DARIAH-DE
Digital Research Infrastructure for the Arts and Humanities

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

CLARIN Federated Content Search workshop

DARIAH-DE
Generic Search

Tobias Gradl
University of Bamberg

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 for individual usage, not the „global picture“

- Valid data origin
  - **Harvestable/crawlable** sources (OAI-PMH, web, etc.)
  - User **upload** (XML, CSV, etc.)

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

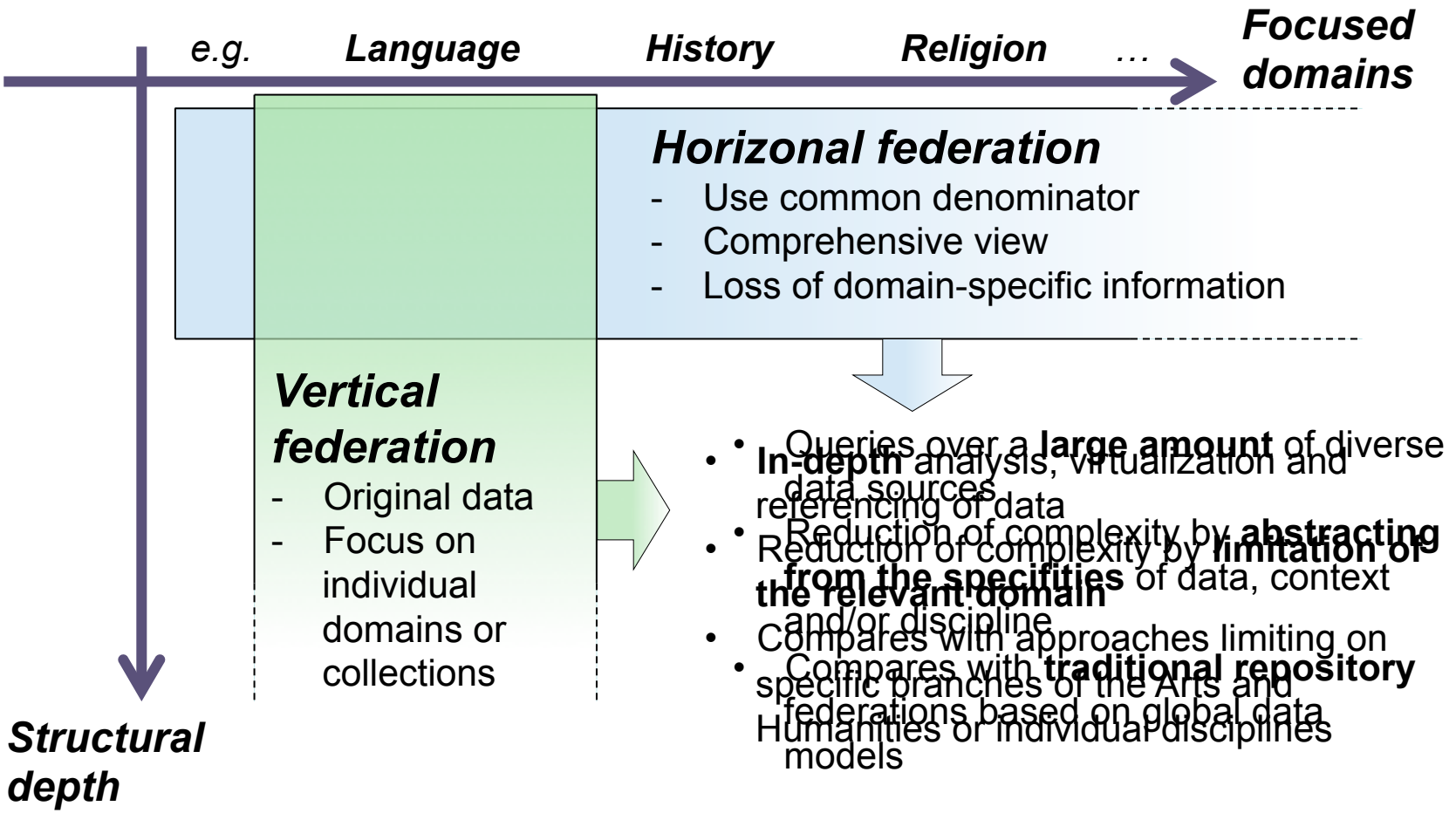
Main ideas for federation

- **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

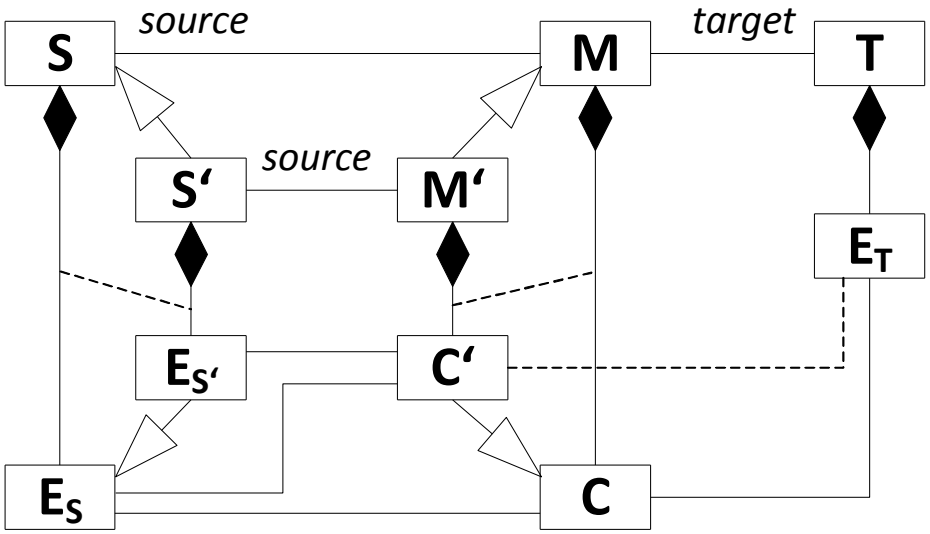
3) CURRENT PROTOTYPES

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

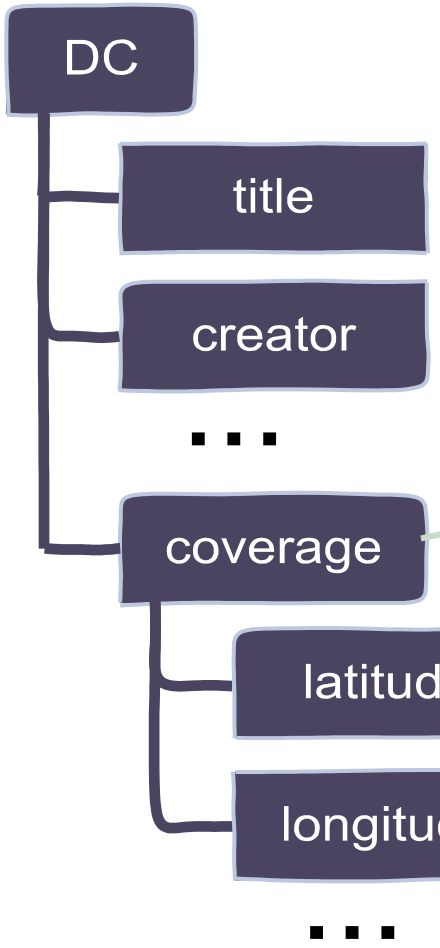


- **Inheritance model** to - allow definition of **global structures and mappings** as well as their derivation as **discipline- or archive specific versions**

