



CLARIN Federated Content Search workshop

DARIAH-DE Generic Search

DARIAH-DE Generic Search

- 1) CONTEXT AND IDEAS**
- 2) MODEL AND ARCHITECTURE
- 3) CURRENT PROTOTYPES

Data sources for DARIAH-DE

- Attributes of relevant data sources
 - relevant for a **research question / individual user**
 - structured and semi-structured data with a **common data model** and structural constraints
 - **no quality limitations** (content and representation)
- Valid data origin
 - **Harvestable/crawlable** sources (OAI-PMH, web, etc.)
 - **User upload** (XML, CSV, etc.)

*Valid for individual usage,
not the „global picture“*

Not „globally“ visible

*No ensured access
method / protocol*

Abstract use cases

■ Some use cases

- **comprehensive** analysis, visualization and search
- **deep** analysis and interaction

*Many collections,
low content-depth*

*highly correlated collections
e. g. in research projects*

↑ *Usage (querying, visualisation etc.)*

↓ *Design and definition*

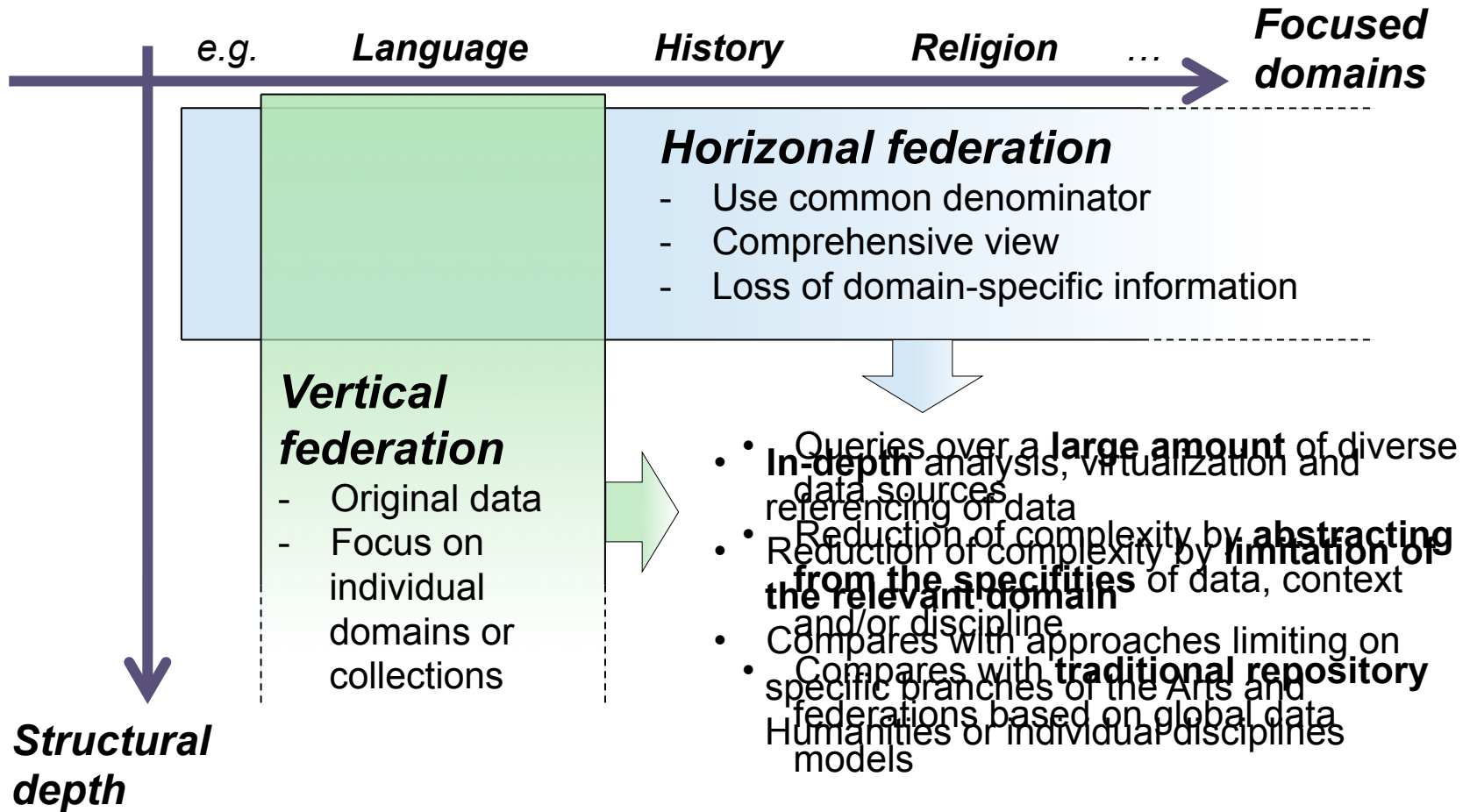
- **definition, classification and semantic relation** of data sources by respective domain experts
- **collaboration and discussion** about semantics of data and correlations between sources

