Digital Research Infrastructures for the Humanities – Approaches and Challenges for Retrieval Applications

Prof. Dr. Andreas Henrich
Otto-Friedrich-Universität Bamberg
Chair of Media Informatics
Overview

1. Digital Humanities
2. DARIAH-DE
3. Searching Federated Collections
4. Curricular Considerations
1. Digital Humanities

Digital ???

Fields according to Wikipedia:

- Anthropology
- Classics
- History
- Geography
- Languages
- Law
- Literature
- Performing arts (Music, Theatre, Dance)
- Philosophy
- Religion
- Visual arts

Heterogeneous in many respects: subject-matter current use of digital methods...

image source: www.colourbox.de, 2652551
Aspect of digital humanities include:

- humanities research enabled through digital media, data mining, software studies, or information design and modeling;
- computer applications in literary, linguistic, cultural, and historical studies, including electronic literature, public humanities, and interdisciplinary aspects of modern scholarship;
- digital arts, architecture, music, film, theatre, new media, digital games, and related areas;
- creation and curation of humanities digital resources;
- social, institutional, global, multilingual, and multicultural aspects of digital humanities; and
- digital humanities in pedagogy and academic curricula.

Call for Papers DH 2015 (http://dh2015.org/cfp/)
2. DARIAH-DE

errichtet vernetzt stellt bereit fördert definiert

\{ e-Infrastruktur Curricula Standards Schulungen Forschungsdaten \}

für \{ Fachwissenschaftler Lehrpersonal Forschungsinstitutionen Datenzentren Forschungsprojekte \}
... supports digitally-enabled research and teaching in the arts and humanities

- generating new research questions
- answering old research questions with new digital methods

... is developing a research infrastructure which offers:

- Tools
- generic and discipline specific services
- access to research data as well as materials for research and education in the Digital Humanities
- curricula and qualification instruments for students, PhD-candidates and academics
DARIAH-DE: DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES

TEACHING
- Workshops on Methods
- Colloquia for Experts
- Workshops for Academics
- Summer Schools
- Curricula

RESEARCH
- Initial Research Interest
- Research Question
- Information Gathering
- Analysis
- Interpretation
- Publication

Digital Methods and Practices
- Demonstrators and Academic Services
  - MEISE
  - DIGIVOY
  - Virtuelles Skriptorium
  - Person Data Repository
  - Interoperability Authority files
  - Monasterium

- Generic Services
  - Development on Demand
  - Geo-Browser
  - Archive in a Box
  - Authority file Services
  - Generic Search
  - Collection Registry
  - Schema Registry

- Software Hosting Services

- Generic Search
  - Collection Registry
  - Schema Registry

RESEARCH DATA
- Metadata
- Legal Aspects
- Ontologies
- Research Data Management
- Crosswalks

TECHNICAL INFRASTRUCTURE
- Basic Services
  - Collaborative Working Environments (Wiki, Developer Portal, Etherpad)
  - PID Service
  - Storage, Archive
  - Databases

- Provision of Virtual Machines

Software Hosting Services

Platform und Infrastructure Hosting Services

Operational IT Services
- AAI
- Security
- Quality Control
- Resources
- Monitoring
- Help Desk

DARIAH-DE
June 2013
Research Data and Collections

- Documents and sources in the field of the Arts and Humanities
- Collected and compiled by academics
- In many ways **multifaceted**:  
  - Different media formats  
  - Provenience and context  
  - Based on the research question and the research methods  
  - Granulation  
  - Language  
  - etc.
Example: epidat – epigraphic database

• **Collection:** Database (25,000 inscriptions, over 130 Jewish Graveyards) from Germany and the Netherlands (1050-2000) including e.g. transcriptions, maps, photos, digital copies and archival resources

• **Question:** Analyzing historical changes of societies

• **Approach:** Visual Analytics of topographical structures; Named-Entity-Recognition; Spatial-temporal Analytics – DARIAH-DE Geo-Browser

![epidat - epigraphische Datenbank](http://steinheim-institut.de/cgi-bin/epidat)
3. Searching Federated Collections

Overview

– Introduction
– Federation
– Generic Search
– Use Case Biographies

image source: www.colourbox.de, 9676365
Dimensions of Search / Information Retrieval

Content
- unstructured text
- structured text
- database records
- images
- videos
- scanned documents
- audio
- music
- 3D models
- motion sequences
- gene sequences
...

