Research Infrastructure for the Humanities

– or –

What the hell is “Digital Humanities”? 

Prof. Dr. Lukas Rosenthaler
Digital Humanities Lab
University of Basel
Humanities

• Most data is of *qualitative* nature
• Texts, descriptions, narrations
• facsimile of texts
• Images (still & moving), sketches, sound
• Annotations, comments, marginalia
Research tools in Humanities... as used to be...
Starting Point

- Many sources in Arts & humanities are available digitally or are being digitized now
- Portals allow for search & viewing of digital objects
- Search by keyword and/or facetting
Research tools in Humanities... today...
Use case (1) of DH: Critical Editions

• How to represent pages like this (or even much worse!) in the digital domain?

• How to constitute a text out of this mess?
  ➡️ we will need variants because scholars never will agree on a canonical text!

• Solution: Topographical transcription and standoff markup in RDF

The last page from Marcel Proust’s manuscript of *À la recherche du temps perdu.*
Topographical transcription

- regions of seemingly consistent text
- record topographical position
- encode text with attributes (underline, strike-through etc.)
Standoff markup

• old school: use TEI = XML = hierarchy (tree structure) = many problems:

This is quite difficult to encode

💀 !!overlapping tags!! 💀

• New school: standoff markup in RDF
  - Text as UTF-8 string
  - markup as separate data structure (RDF-Graph) with position information
Standoff markup

TEI = Text Encoding Initiative
XML based encoding for „texts“

👍 accepted standard in Humanities
👍 X-technologies (XPointer, XSLT etc.)
👎 XML = tree structure
☞ assumption that text structure is hierarchical
👎 Embedded markup
structure and markup are intermingled
👎 very complex „standard“
☞ encoding in TEI is an „art“ and often ambiguous
👎 very verbose

😈 TEI-„Mafia“ very strong… 😈
Standoff markup

• old school: use TEI = XML = hierarchy (tree structure) = many problems:

This is quite difficult to encode

💀!!overlapping tags!!💀

• New school: standoff markup in RDF
  - Text as UTF-8 string
  - markup as separate data structure (RDF-Graph) with position information
Use case (2) of DH: Historical information

First map of Basle (Matthäus Merian, 1615)
What’s the meaning of this plan?
How accurate?

...
Diary of Felix Platter (1536-1614), doctor during the plague, describing the houses he visited.
Use case (3) of DH: Corpus Vitriarium

- Inventory of stained glass in Europe
  - Metadata describing the objects
  - good quality images
    - LCAV eFacsimile project „relighting of stained glass windows“ would be interesting to integrate…😊
  - spatial relationship inside building
  - geolocation
- **Important:** What are the questions of
  - **Art history** (e.g. how was the church lighted at Easter 1126?)
  - **Religious studies** (St. Peter is facing St. Paul – always? why (not) ?)
  - …
3D sketch with spatial relations and annotations

eFacsimile relighting simulation
Humanities research workflow

• The scholar has to relate
  • Text- & documents (-parts)
  • Images(-regions), film(-sequences)
  • 3D-reconstructions
  • …
  to each other with comments, annotations etc.

• Contextualize the results ➔ publication (paper, monograph, …)

• new publication forms to come (paper, hybrid paper-digital, digital only)

Note: I consider a PDF not as a e-publication – it’s just a simulation of paper!

What’s in desperate need:
a generic research platform for the Humanities
Requirements for a *Digital Research Platform*

- **Flexible, versatile repository**
  - Encoded text (TEI, UTF-8,...), music (MEI,...), facsimile, images, sound, moving image,...
  - Versioning & change history ☛ citations

- **Methods and tools**
  - Annotation & Linkage (Nanopublications)
  - Analysis tools

- **Presentation & Visualization**
  - VRE
  - Visualization (2D, 3D, dynamic,...)

- **Collaboration**
  - Communication
  - Sharing

- **Longevity** (decades or more)
knora architecture
(Knowledge organization, representation and annotation)

- SALSAH Frontend
- LIMC Frontend
- R.Walser client
- Ontology Editor
- LOD Frontend

- jQuery GUILibrary / AngularJS (future)
- knora RESTful-API

- RESTful API (JVM/scala)
- knora RESTful-API

- SIPI (C++)
- MLLib (JAVA)
- sparql 1.1
- OAI / REST / sparql / ...

