DARIAH-DE
Digital Research Infrastructure
for the Arts and Humanities

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Die DARIAH-DE Föderationsarchitektur

DARIAH-DE Methodenworkshop

„Quantitative Vorauswahl und Validierung
für ein qualitatives Arbeiten in den
Geisteswissenschaften: Ein iterativer Prozess?“

Tobias Gradl

Überblick

1. Begriffe, Architektur, Überblick
2. Strukturelle Anreicherung

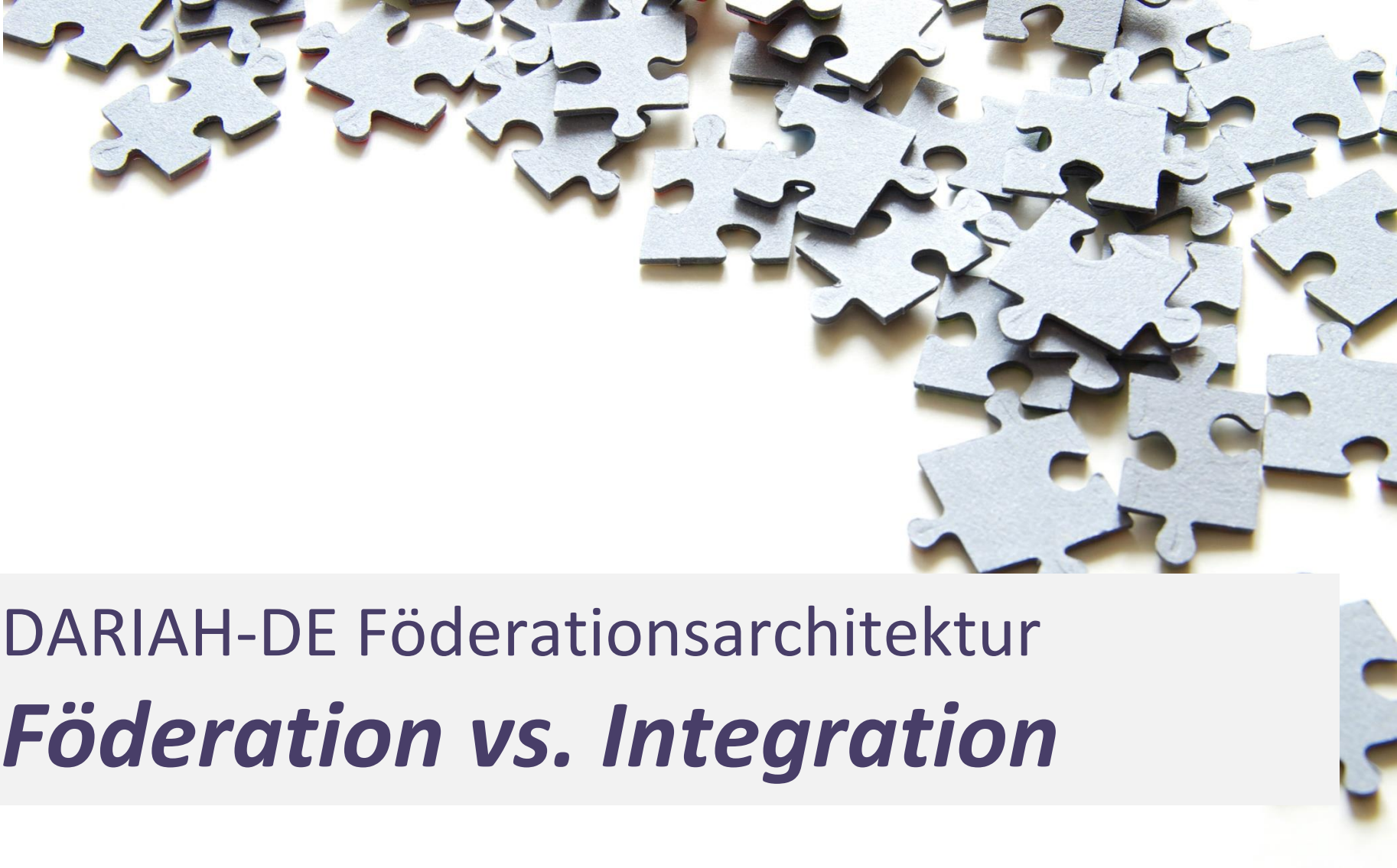
Live Session I

3. Mappings / Transformation

Live Session II

4. Fragen / Diskussion

DARIAH-DE



DARIAH-DE Föderationsarchitektur

Föderation vs. Integration

DARIAH-DE

Dimensionen der Informationsintegration

logisch

geographisch

- *Verteilung*
- *Autonomie*
  - *der Organisationen*
  - *zu integrierender Systeme*

- *Heterogenität*
  - *Schnittstellen*
  - *Daten*

technisch

syntaktisch

strukturell

semantisch

*jeweils **starke Ausprägung** in den Kultur- und Geisteswissenschaften **(bei breiter Betrachtung)***

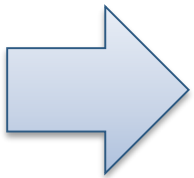
DARIAH-DE

Heterogenität nicht nur „Problem“

- *Integrationsproblem*: Verteilung und Heterogenität verhindern integrierte Sicht auf Daten
- Aber auch:
  - Daten entstehen im Kontext spezifischer Forschung
  - (Entstehungs-)Kontext wichtig für Interpretation

Interdisziplinäre Forschung?

Übergreifende Dienste



*Heterogenität spiegelt **Diversität der Domäne***

DARIAH-DE

Grundannahme

z. B. Standards

akuter Bedarf

Daten sind in einer bewusst oder unbewusst festgelegten Form zugreifbar; eine Umwandlung in eine andere Form ist oft verlustbehaftet.

- Informationsverlust kann akzeptabel sein, wenn Verwendungskontext definiert

vgl. WiWi

- Gegensätzliche Zielsetzungen durch:
  - Diversität der Forschungsdaten, -fragen und -kontexte
  - Notwendigkeit übergreifender Auswertungen

DARIAH-DE

Integration von Daten

- *Datenintegration* versucht die Harmonisierung heterogener Darstellungsformen
  - *eine globale Sicht* für integrative Betrachtung
  - oft als einmaliger Aufwand realisiert
  - Spezifität von Daten irrelevant, wenn nicht in Integrationssicht assoziierbar

z. B. ETL

Lenzerini, M (2002): Data Integration: A Theoretical Perspective. In: Abiteboul, S (Ed.): Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, p. 233. ACM, 2002. ISBN:9781581135077.

Föderation von Daten

- *Datenföderation* beschreibt eine losere Zusammenführung von Daten
  - Daten bleiben im ursprünglichen Zustand erhalten
  - Aufwand je notwendiger Sicht erforderlich
    - globale Sichten auch hier als einmaliger Aufwand
    - aber: mehrere „globale“ Sichten denkbar
  - Integrationsansätze nach spezifischen Bedürfnissen