Applications
- web search
- vertical search (only one domain)
- enterprise search
- desktop search
- forum/blogs search
- P2P search
- literature search

Tasks
- known item search
- question answering
- exploration / overview
- classification
- filtering
- finding correlations
...

What is “typical” in DH
### Dimensions of Search / Information Retrieval for DH

#### Content
- unstructured text *(maybe)*
- structured text
- database records *(maybe)*
- images *(maybe)*
- videos *(maybe)*
- scanned documents *(maybe)*
- audio *(maybe)*
- music *(maybe)*
- 3D models
- motion sequences
- gene sequences

#### Applications
- web search
- vertical search *(only one domain)*
- enterprise search
- desktop search
- forum/blog search
- meta search
- P2P search
- literature search

#### Tasks
- known item search
- question answering
- exploration / overview
- classification
- filtering
- finding correlations

---

*DARIAH-DE*
Basic use-cases

- **Broad search**
  - Digital collections
  - e.g. to find collections

- **Individual Analysis**
  - e.g. “make searchable”

- **Data integration**
  - Actual integration, migration etc.
  - Reduced amount of dimensions

**Deep Search**
Components of the Federation Concept

- Use of interfaces to access and index the data sources
- References to digital collections
- Describing structure of accessible data
- Association of local data models

**Federation Level**

- Collection Registry
- Schema Registry
- Crosswalk Registry

**Accessible Collections**
search for collections
What makes data sources relevant?

- usefulness for research questions of projects, communities or scholars of the arts and humanities

- publication as semi-structured research data:
  - data follows a known data model
  - data is published through a crawlable or harvestable API
  - ideally, linked resources are available in digital form

- however, no a-priori quality limitations with respect to the content and structure of data
Abstract use-cases

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

Usage of resources / collections (querying, visualization etc.)

Enrichment of collections (detailed definition of schemata, crosswalks etc.)

- **definition, classification** and **semantic correlation** of data sources by respective domain experts
- **collaboration and discussion** about semantics of data and correlations between sources
Federating digital resources of the arts and humanities
Horizontal federation

- Use common denominator
- Broad view over heterogeneous data
- Loss of domain-specific information

- Queries over a large amount of diverse data sources
- Reduction of complexity by abstracting from the specificities of data, context and/or discipline
- Compares with traditional repository federations based on global data models

Focused domain(s)

e.g. Linguistics History Religion

Depth of correlations (Structure, Semantics)
Vertical federation

- Original data
- Focus on individual domains or collections

**Vertical federation**

- In-depth analysis, visualization and referencing of data
- Reduction of complexity by limiting the relevant domain
- Compares with approaches focused on specific branches of the Arts and Humanities

**Depth of correlations**
(Structure, Semantics)

*Focused domain(s)*

- e.g. Linguistics
- History
- Religion

DARIAH-DE
Our iterative approach

1. Facilitate deep federation as required by research projects and communities and their specific questions

2. Reuse explicated semantics and correlations to iteratively complement interdisciplinary or global perspectives
Step 1: Explication of semantics

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

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

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

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

\(<\text{dc:coverage}>\)
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
\(</\text{dc:coverage}>\)
Step 1: Explication of semantics

Result: Two perspectives

Parsing-oriented representation

- DC
  - title
  - creator
  - coverage

Mapping-oriented representation

- DC
  - title
  - creator
  - firstName
  - lastName
  - coverage
  - latitude
  - longitude

Required for parsing instances

Provided/referenced by collections

Enriched by domain knowledge
Primary Benefits:

– Local data models can be enriched by context knowledge
  • increases the level of detail for the (automatic) analysis and visualization of data
  • increases overall effectiveness of federated search over heterogeneous digital collections

– Original data is preserved and available for display, usage and citation provenance
Our iterative approach

1. Facilitate deep federation as required by research projects and communities and their specific questions

2. Reuse explicated semantics and correlations to iteratively complement interdisciplinary or global perspectives

Depth of correlations (Structure, Semantics)

Focused domain(s)

e.g. Linguistics History Religion
Idea: research-oriented federation

Scholars / domain experts
1. Identify relevant collections
2. Choose appropriate export data structures (schemata)
3. associate concepts of the selected data structures

Result: Semantically correlated data sources
Step 2: Federation

The premise of federated search: Mappings have been specified
Idea: research-oriented federation

Broad views by means of generic standards
Step 2: Federation

- Mappings as direct associations between collections or models.
  - Deep (local) semantics $\Rightarrow$ 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.
  - Deep (local) semantics $\Rightarrow$ low information loss.

- 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.

Symbols:
- S: Source
- T: Target
- M: Mapping
- E: Element
- C: Correlation
DARIAH-DE Generic Search: Functionality
Basic use-cases

- Broad search
- Digital collections
- Data integration
- Individual Analysis
- Deep Search
1. Phase: Data analysis and indexing

Data fetched from the data sources is then analyzed (e.g. for geo-temporal information or named entities) and gets enriched.

Rules for preprocessing and indexing of content are read from the Schema Registry.

Collection-level metadata is read from the Collection Registry (e.g. where and how?).

Data sources available to DARIAH.

Crawling of the Collections gets triggered manually or based on a schedule.

Enriched content is stored in an index that fits the original structure (with enrichments).
Crosswalks between the selected collections are parsed and utilized to translate the user query. The fit of collections gets assessed based on attributes stored in the Collection Registry and user settings.

User formulates a (faceted) query; registries and user preferences influence the execution of mappings.

Indices are queried, results are merged based on the rules fetched in (3), ranked and returned.

1. Query Processing
2. Schema/Crosswalk Registry
3. Collection Registry
4. Index
DARIAH-DE Generic Search: Prototype
Visual Information-Seeking Mantra
Overview first, zoom and filter, then details-on-demand.
[Shneiderman, 1996]
Finding relevant collections

Collections are selected as expected results

Search facets specified in Dublin Core

Aggregation of search results by collection; more sophisticated visualization techniques are currently being evaluated
Finding relevant resources

Resources are expected

Full-text query

Records are indexed and available in their originally fetched form; here in oai_dc
Custom selection of collections

Favorite collections can be grouped and shared with other users

Collection selection can be performed manually
Extended (faceted) search

Search facets can be specified in different data models

Query is translated into other schemata for which mappings exist!
Basic use-cases

Broad search

Data integration

Digital collections

Individual Analysis

Deep Search
Cosmotool: Problem in the field of History

- Biographical descriptions are often distorted by national perspectives
- Transnational biographies
  - have events in multiple countries / cultural regions
  - are often described from national perspectives and
  - require the combination of appropriate national descriptions to be formed

Data collection and filtering

- Collection of **biographical data**
  - information from **structured data**, which provides a (often coarse) biographical frame and
  - aggregate further biographical “signals” from **unstructured data**

- **Filtering**
  - Determine **probability of detected “signals”** by validating them against the biographical frame

---

*Birth, death, children born etc.*

*e.g. by means of Natural Language Processing (NLP)*
Support qualitative research

- **Selection**: Select groups by common features

- **Aggregation** of individual profiles to describe groups of persons

- **Analysis / Visualisation**: Compare persons groups to detect e.g. common mobility patterns

