeit |7| Wortformen
nichts |282| Wortformen
Nichts=Nutzige |1| Wortformen
auch |828| Lemmata
Frißmich-nicht |1| Lemmata
nicht |1623| Lemmata
nichte |2| Lemmata
nichtig |5| Lemmata
nichtig-flüchtig-eitel* |1| Lemmata
Nichtigkeit |8| Lemmata
nichts |269| Lemmata
Nichts |4| Lemmata
nichtsdestoweniger |7| Lemmata
nichtsnutzig |1| Lemmata
nichtsweniger |1| Lemmata

‹ 11 - 20 ›

das geringste; Zu

befreyt. Der

gecrönte Häupter: vnd

? der auch nicht ? jener

? jener auch nicht ?

Hoffe demnach jeder

kehren wolte / da hat

olle vergessen; wie

17 **Augustini Feuriges Hertz, S. 63**
... r dergleichen Pii zu finden / so würde **auch nicht** die so grosse Vergessenheit einschle ...

18 **Todten-Capelle, Vorrede [S. 5]**
... ! ja / versetzte die Königin / es sind **auch nicht** alle Warheiten von Marmor. Unser P. ...

19 **Todten-Capelle, Vorrede [S. 10]**
... zuvor verkündiget worden. Es brauchte **auch nicht** lange Zeit / so sahe man die gesetzt ...

20 **Todten-Capelle, S. 2**
... n und bevestigen möchten! So würde man **auch nicht** so viel Kosten und Posten auf wohlri ...
... Ausbrecher wenden müssen. Ja es würden **auch nicht** so viel in der Hölle Zähnklappen.

auff die Seiten

Todten Bruderschaft, S. 34 ‹ ›

34.

Enoch in das irdische Paradeyß überlogirt / da=
mit er vor dem Jüngsten Tag wiederumb soll
kommen / vnd die verdambte Lehr deß Antichristi
vernichtigen / sein Mitprediger soll seyn Elias.
Aber O gütigster Gott! wann doch dein Gött=
licher Will also weißlich beschlossen / daß diese
zwey nicht sollen sterben vor der Ankunfft deß
Antichrist / warumb hast du sie unterdessen nicht
der Welt gelassen? sie hätten ja mit ihren evffe=
rigen Predigen den grösten Nutzen geschafft?
darumb: wann diese zwey noch solten auf dem
Erdboden mit vnd neben vns seyn / so thäte man=
cher muthwilliger Mensch ihm thorrecht einbil=
den / können diese zwey so lang leben / warum ich
auch können der **auch nicht**? jener **auch nicht**?
Hoffe VMFIN ch jeder ein langes Leben; Damit
dann GOtt die Gedächtnuß deß Todts bey den
Menschen liesse / hat er ihnen diese zwey aus den
Augen geraumbt / zum Zeigen es soll ihme keiner
lange Jahr freventlich einbilden.

Habt ihrs vernommen / ihr hochwitzige vnd
Kunstreiche Leuth? lange Täg habt ihr euch
nicht zu machen / vielleicht ist heut die Stund
gesetzt / an dero ihr müsset vor GOttes Gericht;
bildet euch nur vor / euer Leben seye eine Copey
von dem Sodomitischen Apffel; dann vieler
Scribenten Außsag ist / als stincke es noch in der
gantzen Gegend der Stadt Sodoma vnd Go=
morra

Zitation: Abraham à Sancta Clara: Grosse Todten Bruderschaft. Wien, 1681. (Digitale Ausgabe) S. 34. In: ABaC:us – Austrian Baroque Corpus. Hrsg. von Claudia Resch und Ulrike Czeitschner. <https://acdh.oeaw.ac.at/abacus/get/abacus.7_46> abgerufen am 29. 4. 2016

Use Cases - Structured data ?

lots of data in **relational databases**

- Adlib - archival records
- DEFC - archaeological finds and sites
- APIS - biographical data of the ÖBL

Challenges:

- What is data / what is metadata?
- Granularity (one entry for whole application, or for individual items?)
- What target format to map to?