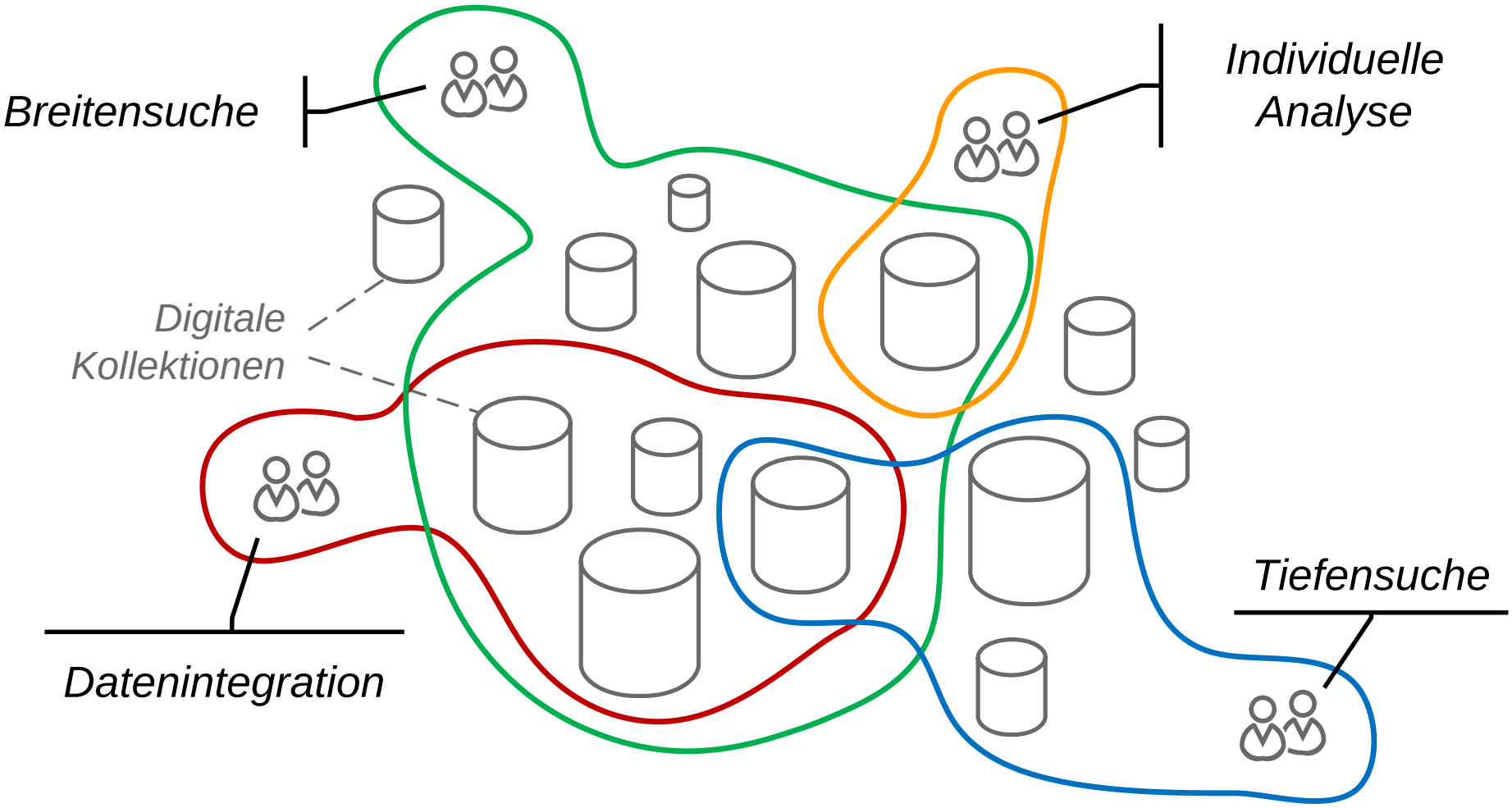
Standards

domänenspezifisch

Gradl, T.; Henrich, A. (2014): A novel approach for a reusable federation of research data within the arts and humanities, in: Digital Humanities 2014: Book of Abstracts. Lausanne: Ecole polytechnique fédérale de Lausanne; Université de Lausanne 2014, S. 382–384.

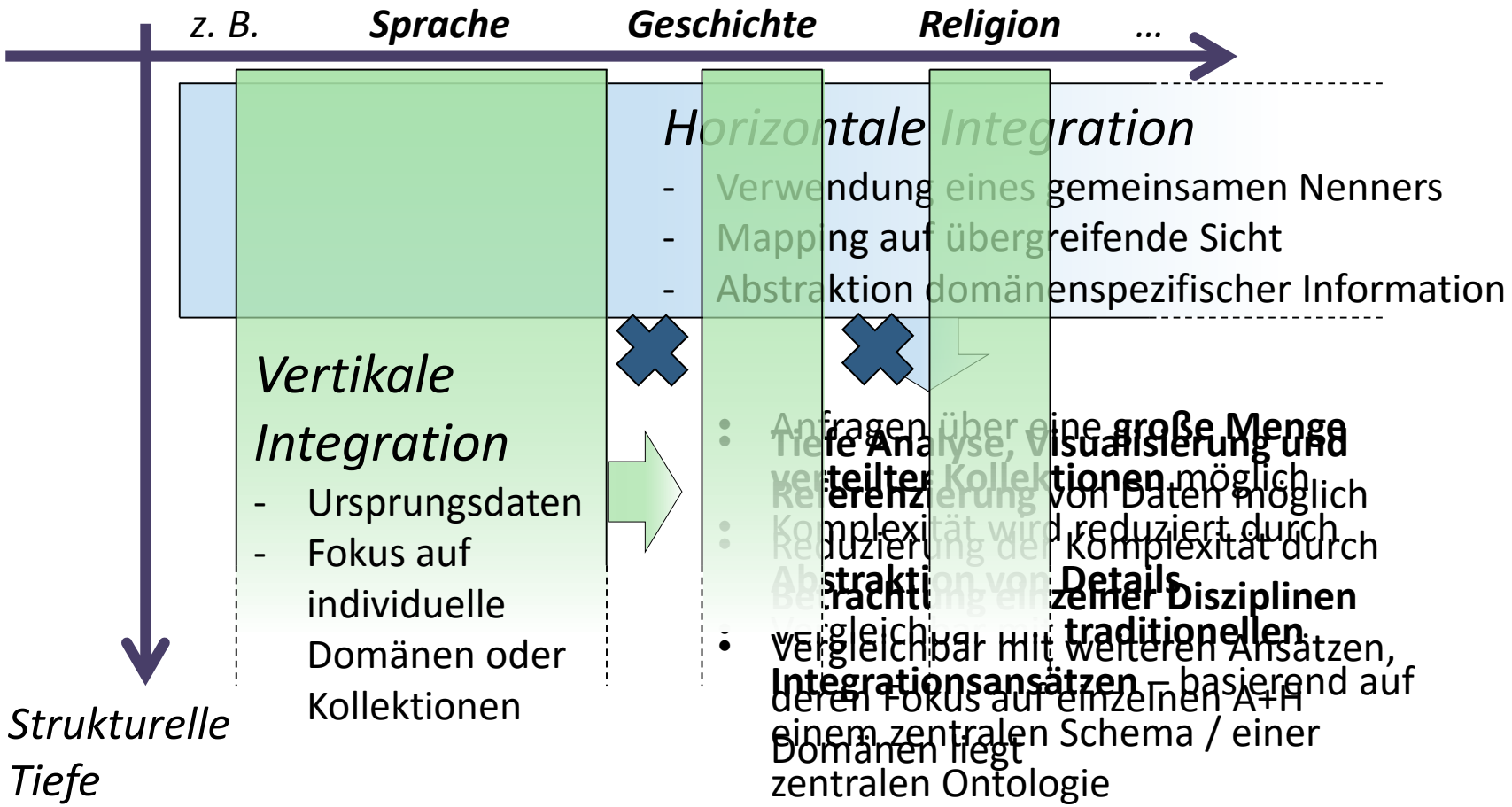
DARIAH-DE

Grundlegende Anwendungsfälle



DARIAH-DE

Horizontale und vertikale Integration



DARIAH-DE

DARIAH-DE Föderationsarchitektur

Ziele und Funktionen

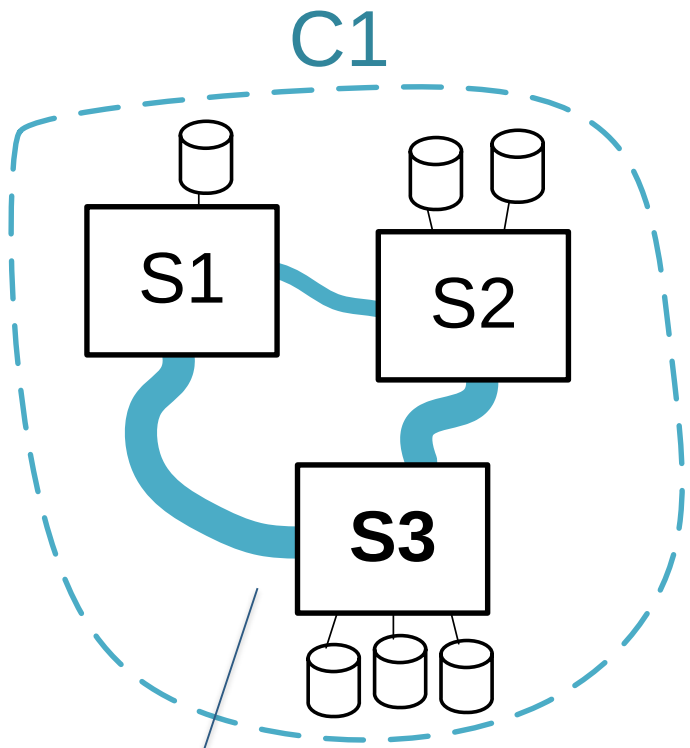
DARIAH-DE

Ziele der Föderationsarchitektur

- Verzeichnung von *Forschungsdaten-sammlungen* und verwendeten *Datenmodellen*
- *Kombinierte Unterstützung:*
  - *Domänenspezifische Integration auf Basis spezifischer inhaltlicher Bedürfnisse*
  - *Übergreifender Zugriff* auf Inhalte heterogener Sammlungen durch Standardschemata
- (Exemplarische) Anwendungen der Suche und Visualisierung von Kollektionen und Inhalten