*Simple features: birth, death, places etc.*

*But also more vague (affiliation, profession etc.)*
Integration principle

**S1:** Wikidata
- birth
- death
- children
- etc.

**S2:** Wikipedia
- article structure and
- Structure of detected biographical claims

**S3:** „Biographical profile“
- quadruples of place, time, person, event
- confidence / probability

Analysis and visualization based on integrated data
Cosmotool prototype
Cosmotool prototype

from structured data

Extracted from full-text
Cosmotool prototype

Matched quadruple: 
*Time, loc, subject, event*
4. Curricular Considerations

http://www.dig-hum.de/arbeitsgruppe-referenzcurriculum-digital-humanities

image source: www.colourbox.de, 9718752
Positioning of Programs

- Theories
- Methods
- Research Questions

Humanities

inter-/trans-disciplinarity

DH

application domain orientation

Computer Science

Applied CS

state of the art H-programs

DH-programs

Appl. CS for H

a b c d e
Objectives for a Reference Curriculum

- [obviously: Designing good programs]

- **Standardization** of programs with centralized exams
  - Example: computer science teachers

- **Definition and delimitation**
  - What is a computer science degree program? What type?
  - Example: GI guidelines

- **Clear identification of contents**
  - For positioning
  - For orientation in heterogeneous offers
  - For simple approval of credits, Master access, ...
Example module descriptions in VAWi:

- Advantages:
  - One can deal well with optional moduls
  - Clear labelling of modules
  - Specifying corridors for a degree program according to students choice
  - Confirmation of shares also possible for individual students (\(\Sigma\) over all testified modules)

---

### Zuordnungen im 3^+ -Säulen-Modell

- 10% Wirtschaftswissenschaften
- 25% Informatik
- 60% Kerngebiete der Wirtschaftsinformatik
- 5% allg. Grundlagen & Schlüsselqualifikationen
Considerations for d and e

**Abstrakte Grundkenntnisse**
- Lifecycle Management
- Algorithmen und Datenstrukturen, insbesondere: Komplexitätsbetrachtung, Sortier- und Suchverfahren, Graphen, algorithmische Prinzipien
- Technische Grundlagen: von Neumann Architektur, Grundlagen von Betriebssystemen, Netze
- Medienformate
- Wahlfreie Ergänzungen

**Skills / Programmieren**
- Beherrschung grundlegender imperativer Prozesse, Objektorientierung. Iteration, Rekursion.
- X-Technologien (XSLT, ...)

**Modelle und Modellieren**
- Datenmanagement, einschl. Daten(bank)modelle, Datawarehousing, NRDBM
- Markupsprachen: Basistechnologien, ausgewählte Standards z.B. TEI

**Inhaltliche Anwendungen**
- Wahlpflichtkombinationen: Fachspezifische Anwendungen, darunter mindestens zwei der folgenden Anwendungsbereiche:
  - Bild
  - Audio
  - Text
  - Video
  - 3D Modellierung
  - Geodaten

**Option 1: Skriptsprache** (PHP, Python, JavaScript) unter Beachtung obiger Pflicht.
**Option 2: Objektorientierte Programmiersprache** (Java, C++).
**Option 3: Funktionale und / oder Logikprogrammierung.**

**Modellieren (z. B. mit UML)**
### Mapping concrete modules

#### Mapping to the DH reference model:

<table>
<thead>
<tr>
<th>I.1.c</th>
<th>div. media formats</th>
<th>60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.2.b</td>
<td>XML, DTD, XSLT</td>
<td>20%</td>
</tr>
<tr>
<td>I.3.b</td>
<td>DC, TEI</td>
<td>10%</td>
</tr>
<tr>
<td>I.4</td>
<td>3D modelling</td>
<td>10%</td>
</tr>
</tbody>
</table>

#### Remarks:

- Sum can be < 100%
- Level of detail in this example might be too high
- A derivation of values for “contains \(x\)% of the recommended dose for DH programs” is easily possible 😊.
If you plan an program related to DH

- Position your program

- Label your modules
Thank you – Discussion

- Verband DHd – „Digital Humanities im deutschsprachigen Raum“
  - http://www.dig-hum.de/

- AK „Informatik und Digital Humanities“ in der GI
  - Im FB Informatik und Gesellschaft (IUG)