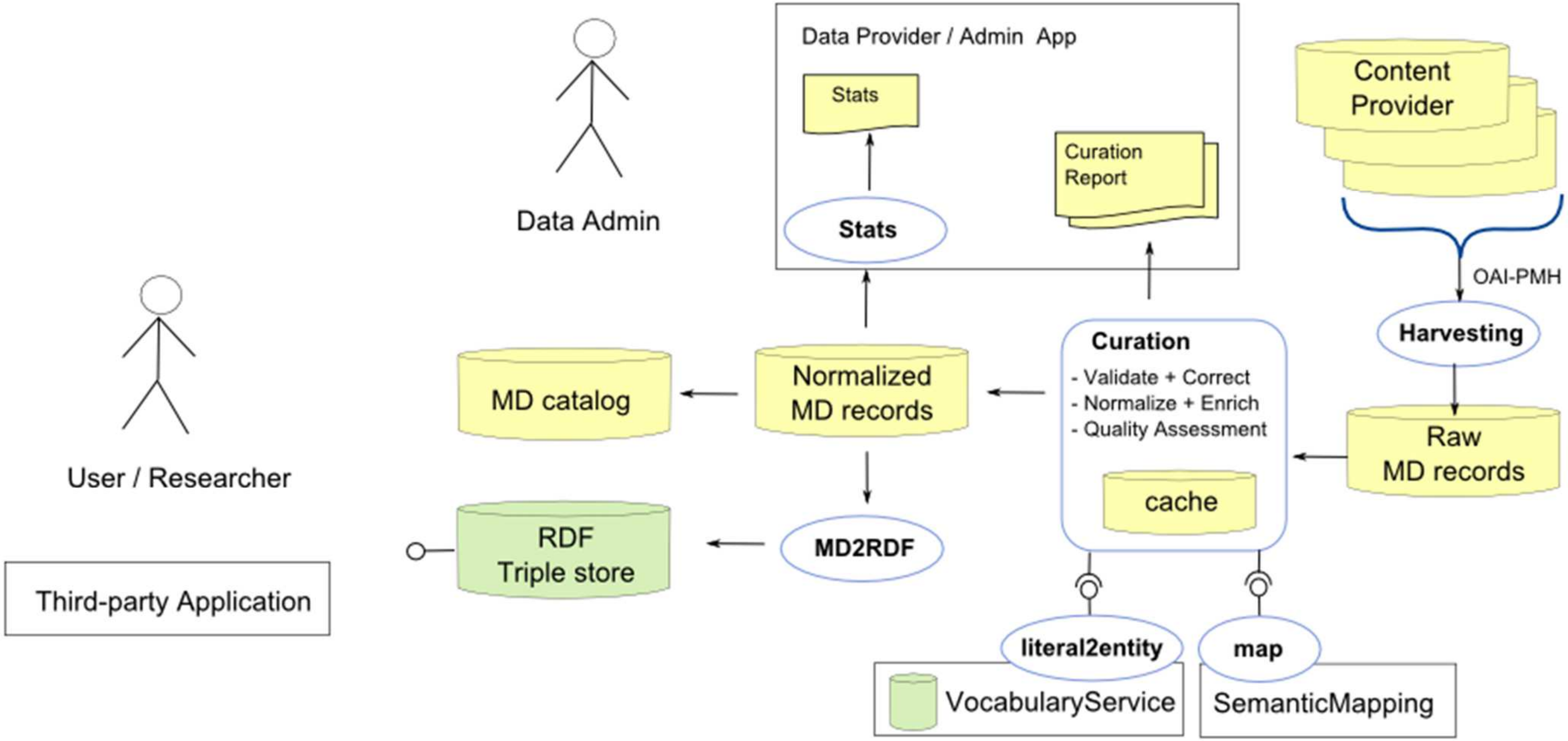
=> RDF ?