DARIAH-DE

Idee: *Forschungsorientierte Föderation*



Beispiel: drei Repräsentationsformen biographischer Daten

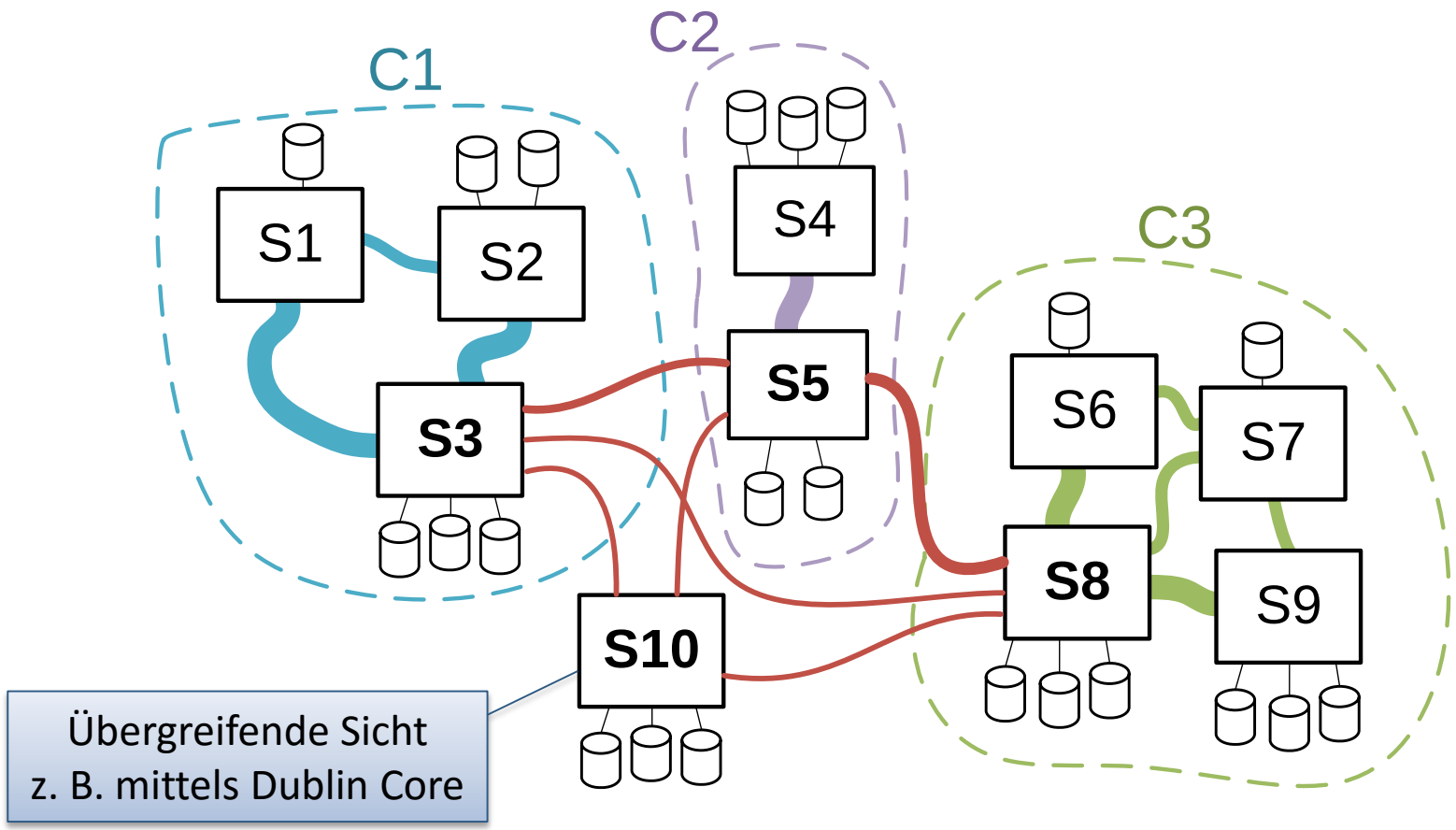
Fachwissenschaftler…

1. identifizieren relevante Kollektionen
2. wählen semantisch „passende“ Export-schemata aus
3. assoziieren Konzepte der gewählten Schemata