*e. g. in order to identify
valid but conflicting
disciplinary interpretations*

Main ideas for federation

- **Encapsulate technical details** of data interoperability
- Supply interfaces for both **generic and research-specific applications**
 - Minimize information loss for **deep data analysis** (by allowing immediate integration of collections)
 - Provide **federated views** on data adaptable to the **specific needs and perspective** of individual disciplines Wide or deep search
- Allow the coexistence of **valid but opposing discipline-specific perspectives** on data and correlations
 - Let the researcher decide on the interpretation of semantics
 - Allow local integration projects that do not interfere with the global picture on data federation in DARIAH

Opposing integrated views

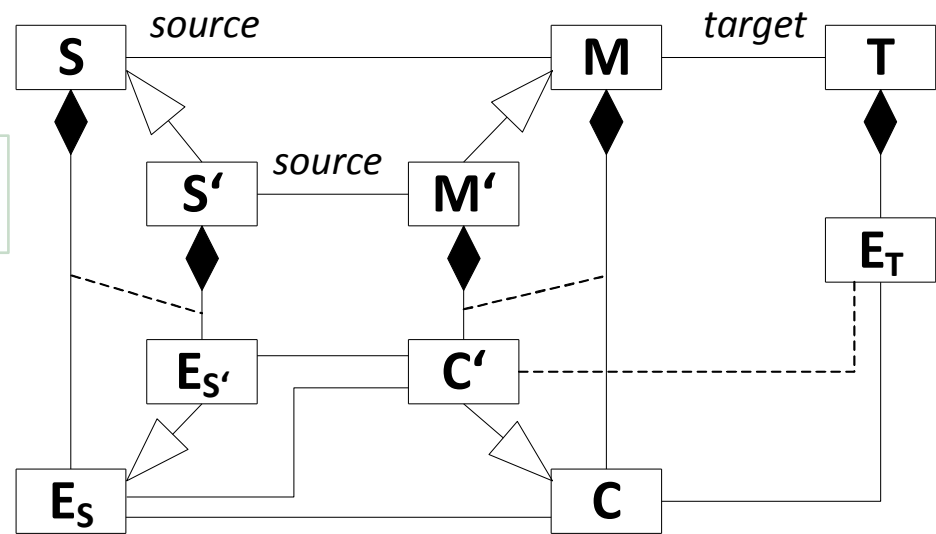


DARIAH-DE Generic Search

- 1) CONTEXT AND IDEAS
- 2) MODEL AND ARCHITECTURE**
- 3) CURRENT PROTOTYPES

Integration model

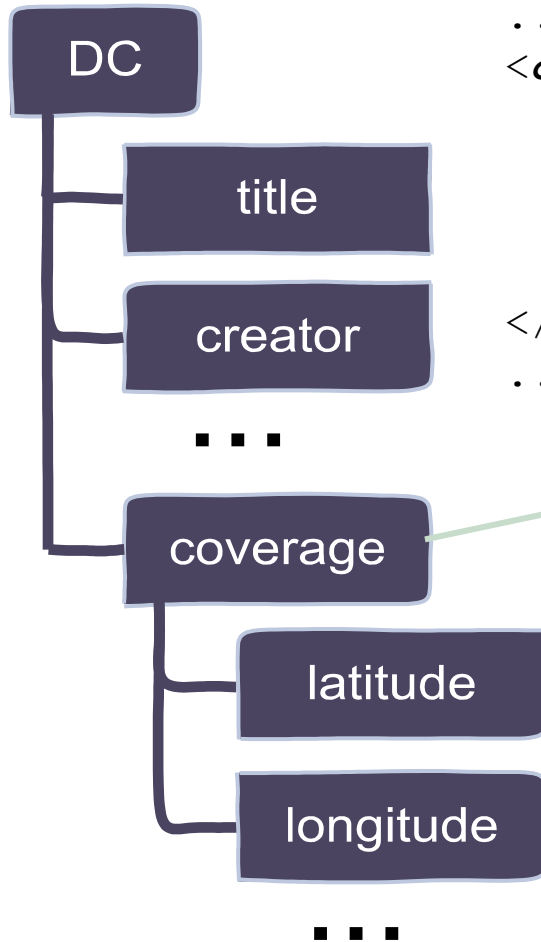