Research Infrastructures / Aggregators

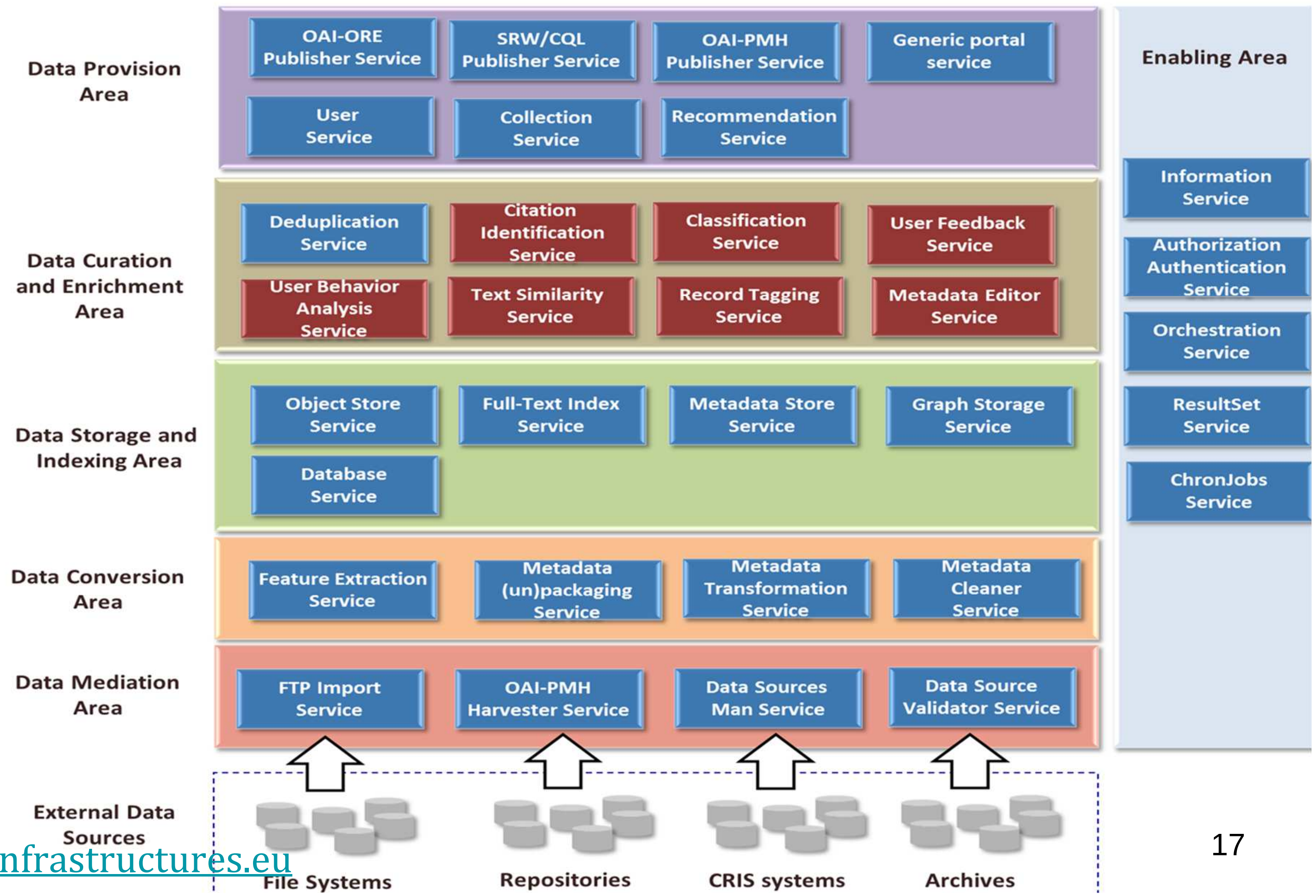
- CLARIN
- DARIAH
- PARTHENOS
- OpenAIRE

  - starting point: registry of repositories?
  - crosswalks between formats - usually least common denominator
    => reduction during aggregation
  - try union with lacunae ?
  - establishing entity-identity
    esp. resurrect entities from string values! (organisations, persons)
  - granularity? (just collections, or individual items?)