Ergebnis: Semantisch eng korrelierte Datenbasis

DARIAH-DE

Idee: *Forschungsorientierte Föderation*

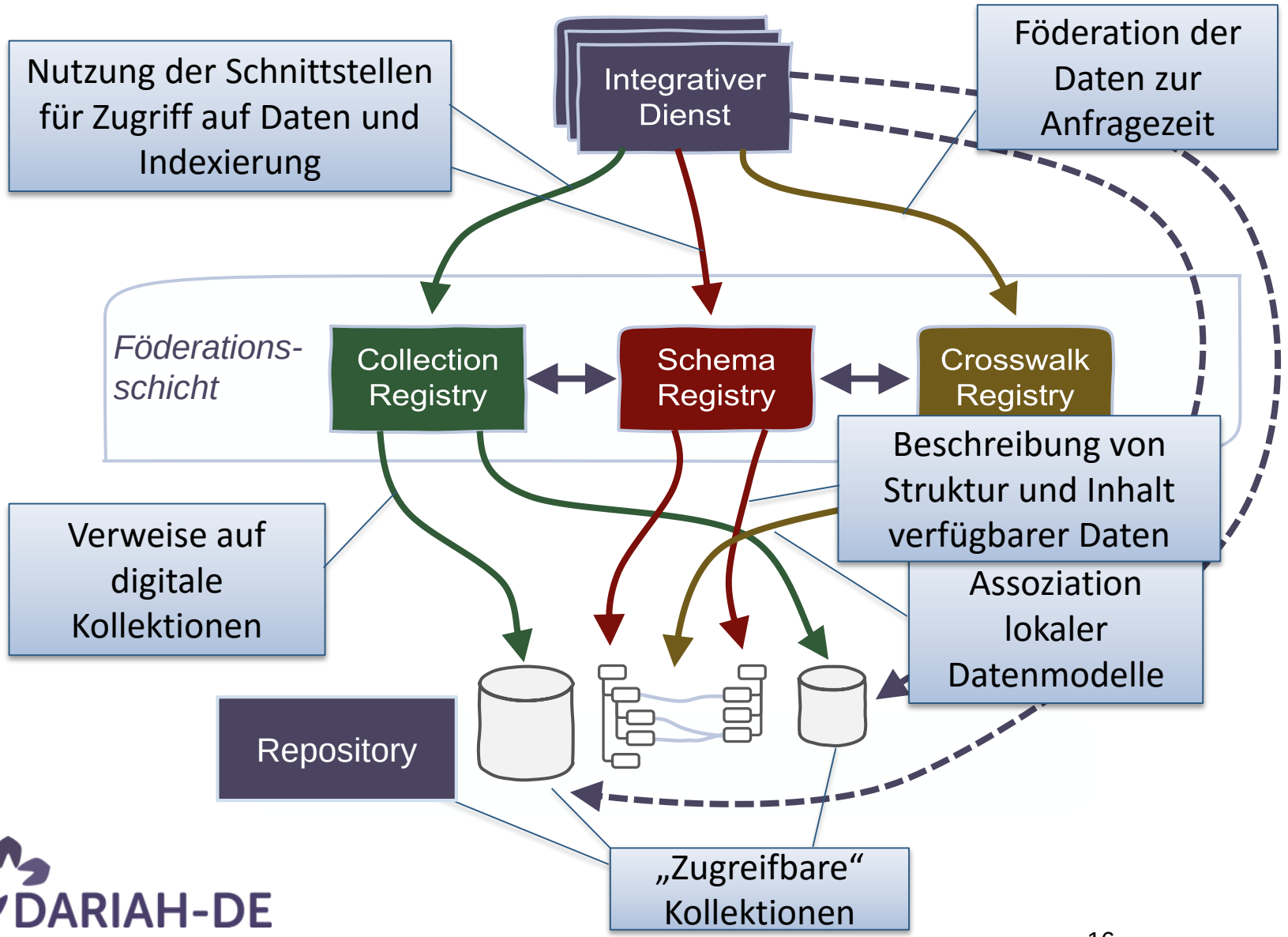


DARIAH-DE

DARIAH-DE Föderationsarchitektur
Architekturkomponenten

DARIAH-DE

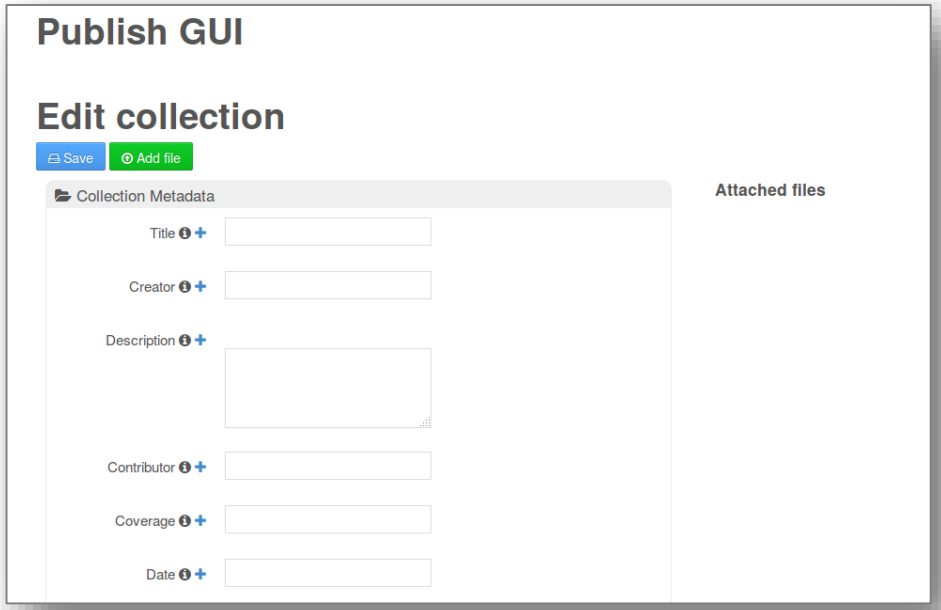
Komponenten des Föderationskonzepts



DARIAH-DE

Repository

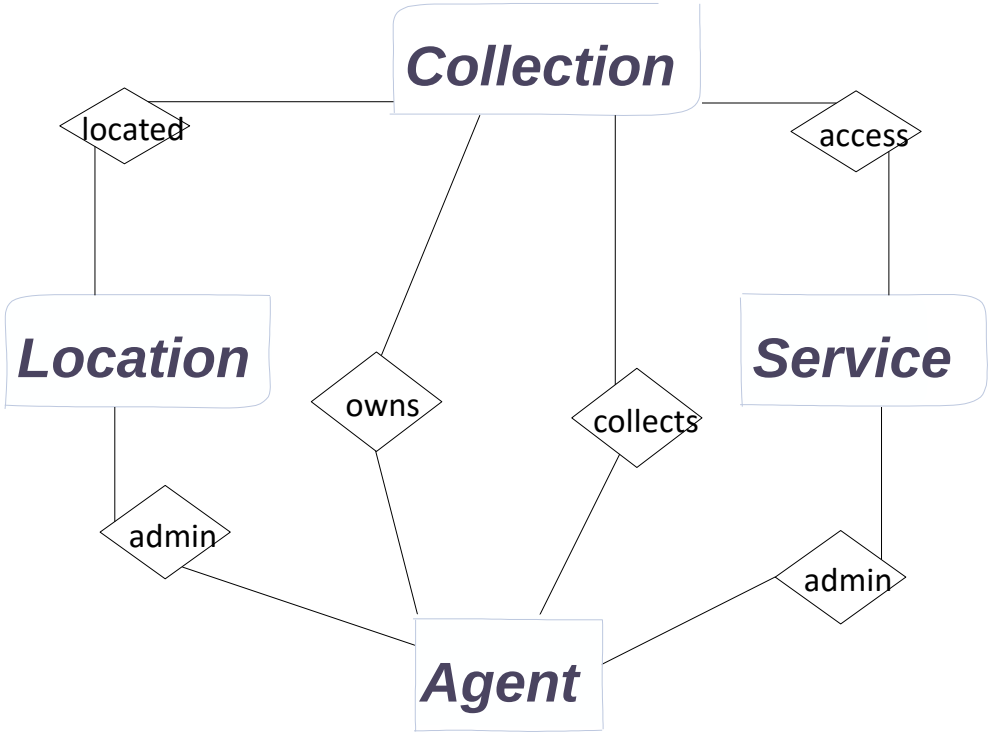
- Forschungsdatenarchiv mit *Publish-GUI* zur Publikation von Forschungsdaten
- Daten sind *sicher* und *nachhaltig referenzierbar* (PIDs) gespeichert
- Metadaten in *DC*
- *OAI-PMH* Export



DARIAH-DE

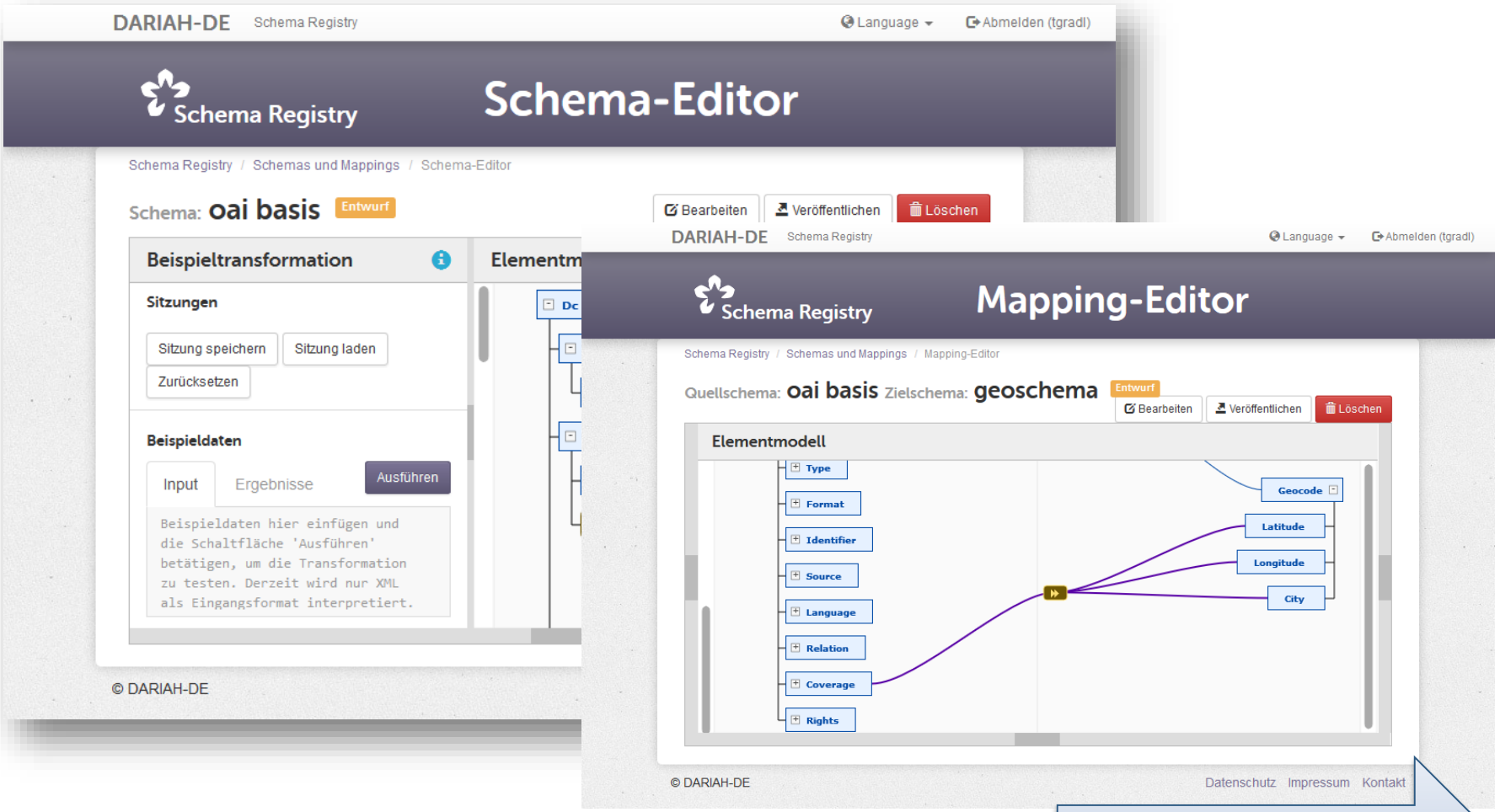
Collection Registry

- Verzeichnis für Forschungsdaten-sammlungen
- Suchen, Bearbeiten, Reviewen
- Zugriffsdienste zu Sammlungen
- OAI-PMH Export