- Mappings as **direct associations** between collections or models *Deep (local) semantics
→ low information loss*
- Main Problems: Does not scale well, no global picture
- **Inheritance model** to - allow definition of **global structures and mappings** as well as their derivation as **discipline- or archive specific versions**



Results: Not every action needs to be done on a global level; researchers can „play“ with data; sub-schemas are (through hierarchy) usable for global/generic mappings

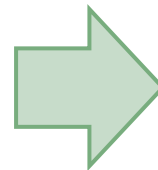
S: Source
T: Target
M: Mapping
E: Element
C: Correlance

Example usage of derivation



```
...  
<dc:coverage> http://doi.pangaea.de/10.1594/PANGAEA.51915  
  LATITUDE: -46.069333 * LONGITUDE: 90.111167  
  * MINIMUM AGE: 4.610 ka BP * MAXIMUM AGE:  
  201.000 ka BP * MINIMUM DEPTH, sediment: 0.0  
  m * MAXIMUM DEPTH, sediment: 11.7 m  
</dc:coverage>  
...
```

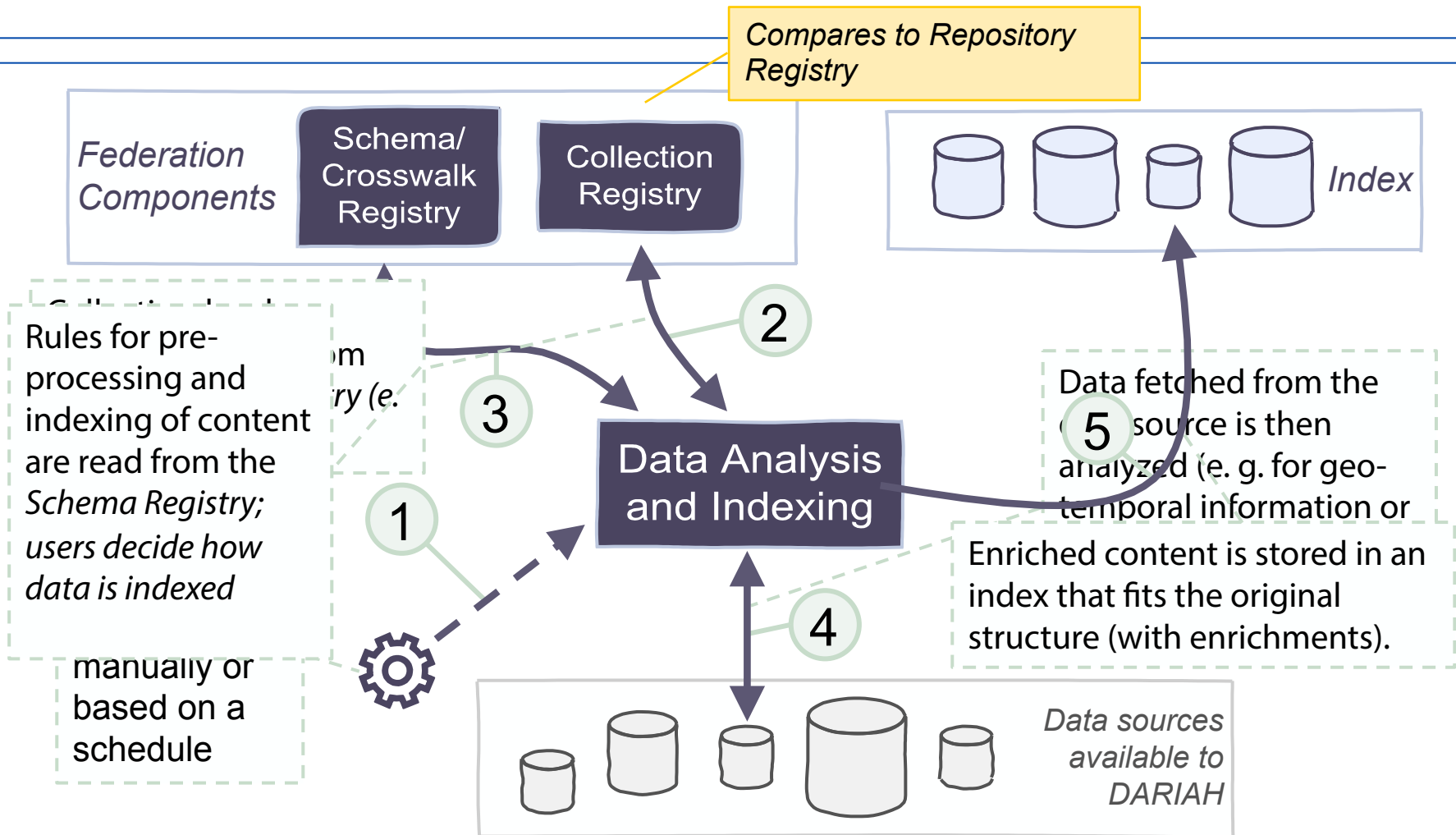
Explicate semantics to enrich local data models at indexing-time; Example here:
`(?<Key>\w[\s\w-,]*) : (?<Value>[^\s]*)/*`