- Other Repositories (e-codices, e-lib, ...)
- Federal Archives
- OAIS Archive
- SIARD

- knora repository
  - private SPARQL endpoint (RDFS/OWL)
  - bitstream server from filesystem

- bitstream server from filesystem
- DISTARNET Archive
- public SPARQL endpoint
knora architecture
(Knowledge organization, representation and annotation)

- SALSAH Frontend
- LIMC Frontend
- R.Walser client
- Ontology Editor

- jQuery GUI Library / AngularJS (future)
- knora RESTful-API
- knora RESTful-API

- RESTful API (JVM/scala)

- SIPI (C++)

- Other Repositories
  - (Fedora etc.)

- Other Repositories
  - (e-codices, e-lib, ...)

- Federal Archives
  - SIARD

- OAIS Archive

- knora repository
  - private SPARQL endpoint (RDFS/OWL)
  - bitstream server from filesystem

- MLLib (JAVA)
- sparql 1.1

- LIMC
- LIMC

- MLlib (JAVA)

- R.Walser client

- public SPARQL endpoint

- DISTARNET Archive
Repository layer

• Problem:
  - Encoded text (TEI, UTF-8,…), music (MEI,…), facsimile, images, sound, moving image,…
  - soft, qualitative „values“ (e.g. a „date“ as „point“ in time)
  - Versioning & change history ⇾ citations / permalink
  - linking, annotations, …

• Solution:
  - RDF-triplestore
  - base ontology (RDFS/OWL)
    ⇾ versioning, complex, soft values
  - project specific ontologies (derived) for metadata and knowledge snippets (annotations, nano publications etc.)
  - Integrated media server for images, sound, film etc.
Semantic Web Technologies

- **Ontology**
  formal naming and definition of the concepts, types, properties, and interrelationships of the entities of the "world"

- **Resource Description Framework (RDF)**
  - W3C standard for data modeling
  - Statements about "Resources" with "triples":
    - `subject` – `predicate` – `object`
      - `<This book> has the name "Stultifera Navis"`
      - `<This book> has the author <Sebastian Brant>`
      - `<This Book> was published "1494"`
      - `<Sebastian Brant> has the name "Sebastian Brant"`
      - `<Sebastian Brant> is born "1457"`
      - `<Sebastian Brant> died "10 May 1521"`

- Directed graph ⊨ triple store = graph database

- **RDF Schema (RDFS)**
  Schema for simple ontologies (in RDF notation)

- **Web Ontology Language (OWL)**
  Language for semantic knowledge representation (RDF notation)
Repository layer

• Problem:
  - Encoded text (TEI, UTF-8,…), music (MEI,…), facsimile, images, sound, moving image,…
  - soft, qualitative „values“ (e.g. a „date“ as „point“ in time)
  - Versioning & change history ◄ citations / permalink
  - linking, annotations, …

• Solution:
  - RDF-triplestore
  - base ontology (RDFS/OWL) ◄ versioning, complex, soft values
  - project specific ontologies (derived) for metadata and knowledge snippets (annotations, nano publications etc.)
  - Integrated media server for images, sound, film etc.
Repository layer

- **Problem:**
  - Encoded text (TEI, UTF-8,...), music (MEI,...), facsimile, images, sound, moving image,...
  - soft, qualitative „values“ (e.g. a „date“ as „point“ in time)
  - Versioning & change history ➔ citations / permalink
  - linking, annotations, ...

- **Solution:**
  - RDF-triplestore
  - base ontology (RDFS/OWL) ➔ versioning, complex, soft values
  - project specific ontologies (derived) for metadata and knowledge snippets (annotations, nano publications etc.)
  - Integrated media server for images, sound, film etc.
Advantages of semantic web

- Universal knowledge representation (using RDF)
- Reusable data models (using RDF, RDFS and OWL) hierarchy of ontologies
- W3C standard query language (SPARQL 1.1)
- rule based inference engines and reasoning
Ontology

• <http://www.knora.org/ontology/knora-base> base-ontology from which project specific ontologies are derived

• Requirements:
  - high flexibility
  - support of standards (e.g. dublin core,...)
  - soft, complex values (e.g. date, time,...)
  - timestamp based versioning ➔ permalinks
  - language agnostic
Base-ontology