DARIAH-DE

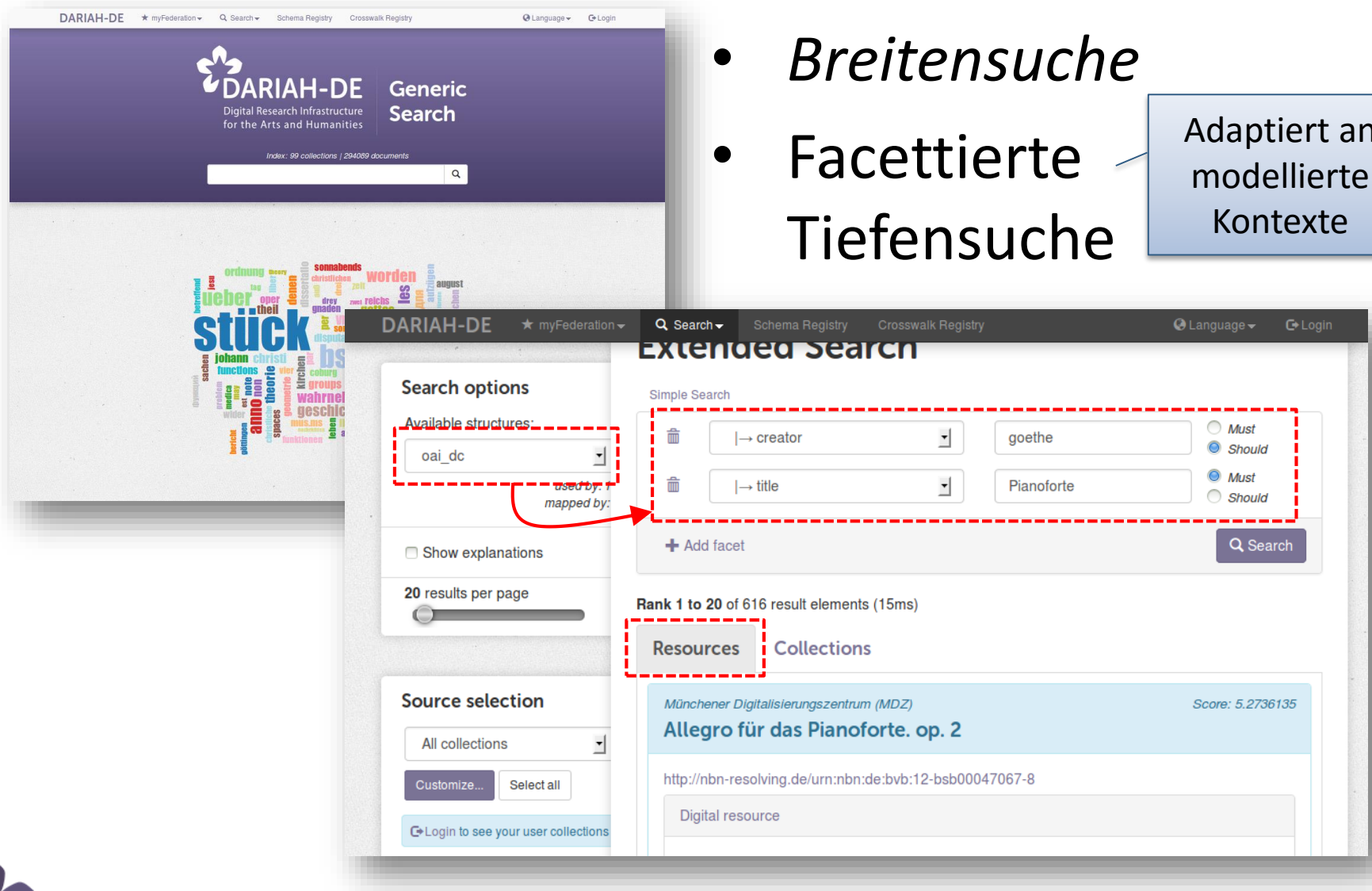
Schema Registry



Live-Sessions

DARIAH-DE

Generische Suche

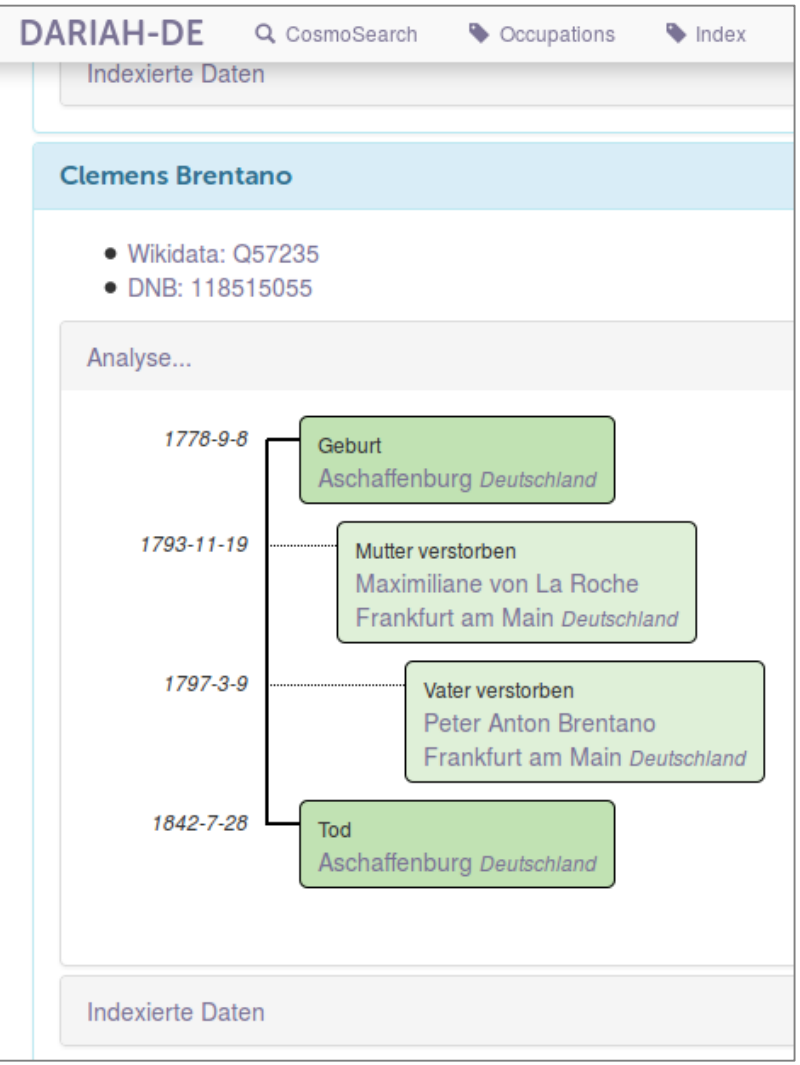


- *Breitensuche*
- Facettierte Tiefensuche

Adaptiert an modellierte Kontexte

DARIAH-DE

aber auch: Cosmotool



- *Biographische Profile aus verschiedenen Quellen*
- Kontextspezifische Auswertung der Daten
- Transformation in integriertes Datenmodell
- Assoziationen der Schema Registry

DARIAH-DE

Funktionsprinzip

Die Sammlung wird eingetragen und die Zugriffsschnittstelle auf Daten der Sammlung beschrieben.

Forscherin möchte Daten einer Sammlung, für die es keine geeigneten Suchmöglichkeiten gibt, analysieren und vergleichen.

Collection Registry

Sammlung beschreiben

1

2

Das Schema der Daten wird ausgewählt oder neu angelegt.

Titel

Author

Schema definieren

Schema Registry

3

Mappings erstellen

Crosswalk Registry

4

Generische Suche

Suche

Schemata werden miteinander assoziiert.

Die generische Suche indiziert die Daten der Kollektion.

Nachnutzung

Forscherinnen und Forscher können Sammlungsbeschreibung, Schema, Mapping und Suche für ähnliche Anfragen nachnutzen.

DARIAH-DE

DARIAH-DE Föderationsarchitektur

Strukturelle Anreicherung

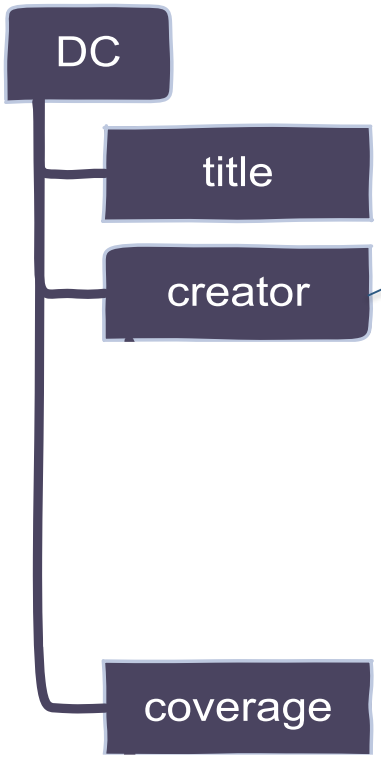
DARIAH-DE

Fachliche Datenaufbereitung

- Schema Registry unterscheidet *zwei Phasen*:
  - *Datenaufbereitung*: Daten nach Vorgaben des Modellierers evaluieren und in erweiterter Form abgespeichert
    - dadurch „Integrationsvorbereitung“ bei Import
    - Original bleibt erhalten
    - generisch oder kontextspezifisch
  - *Mappings & Transformation*: Überführung von (aufbereiteten) Daten in Zielform
    - Aufwand reduziert
    - kontextspezifisch

DARIAH-DE

Beispiel: Pangaea



Beispiele:
- *Substrukturen z. B. Listen*
- *Muster wie „Nachname, Vorname“*

Beispiel für Substruktur

```
...
<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

DARIAH-DE

Aufgabe 1: Datenbeschreibung

- *Beschreibung der Daten* mit dem Ziel der Einarbeitung von Hintergrundwissen zu Daten in ihrem Erstellungskontext
  - durch Domänenexperten
  - Daten werden ohne Rücksicht auf spätere Verwendung erweitert
  - Kombination explizit vorliegender Strukturinformationen und Wissen um Instanzen

DARIAH-DE

Aufgabe 1: Datenbeschreibung

- Anwendung von Methoden aus dem Bereich *sprachbasierter Anwendungen*:
  - Kernidee: Daten werden durch grammatikalische Regeln definiert und erklärt
  - Es werden *domänenspezifische Sprachen* (domain specific language, DSL) konstruiert und mit schematischen Informationen hinterlegt
  - Explikation von Wissen auf Instanzebene

DARIAH-DE

Aufgabe 1: Datenbeschreibung

Kontextfreie Grammatik

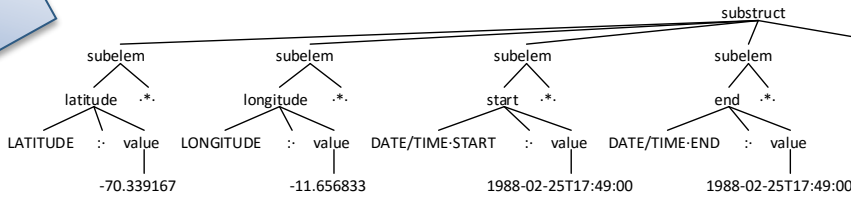
EBNF