metadata workflow - harvesting, curation, publishing

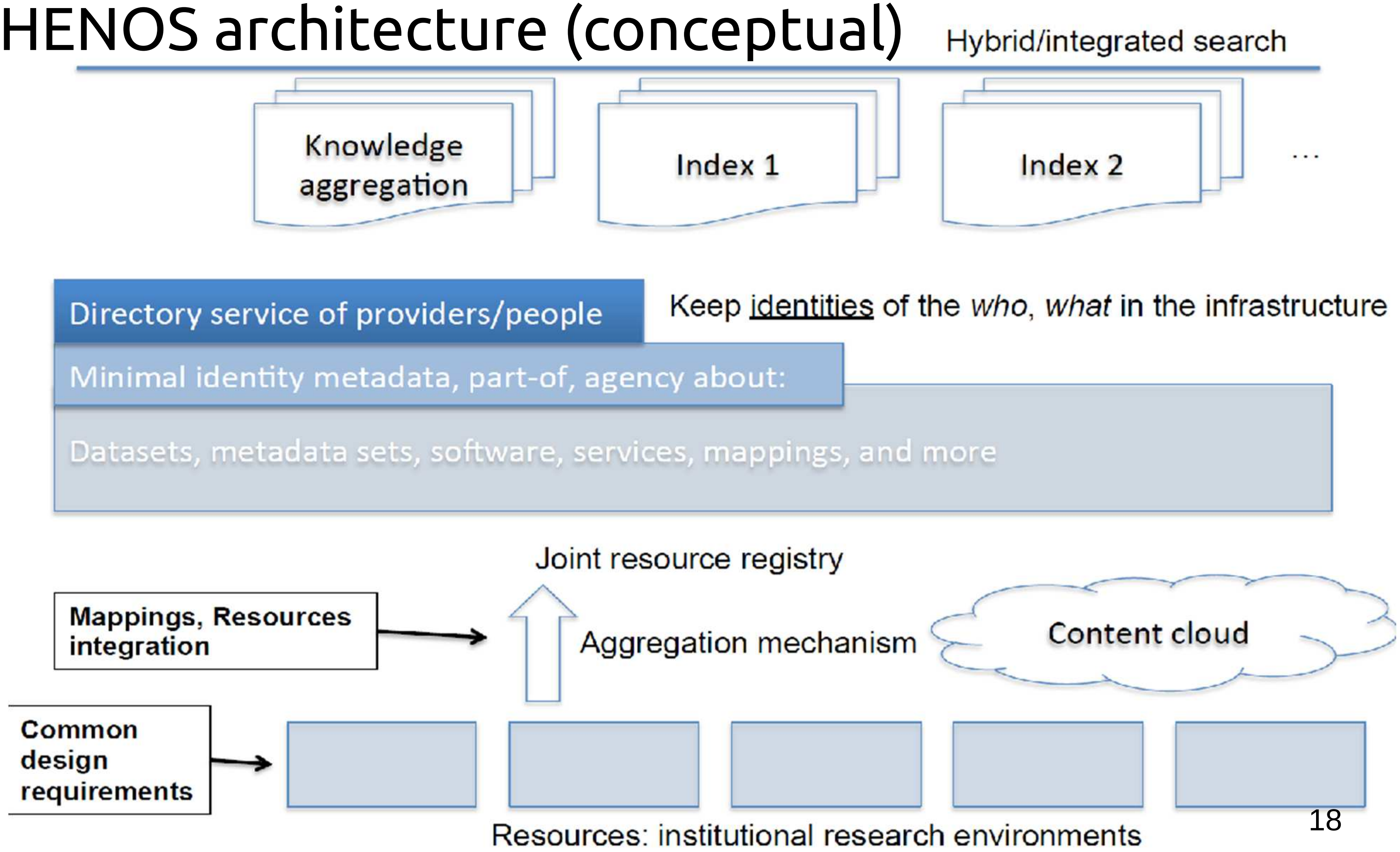


D-Net



d-net.research-infrastructures.eu

PARTHENOS architecture (conceptual)