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>
    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>
...
```

http://doi.pangaea.de/10.1594/PANGAEA.51915

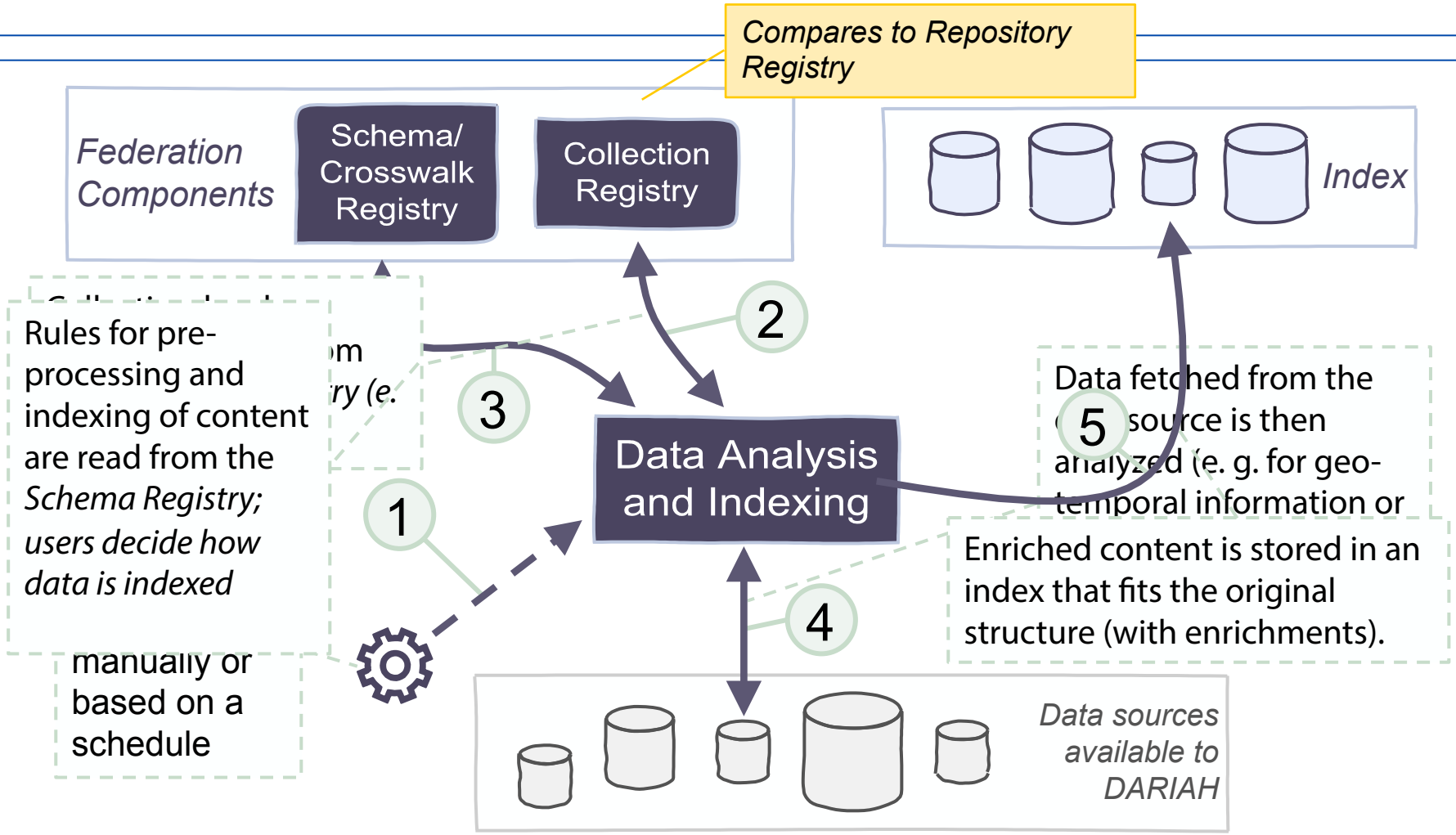
Explicate semantics to enrich local data models at indexing-time; Example here:

```
(?<Key>\w[\s\w-,]*):(?<Value>[^\*]*)/*
```

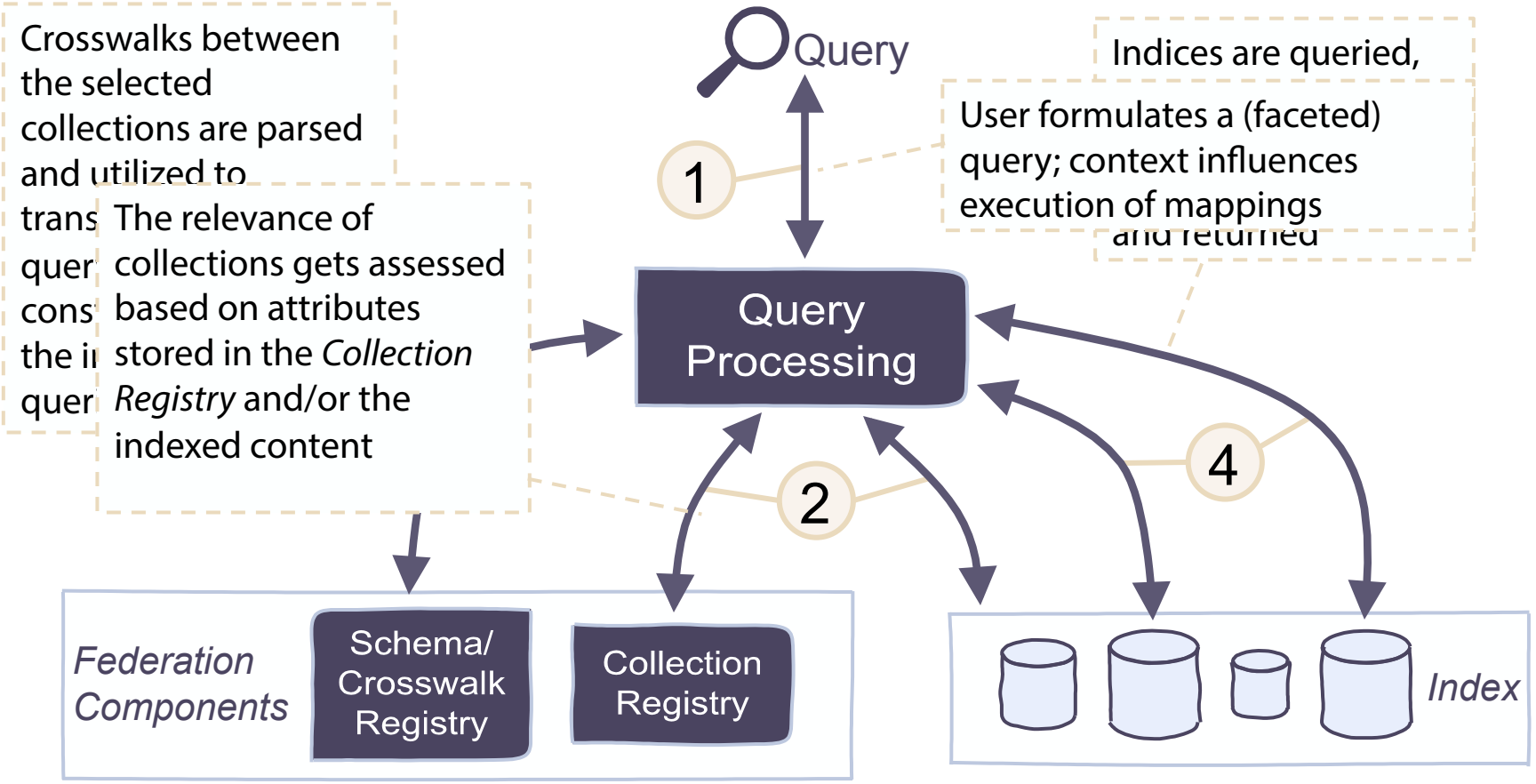
Enriched data available for analysis and/or mappings to other data sources