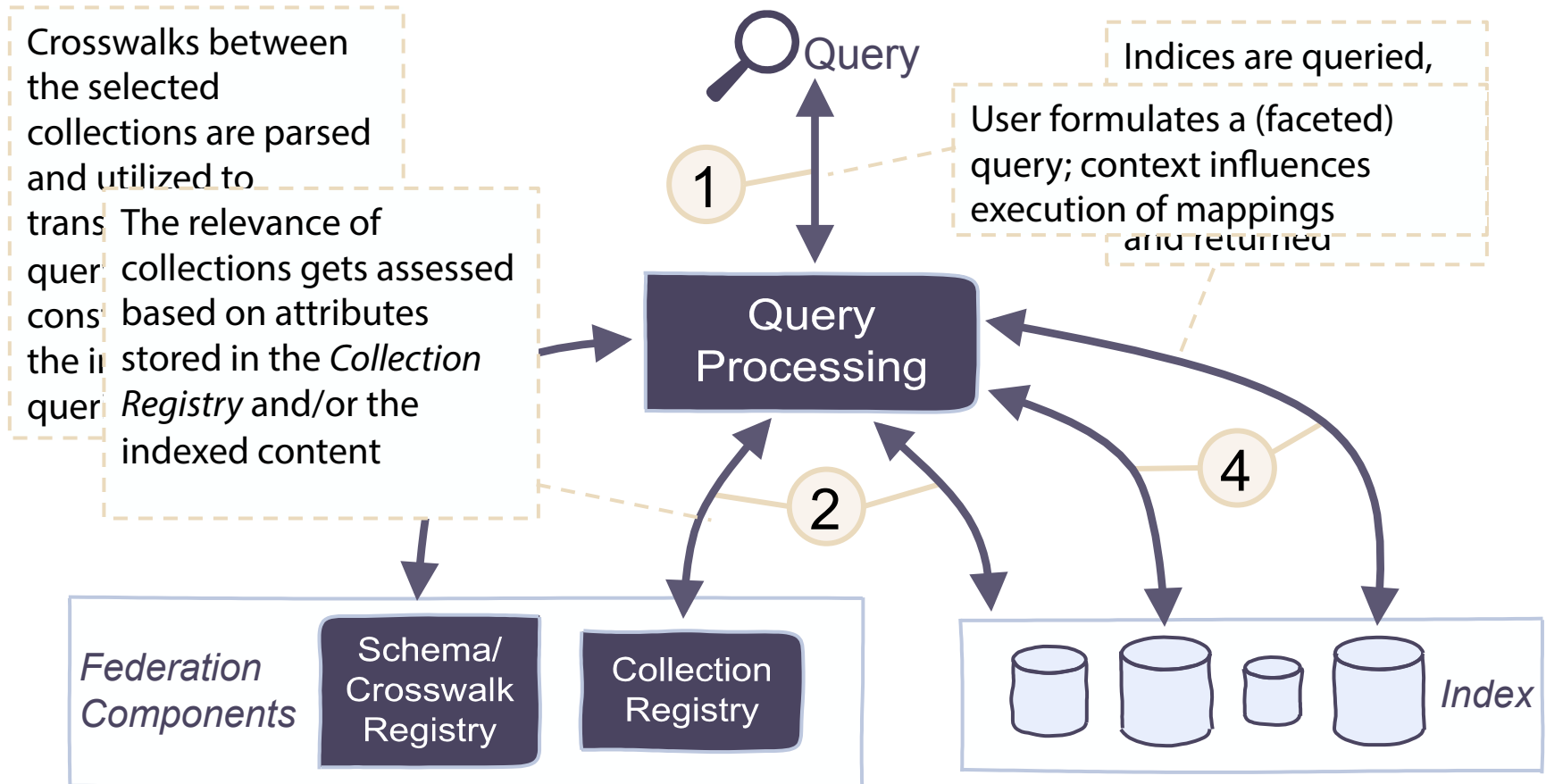
Enriched data available for analysis and/or mappings to other data sources for the **derived local schema**