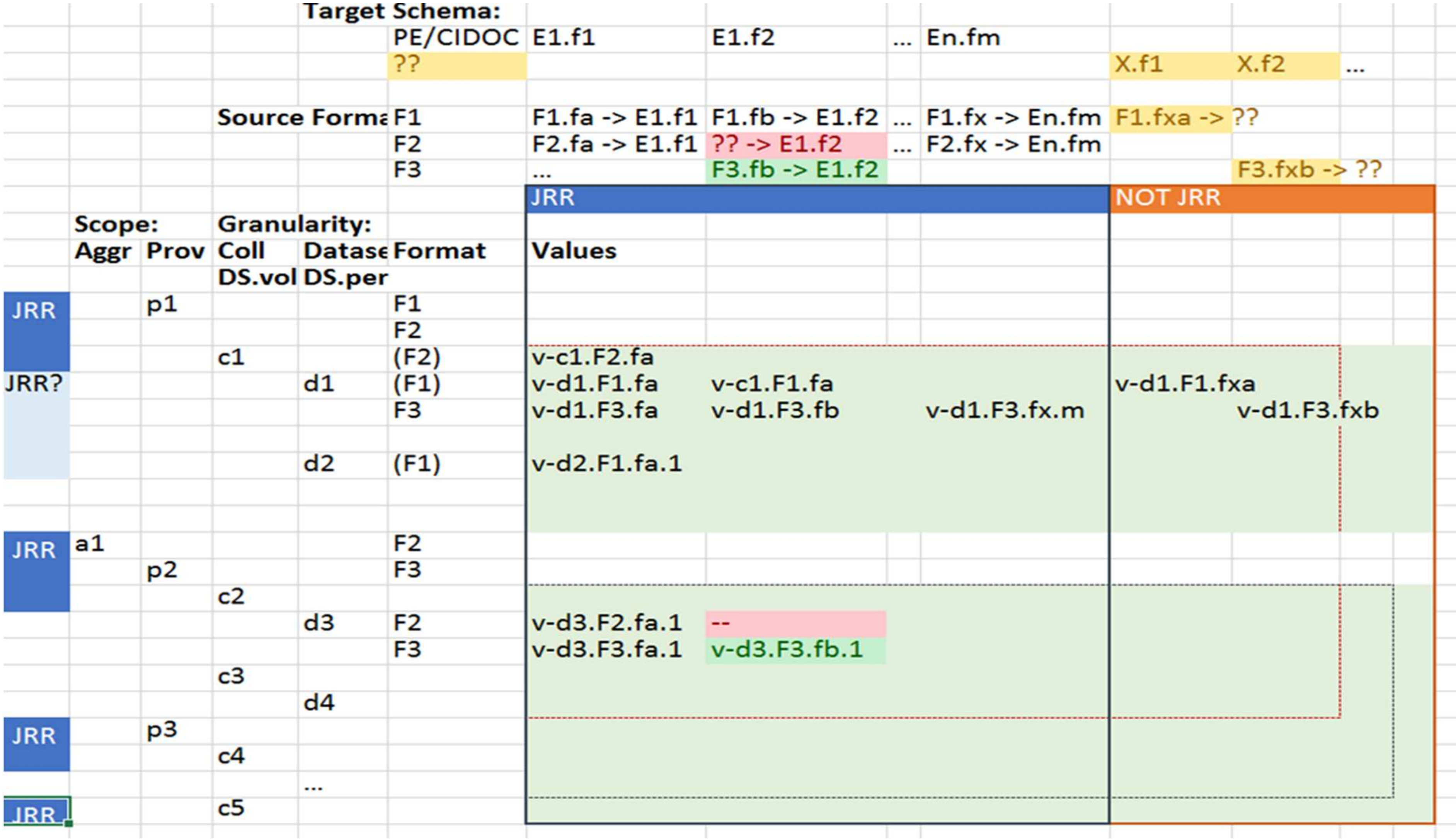

Hybrid/integrated search
Knowledge aggregation
Index 1
Index 2
...
Directory service of providers/people
Keep identities of the *who*, *what* in the infrastructure
Minimal identity metadata, part-of, agency about:
Datasets, metadata sets, software, services, mappings, and more
Joint resource registry
Mappings, Resources integration
Aggregation mechanism
Content cloud
Common design requirements
Resources: institutional research environments

How big is the data space?

- How big is the data space PARTHENOS (or any aggregator) aims to offer?
- In terms of: Scope vs. Granularity vs. Format
- **Scope**: Aggregator/Provider/"Collection"

  Repositories, Aggregators, reference resources?
- **Granularity**: "Collection"/.../"File"
- **Format:**
  - Each aggregator/provider can offer MD in different formats
  - Some entities may be just (string) values in source format (esp. Actor, Service)

    -> need to **generate entities** from values! (aggregate, normalise, disambiguate)
  - Provider tend to have more detailed formats than aggregators



				Target Schema:								
					PE/CIDOC	E1.f1	E1.f2	...	En.fm			
					??					X.f1	X.f2	...
			Source Forma		F1	F1.fa -> E1.f1	F1.fb -> E1.f2	...	F1.fx -> En.fm	F1.fxa ->	??	
					F2	F2.fa -> E1.f1	?? -> E1.f2	...	F2.fx -> En.fm			
					F3	...	F3.fb -> E1.f2				F3.fxb -> ??	
						JRR				NOT JRR		
	Scope:		Granularity:									
	Aggr	Prov	Coll	Datase	Format	Values						
			DS.vol	DS.per								
JRR		p1			F1							
					F2							
			c1		(F2)	v-c1.F2.fa						
JRR?				d1	(F1)	v-d1.F1.fa	v-c1.F1.fa			v-d1.F1.fxa		
					F3	v-d1.F3.fa	v-d1.F3.fb		v-d1.F3.fx.m		v-d1.F3.fxb	
				d2	(F1)	v-d2.F1.fa.1						
JRR	a1				F2							
		p2			F3							
			c2									
				d3	F2	v-d3.F2.fa.1	--					
					F3	v-d3.F3.fa.1	v-d3.F3.fb.1					
			c3									
				d4								
JRR		p3										
			c4									
				...								
JRR			c5									

ÖAW AUSTRIAN ACADEMY OF SCIENCES

ACDH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES

Thank you!

Questions?

Matej Ďurčo @ ACDH Tool Gallery 2.1, 2016-03-16