for the **derived local schema**

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

Analysis and indexing



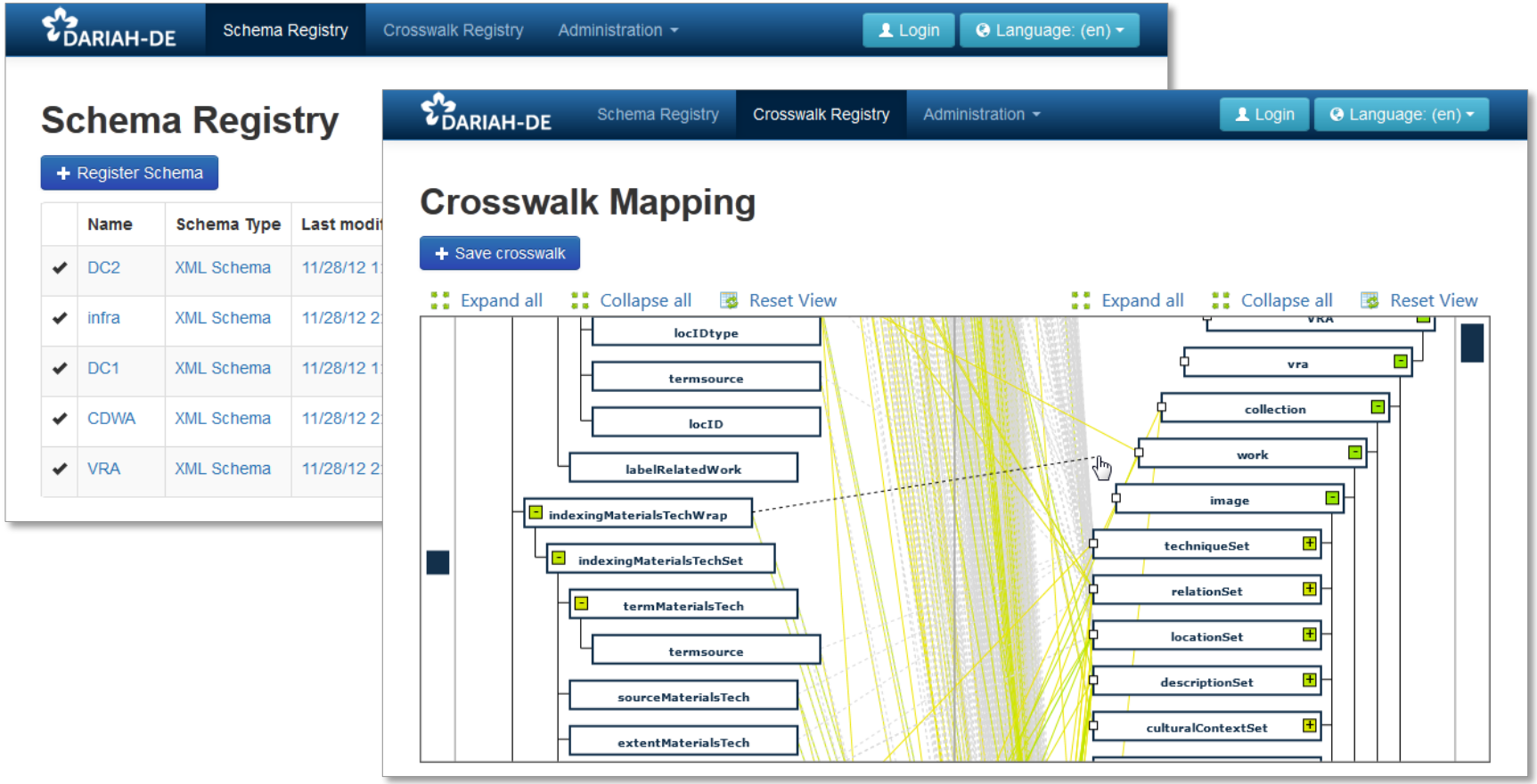