Other examples: array of names in creator, 18xx as uncertain representation of years

Analysis and indexing



Query Processing



DARIAH-DE Generic Search

- 1) CONTEXT AND IDEAS
- 2) MODEL AND ARCHITECTURE
- 3) CURRENT PROTOTYPES**

Schema and Crosswalk Registry

DARIAH-DE Schema Registry Crosswalk Registry Administration [Login](#) Language: (en) ▾

Schema Registry

[+ Register Schema](#)

	Name	Schema Type	Last modified
✓	DC2	XML Schema	11/28/12 1
✓	infra	XML Schema	11/28/12 2
✓	DC1	XML Schema	11/28/12 1
✓	CDWA	XML Schema	11/28/12 2
✓	VRA	XML Schema	11/28/12 2

Crosswalk Mapping

[+ Save crosswalk](#)

[Expand all](#) [Collapse all](#) [Reset View](#) [Expand all](#) [Collapse all](#) [Reset View](#)

The diagram illustrates a crosswalk mapping between two schema structures. On the left, a tree structure includes elements like `locIDtype`, `termsource`, `locID`, `labelRelatedWork`, `indexingMaterialsTechWrap`, `indexingMaterialsTechSet`, `termMaterialsTech`, `termsource`, `sourceMaterialsTech`, and `extentMaterialsTech`. On the right, a tree structure includes `vra`, `collection`, `work`, `image`, `techniqueSet`, `relationSet`, `locationSet`, `descriptionSet`, and `culturalContextSet`. Numerous yellow lines connect corresponding elements between the two trees, representing the crosswalk mapping. A dashed line highlights a specific connection between `indexingMaterialsTechWrap` on the left and `work` on the right.

Search: Facetted / simple

The image shows two overlapping screenshots of the DARIAH-DE search interface. The top screenshot displays a facetted search with the following details:

- Search facets: ANY, Bamberg, dc:subject, NOT VD18
- Query returned 348 results in 762ms
- Results list:
 - Bamberg. deutsche Stadt der Wunder und Träume**
Bayerische Staatsbibliothek (BSB): Münchener Digitalisierungszentrum
<http://nbn-resolving.de/urn:nbn:de:bvb:12-bsb00011868-0>
Content
 - Illustrierter Führer durch Bamberg und Umgebung. m**
Bayerische Staatsbibliothek (BSB): Münchener Digitalisierungszentrum
<http://nbn-resolving.de/urn:nbn:de:bvb:12-bsb00012018-8>
Content
 - Ueber Käfermilben und Bamberg.**
GDZ - Göttinger Digitalisierungszentrum
http://resolver.sub.uni-goettingen.de/purl?PPN623918188_0012/DMDL
Content
 - Statistische Entscheidungstheorie - BAMBERG, G.**
GDZ - Göttinger Digitalisierungszentrum
<http://resolver.sub.uni-goettingen.de/purl?GDZPPN002453924>
Content