@prefix : <http://www.knora.org/ontology/knora-base#> 

Basic design pattern:
- Each „value“ is a resource
- A value may have several literals to describe the „value“
- Each value has a timestamp
- :previousValue points to previous versions ➡️ history
Base-ontology

@prefix : <http://www.knora.org/ontology/knora-base#>
Deriving ontologies

Resources

Properties for values

Properties for literals
Example: Ship of fools

inc:stultifera

inc:hasTitle

inc:hasPubdate

knora:DateValue

knora:hasResPtr

knora:startJDC

knora:endJDC

knora:hasDatePrecision

knora:calendar

"2266457"^^int

"MONTH"

"GREGORIAN"

"DAY"

inc:f3r

inc:inBook

inc:pagename

knora:TextValue

knora:valueHasString

"f 3 recto"@de

inc:transcription

knora:valueHasString

""""De custodia mulierum...""""@lt

knora:hasStandoff

knora:valueHasString

"bold", [15,16], "...

knora:hasGeom

"rect": [.77,.33,...]"
Example: Ship of fools
Example: Ship of fools

Incunable „Ship of fools“

Title „Stultifera Navis“
Example: Ship of fools

Incunable „Ship of fools”

Title „Stultifera Navis“

Publication date

Inc:f3r
inc:hasTitle
inc:hasPubdate
inc:hasResPtr
inc:inBook
inc:hasTranscription
inc:hasPageName

Example: Ship of fools

Title "Stultifera Navis"

Publication date

Page "f 3 recto"

Incunable "Ship of fools"
Example: Ship of fools

Title „Stultifera Navis“

Publication date

Page „f 3 recto“
Example: Ship of fools

Title: "Stultifera Navis"

Page identifier: "f 3 recto"

Publication date: "2266490"^int, "MONTH", "GREGORIAN", "DAY"

Page is in incunable

Incunable: "Ship of fools"
Example: Ship of fools

Title "Stultifera Navis"

Publication date "2266457"^^int

Page "f 3 recto"

Page identifier

Transcription with standoff markup and topographical position

Page is in incunable

Incunable „Ship of fools“
Additional features

• Access management
  - user and permission management in RDF
  - session management using cookies and MongoDB
  - Integration to switch/AAI planned

• Bitstream server IIIF-compliant (C++ implementation) [IIIF = International Image Interoperability Framework]
  - filesystem based
  - images: JPEG2000 & on-the-fly conversion
    ‣ multiple file formats (JPG, TIFF, PNG), resolutions and compression qualities, watermarking
  - preserves all embedded metadata (!)
  - enrich metadata with data from repository
knora architecture

(Knowledge organization, representation and annotation)

SALSAH Frontend

LIMC Frontend

R.Walser client

Ontology Editor

LOD Frontend

jQuery GIULibrary / AngularJS (future)

knora RESTful-API

knora RESTful-API

RESTful API (JVM/scala)

OAI / REST / sparql / …

SIPI (C++)

sparql 1.1

MLLib (JAVA)

knora repository

private SPARQL endpoint (RDFS/OWL)

bitstream server from filesystem

DISTARNET Archive

Federal Archives

Other Repositories (e-codices, e-lib, …)

bitstream server from filesystem

OAIS Archive SIARD
DISTARNET

DISTributed ARchival NETwork

• is a distributed, autonomous long-term digital preservation system based on a self-organizing P2P network

• At data level, DISTARNET supports complex data objects, the management of collections, annotations, and arbitrary links between digital objects

• At process level, dynamic replication management, consistency checking, and automated recovery of the archived digital objects

• autonomic behavior governed by preservation policies without any centralized component
DISTARNET key features

- Self-Configuration
- Self-Healing
- Self-Optimization
- based on RDF
- implementation scala/AKKA
knora architecture

(Knowledge organization, representation and annotation)

- SALSAH Frontend
- LIMC Frontend
- Knora RESTful-API
- jQuery GuILibrary / AngularJS (future)

Other Repositories
(e-codices, e-lib, ...)

OAI / REST / sparql / ...