Query Processing



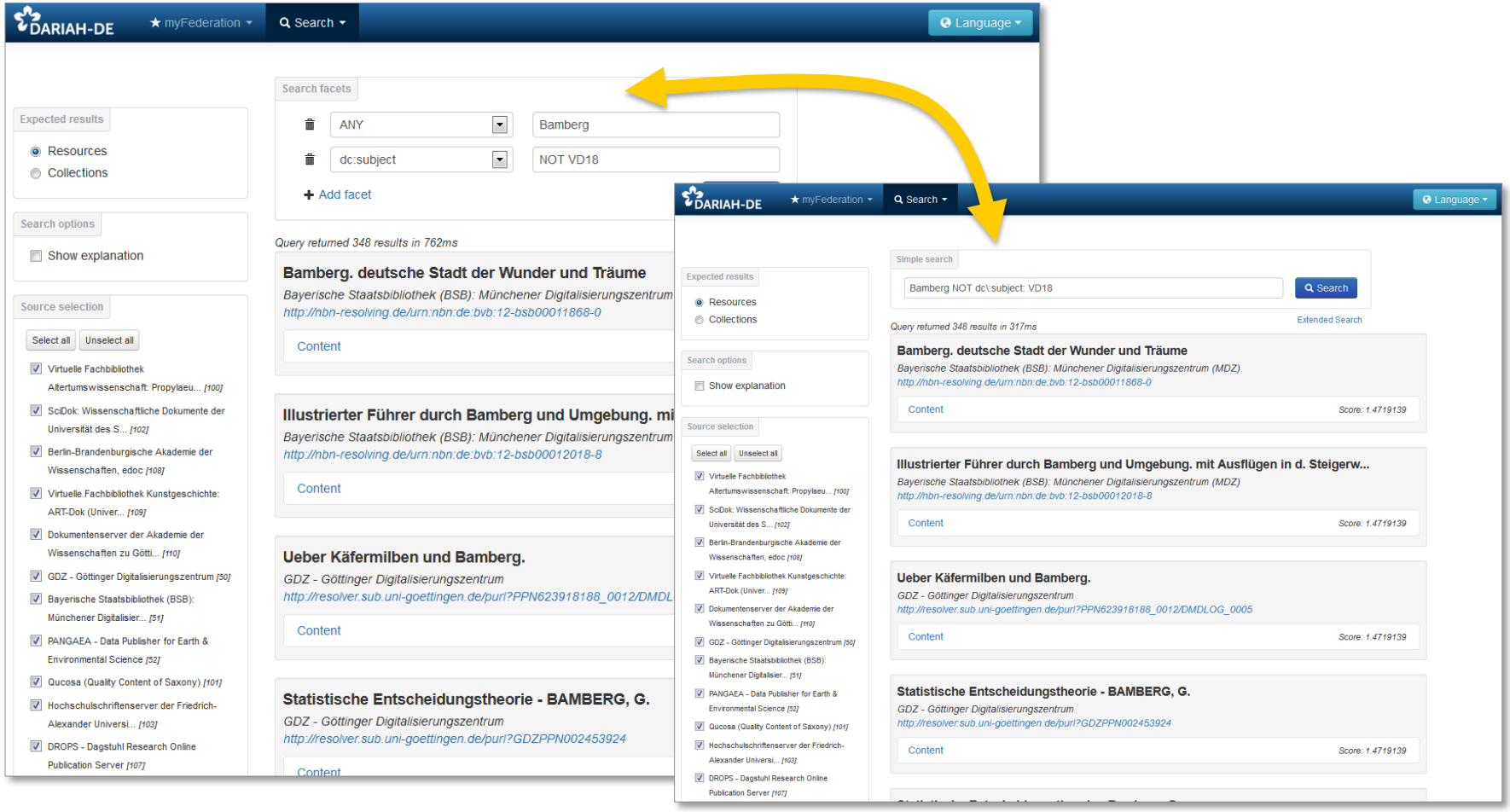
DARIAH-DE Generic Search

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

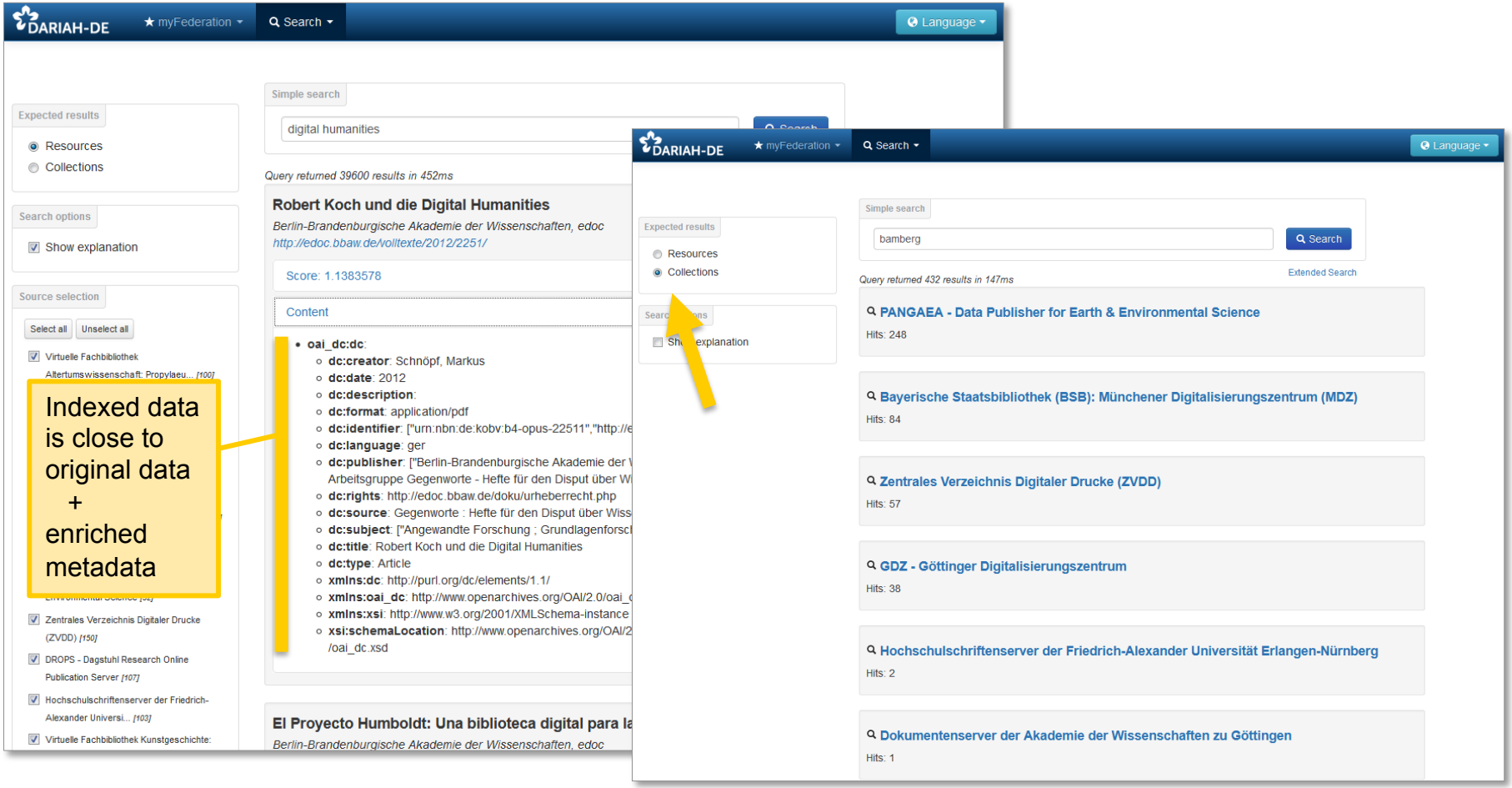
Schema and Crosswalk Registry



Search: Facetted / simple



Search: Result presentation



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