The bottom screenshot shows a simple search interface with the following details:

- Search facets: Bamberg NOT dc:subject: VD18
- Query returned 348 results in 317ms
- Results list:
 - Bamberg. deutsche Stadt der Wunder und Träume**
Bayerische Staatsbibliothek (BSB): Münchener Digitalisierungszentrum (MDZ)
<http://nbn-resolving.de/urn:nbn:de:bvb:12-bsb00011868-0>
Content (Score: 1.4719139)
 - Illustrierter Führer durch Bamberg und Umgebung. mit Ausflügen in d. Steigerw...**
Bayerische Staatsbibliothek (BSB): Münchener Digitalisierungszentrum (MDZ)
<http://nbn-resolving.de/urn:nbn:de:bvb:12-bsb00012018-8>
Content (Score: 1.4719139)
 - Ueber Käfermilben und Bamberg.**
GDZ - Göttinger Digitalisierungszentrum
http://resolver.sub.uni-goettingen.de/purl?PPN623918188_0012/DMDLOG_0005
Content (Score: 1.4719139)
 - Statistische Entscheidungstheorie - BAMBERG, G.**
GDZ - Göttinger Digitalisierungszentrum
<http://resolver.sub.uni-goettingen.de/purl?GDZPPN002453924>
Content (Score: 1.4719139)

A yellow arrow points from the facetted search results to the simple search interface, indicating the transition between the two search modes.

Search: Result presentation

Indexed data is close to original data + enriched metadata

Left Screenshot: Search Results for 'digital humanities'

Simple search: digital humanities
Query returned 39600 results in 452ms

Robert Koch und die Digital Humanities
Berlin-Brandenburgische Akademie der Wissenschaften, edoc
<http://edoc.bbaw.de/volltexte/2012/2251/>

Score: 1.1383578

Content

- oai_dc:dc:
 - dc:creator: Schnöpf, Markus
 - dc:date: 2012
 - dc:description:
 - dc:format: application/pdf
 - dc:identifier: ["urn:nbn:de:kobv:b4-opus-22511", "http://edoc.bbaw.de/volltexte/2012/2251/"]
 - dc:language: ger
 - dc:publisher: ["Berlin-Brandenburgische Akademie der Wissenschaften", "Arbeitsgruppe Gegenworte - Hefte für den Disput über Wissen"]
 - dc:rights: http://edoc.bbaw.de/doku/urheberrecht.php
 - dc:source: Gegenworte : Hefte für den Disput über Wissen
 - dc:subject: ["Angewandte Forschung ; Grundlagenforschung"]
 - dc:title: Robert Koch und die Digital Humanities
 - dc:type: Article
 - xmlns:dc: http://purl.org/dc/elements/1.1/
 - xmlns:oai_dc: http://www.openarchives.org/OAI/2.0/oai_dc.xsd
 - xmlns:xsi: http://www.w3.org/2001/XMLSchema-instance
 - xsi:schemaLocation: http://www.openarchives.org/OAI/2.0/oai_dc.xsd

Right Screenshot: Search Results for 'bamborg'

Simple search: bamborg
Query returned 432 results in 147ms

- PANGAEA - Data Publisher for Earth & Environmental Science
Hits: 248
- Bayerische Staatsbibliothek (BSB): Münchener Digitalisierungszentrum (MDZ)
Hits: 84
- Zentrales Verzeichnis Digitaler Drucke (ZVDD)
Hits: 57
- GDZ - Göttinger Digitalisierungszentrum
Hits: 38
- Hochschulschriftenserver der Friedrich-Alexander Universität Erlangen-Nürnberg
Hits: 2
- Dokumentenserver der Akademie der Wissenschaften zu Göttingen
Hits: 1

Links

Thank you!

DARIAH-EU

→ <http://dariah.eu>

DARIAH-DE

→ <http://de.dariah.eu>

Generic Search

→ <http://demo2.dariah.eu:8080/search>

Registries

→ <http://demo2.dariah.eu:8080/schereg>