```
grammar PangaeaCoverage;

substruct     : subelem+;
subelem       : (longitude | latitude | start
                 | end | minDepth | maxDepth
                 | otherElem) SEPARATOR?;

longitude : 'LONGITUDE' ': ' value;
latitude : 'LATITUDE' ': ' value;
...
otherElem : key ': ' value;

key     : ID;
value   : DATE
        | ID;
...
```

Modelliert durch Domänenexperten

22 Zeilen insgesamt

Syntaxbaum



DARIAH-DE

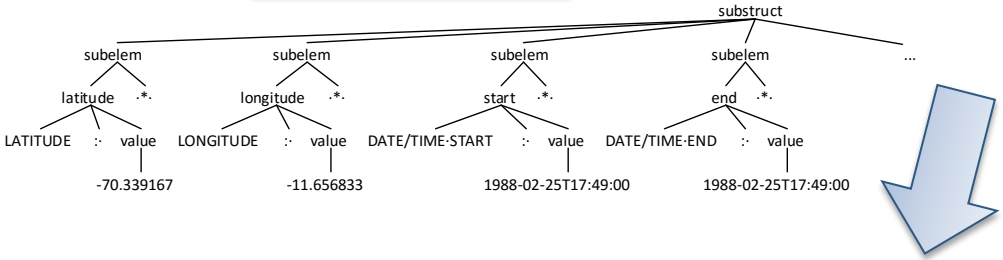
Aufgabe 2: Datentransformation

- *Beschreibung von Transformationsregeln* zur Verfeinerung, Erweiterung, Bereinigung von Daten
  - Ebenfalls durch Domänenexperten
  - wenn möglich, auch hier ohne konkrete Ausrichtung auf Verwendungskontext
  - Vorbereitung der Umwandlung von Daten in andere Formate, Befragung externer Quellen wie Thesauri, Ontologien etc.

DARIAH-DE

Aufgabe 2: Datentransformation

Syntaxbaum




```
lon = @longitude.value;

lat = {
        value = @latitude.value;
};

combined = CONCAT("[", @latitude.value, "-"
                , @longitude.value, "]");
```

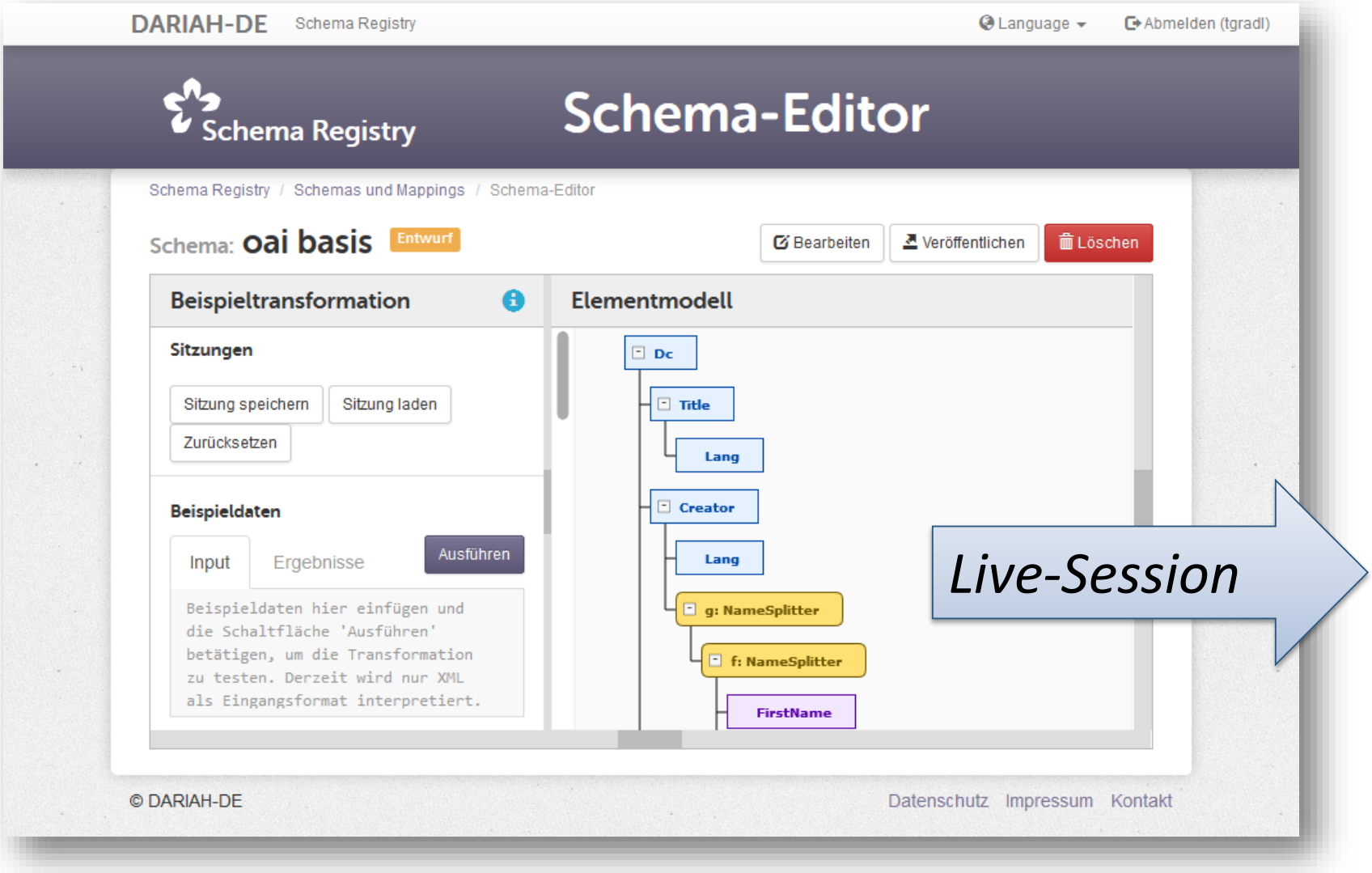
Zuweisung

Generierung untergeordneter Elemente

Erweiterbarer Funktionssatz

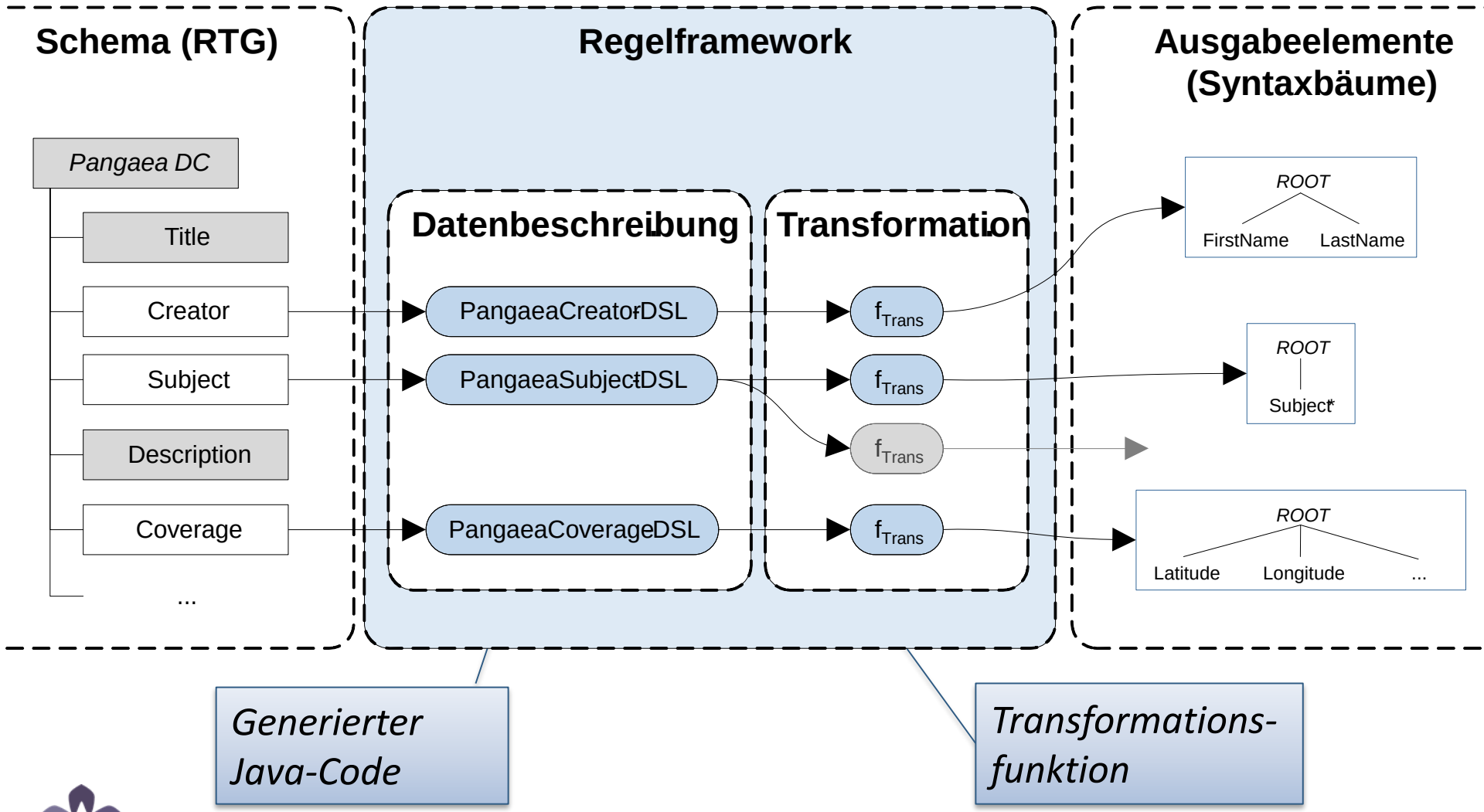
DARIAH-DE

Session 1: Datenbeschreibung



DARIAH-DE

Gesamtüberblick



Beispiel: Wikipedia

```
<page>
    <title>Lujo Brentano</title>
    <ns>0</ns>
    <id>142397</id>
    <revision>
    <id>134487883</id>
    <parentid>133924296</parentid>
    <timestamp>2014-09-30T13:12:04Z</timestamp>
    <contributor>
        <username>Gelehrter11</username>
        <id>993293</id>
    </contributor>
    <minor/>
    <comment>/* Leben */ Spätere Distanzierung vom "Manifest der 93" gem. dem
    Artikel "Manifest der 93"</comment>
    <text xml:space="preserve">...
    * [[19. Dezember]] [[1844]] in [[Aschaffenburg]]; † [[9. September]]
    [[1931]] in [[München]]) war ein [[Deutschland|deutscher]]
    [[Volkswirtschaftslehre|Nationalökonom]] und [[Sozialreform]]er.