ImgLib (C++)

sparql 1.1

MLLib (JAVA)

knora repository

private SPARQL endpoint (RDFS/OWL)

bitstream server from filesystem

DISTARNET Archive

Federal Archives

OAIS Archive

SIARD

Ontology Editor

R.Walser client

SALSAH

LIMC

R.Walser client

frontend

Dokubib

Frontend

PARZIVAL

Frontend

LIMC

Frontend

LOD Frontend

knora RESTful-API

sparql 1.1

public SPARQL endpoint

private SPARQL endpoint (RDFS/OWL)

bitstream server from filesystem

DISTARNET Archive
RESTful API layer

• Requirements:
  - modern, sustainable design
  - scalability, paralleism

implementation using scala, akka, spray
  (reactive design, futures)

• Provides simple high-end API to applications
• primary, recommended access method to repository
• expandable & flexible
knora architecture
(Knowledge organization, representation and annotation)
Frontend-toolbox

- designed for single-load web application design
- using jQuery + AJAX
- about 40 plugins which simplify the interaction with rapier
  ➤ fast and efficient design of customized, project specific user interfaces
- considering migration to Angular/JS or Angular/DART
SALSAH

- generic web-based front end
- HTML5 / AJAX / RESTful
- Window/desktop metaphor within browser
  - simultaneous view of multiple objects
  - comparison of objects
  - annotation and easy linking (drag/drop)
intuitive media viewer

- zoom: mousewheel
- pan: drag with mouse
- navigator window
  - overview
  - zoom control
  - pan control
Project specific GUI: Parzival

Courtesy of Prof. M. Stolz, University of Bern

original

knora/salsah-version
Graph view
Knora today…

• 1.1 mil. resources
• 12 mil. triples
• 0.5 mil. images (& a few films/video, ca. 500)
• attached repositories
  - e-codices (medieval manuscript library of Switzerland)
  - e-rara
  - e-manuscripta
• Data provider for
  - ORD@CH
Current users of knora

• in productive use
  - Anton Webern Edition
  - Swiss Society of Folklore
  - Robert Walser Edition
  - Viat Image
  - Documentation Library of St. Moritz
  - Kunsthal Basel
  - Text-image relations within the „Ship pf Fools“
  - Vitro Centre Romont
  - Bruno Manser Fond
  - Lexicon Iconographicum Mythologiæ Classicæ (since last Friday :-)
  - History of Russian Postcards
  - Parzival Edition

• in preparation
  - Bernoulli-Euler Online Edition
  - Alexander Humboldt Edition
  - HyperHamlet
  - New Testament
  - Commentary of Petrus Lombardus
  - Homer Commentary (Center for Hellenic Studies, Harvard University)
  - …
  - Montreux Jazz Archive?
  - Cinémathèque Suisse?
Hardware

Dalco 6-Core 64GB RAM
Virtual machine CentOS
Virtual machine CentOS

Dalco 6-Core 64GB RAM
Virtual machine CentOS
Virtual machine WinServer

QNAP 72TB RAID5+ NAS
NFS, CIFS,…

10 GB Ethernet
1 GB Ethernet

Virtual machine
virtual machine

Large Scale Storage
100TB+

IT-Services

Tape

10 GB Ethernet
1 GB Ethernet

Klingelbergstrasse 30
Bernoullistrasse 32
Goal: Keeping research data accessible

• for a indefinite time
• in a way that it is accessible/useable to others
  ☛ open access (if possible)
• after the funding of a project ceased
• after the people left

• development of new tools & digital methods
• Support of Digital Humanities Research in Switzerland

• based on knora
Planned developments

• knora as a P2P network of servers at participating organizations

• Integration of WebGL based 3D visualization
  - GIS for archeology
  - 3D-modelling of monuments, objects etc.

• Reflectance Transformation Imaging (RTI)
• Better tools for annotation of moving images
• Tools for critical editions (literature, Science, music,…)
• Tools for „real“ e-publications (with DeGruyter?)
• Integrated image analysis tools
Team-Basel

André Kilchenmann
Media Science
(PhD-student)

Dr. Benjamin Geer
Computer science &
oriental studies

Dr. Ivan Subotic
Computer Science &
DISTARNET

Andreas Aeschlimann
Programming &
Mathematics

Tobias Schweizer
Historien, Editions &
Software development

Lukas Rosenthaler
Linking Data to Knowledge
Linking Data to Knowledge

The Beginning
(and not the end!)