    == Leben ==
    Lujo Brentano wurde hineingeboren in die prominente, aus Italien stammende
    katholische Intellektuellenfamilie Brentano: ...
```

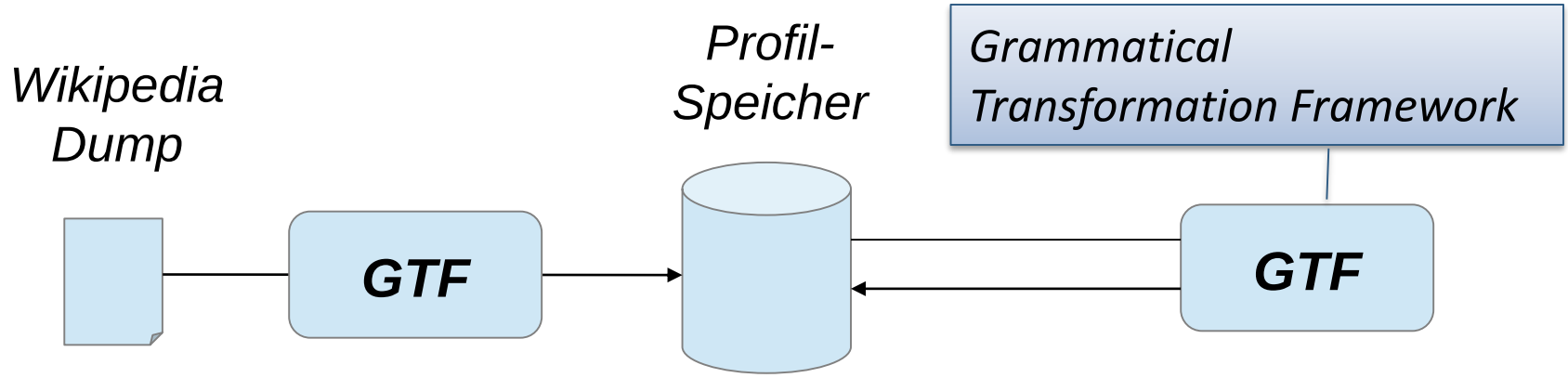
Metadaten eher irrelevant

Biographische Daten im Fließtext

Aber: relevante Substruktur

DARIAH-DE

Beispiel: Wikipedia → Cosmotool

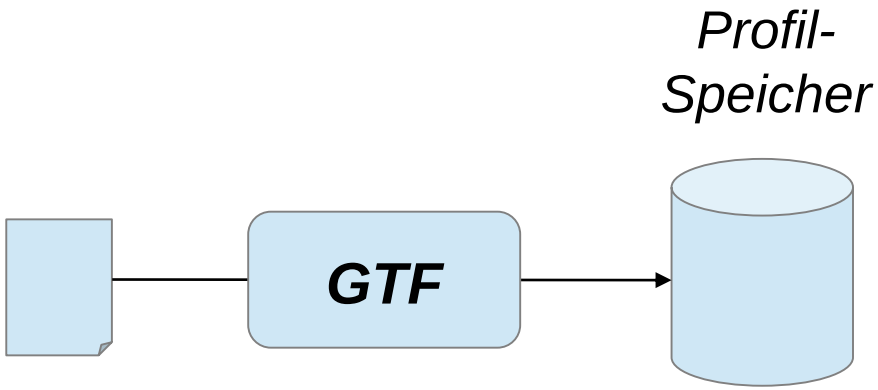


- *Grammatik: Definition sprachlicher Elemente*
- *Extraktion von „Inhalt“*
- *Transformation des Inhalts in Elemente des Profils*
- *Ergebnis: Paare von Überschriften und Text*

- *Anwendung technischer Verfahren*
- *Iteration im Beispiel:*
  - *Satzerkennung*
  - *Anwendung typischer NLP-Verfahren (mehrfach)*
  - *Ableitung von Korrelationen*

Wiki-Entities

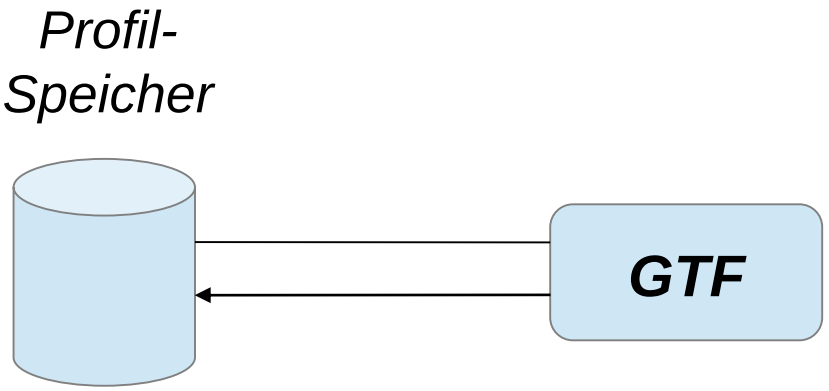
Implementierte Adapter



- *Datei*: wird genutzt für Wikipedia/Wikidata-Dumps
- *OAI-PMH* (aus generischer Suche): derzeit noch nicht verwendet
- *HTTP-GET*: Abruf aus Webseiten

Bezeichner als GET-Parameter

Implementierte Analysefunktionalität



Generisch wiederverwendbar

- Core
  - String-Funktionen
  - Logische Funktionen
- NLP Adapter
  - Stanford
  - OpenNLP
- Biographische Korrelationsanalyse

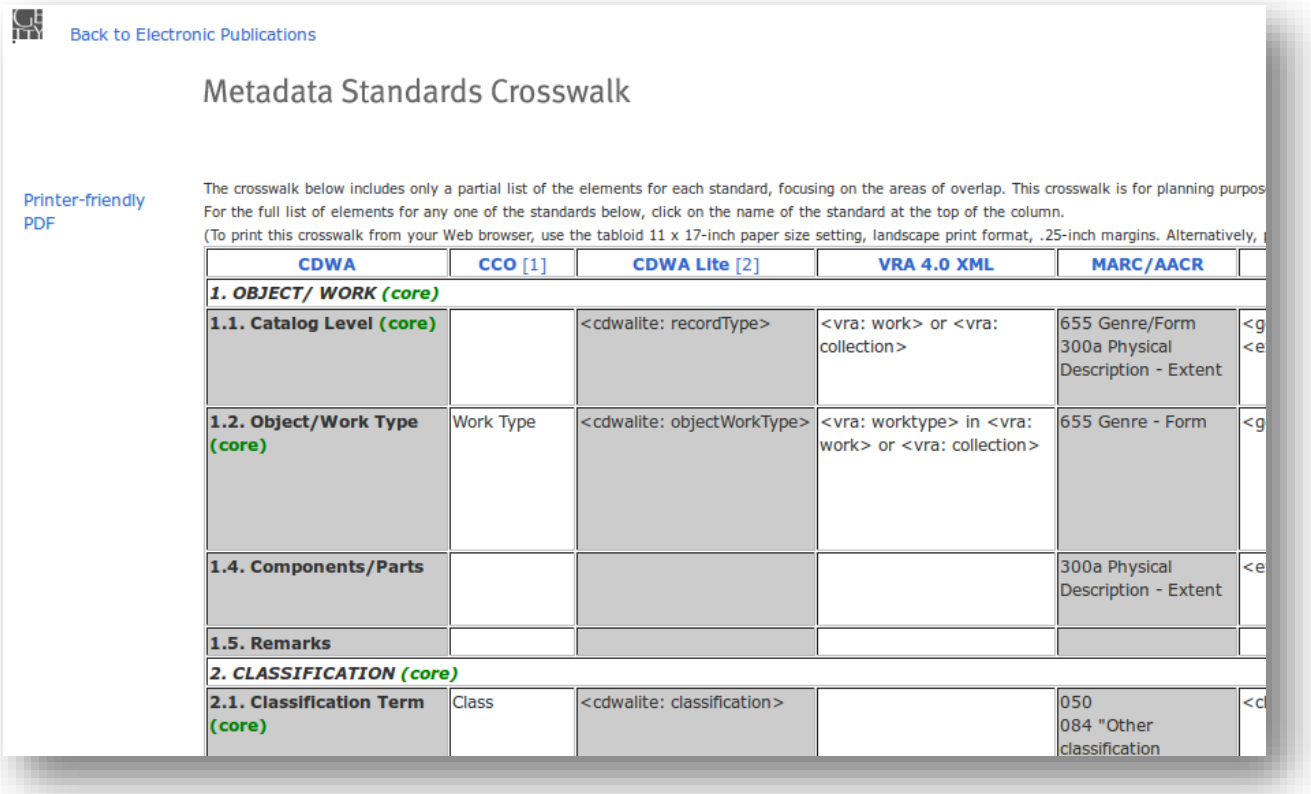
DARIAH-DE Föderationsarchitektur

Datentransformation

DARIAH-DE

Einschränkung generischer Crosswalks

- Keine Abbildung kollektionsspezifischer Verwendung
- Nur einfache Assoziationen zw. Elementen möglich



Back to Electronic Publications

Metadata Standards Crosswalk

Printer-friendly PDF

The crosswalk below includes only a partial list of the elements for each standard, focusing on the areas of overlap. This crosswalk is for planning purpos
For the full list of elements for any one of the standards below, click on the name of the standard at the top of the column.
(To print this crosswalk from your Web browser, use the tabloid 11 x 17-inch paper size setting, landscape print format, .25-inch margins. Alternatively,

CDWA	CCO [1]	CDWA Lite [2]	VRA 4.0 XML	MARC/AACR	
1. OBJECT/ WORK (core)					
1.1. Catalog Level (core)		<cdwalite: recordType>	<vra: work> or <vra: collection>	655 Genre/Form 300a Physical Description - Extent	<g <e
1.2. Object/Work Type (core)	Work Type	<cdwalite: objectWorkType>	<vra: worktype> in <vra: work> or <vra: collection>	655 Genre - Form	<g
1.4. Components/Parts				300a Physical Description - Extent	<e
1.5. Remarks					
2. CLASSIFICATION (core)					
2.1. Classification Term (core)	Class	<cdwalite: classification>		050 084 "Other classification	<cl

http://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html

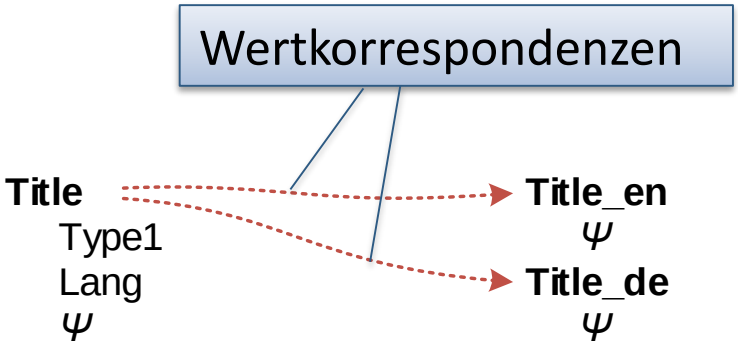
DARIAH-DE

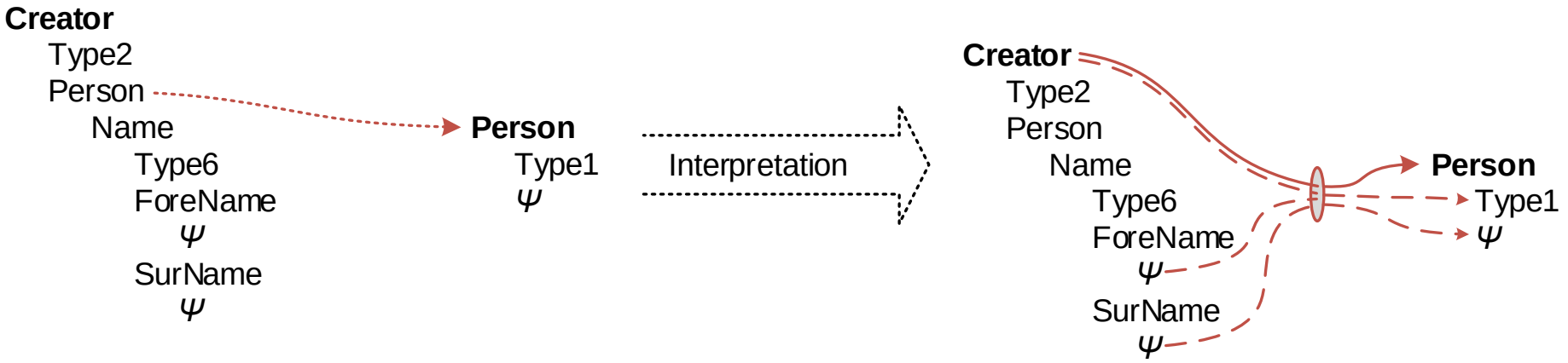
Aufgaben der Mappings

- wie bei Erweiterung von Schemata sind auch hier *Beschreibung und Transformation der Daten möglich*
- Hier aber: Fokus auf *Verwendungskontext*
- Modellierer: *Experte der Anwendungsdomäne*

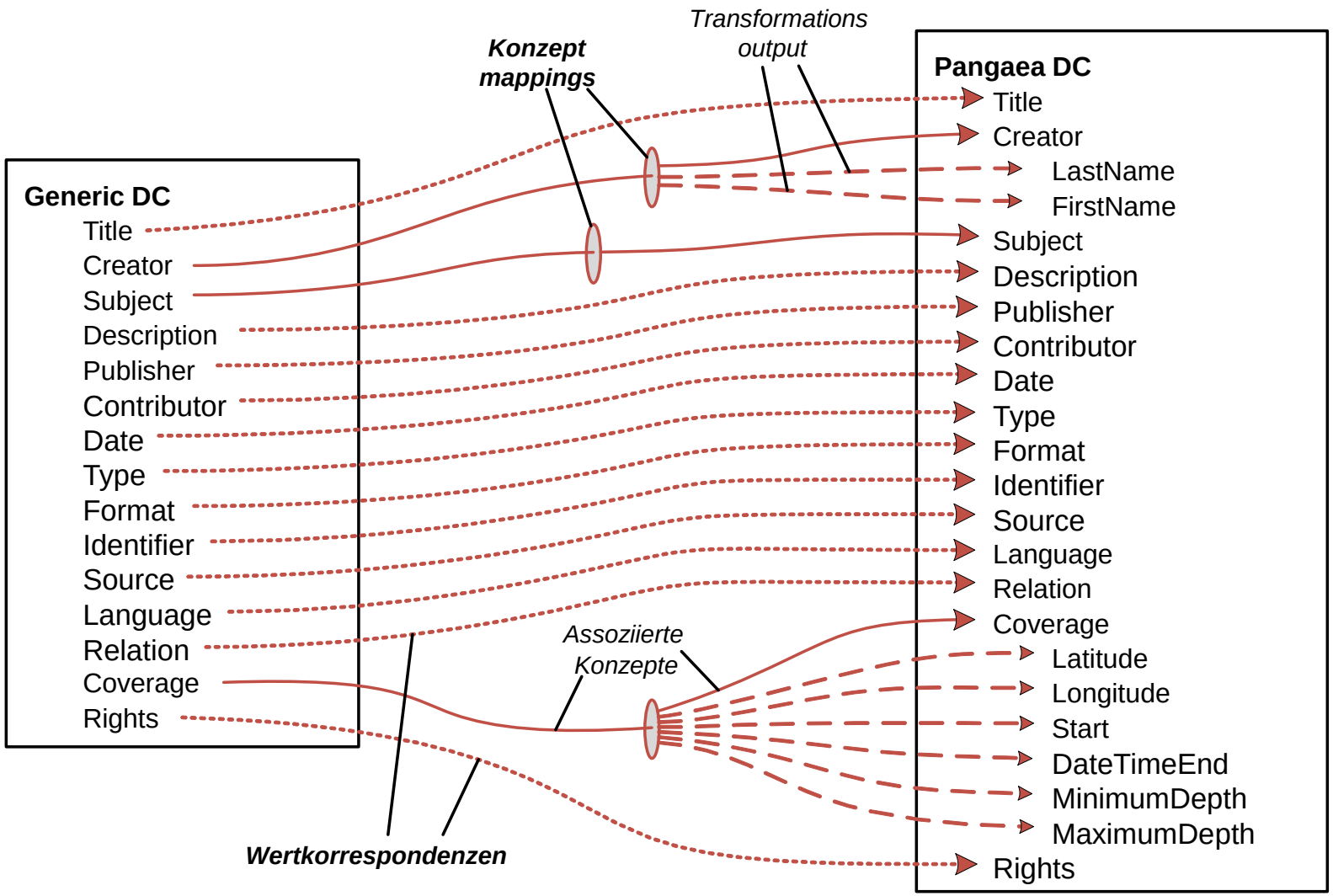
DARIAH-DE

Mappings verfeinern





Mappings vs. Anreicherung



DARIAH-DE

Session 2: Mappings

DARIAH-DE

Vielen Dank! Fragen?

DARIAH-DE