ZENTRUM FÜR
INFORMATIONSMODELLIERUNG
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES

KARL-FRANZENS-UNIVERSITÄT GRAZ
UNIVERSITY OF GRAZ

UNI
GRAZ

Handschriften digital repräsentieren (Standards, Archivierung & Linked Open Data)

Gastvortrag in LV 100200 KO ÄdL: Digitale Transformation in der Handschriftenforschung: (2022S)– Einführung in die Methoden der Computer Vision und der Materialanalyse

Sarah Lang

Universität Wien, 19.05.2022

Overview

1. Goals for this session
2. GAMS
3. Digital Scholarly Editing
4. Digital Archives
5. Annotating with XML markup
6. Text Encoding Initiative
7. Primer on metadata formats
8. Linked Open Data (LOD) in the Semantic Web
9. Archiving Image Data

Goals for this session

Goals

How to archive heritage imaging data in GAMS?

Primary goal: understand the M3R Content Model. ⬇

- What is GAMS? (*Geisteswissenschaftliches Asset Management System*)
- What's the difference between a digital edition & a digital archive?
- What data goes into the GAMS Multispectral Content Model & how does it all work?

Long-term archiving

1. the GAMS repository
2. Digital Editions
3. Digital Archives

XML/TEI

1. TEI for storing text
2. metadata in the *`<teiHeader>`*
3. TEI for digital editing

Linked Open Data

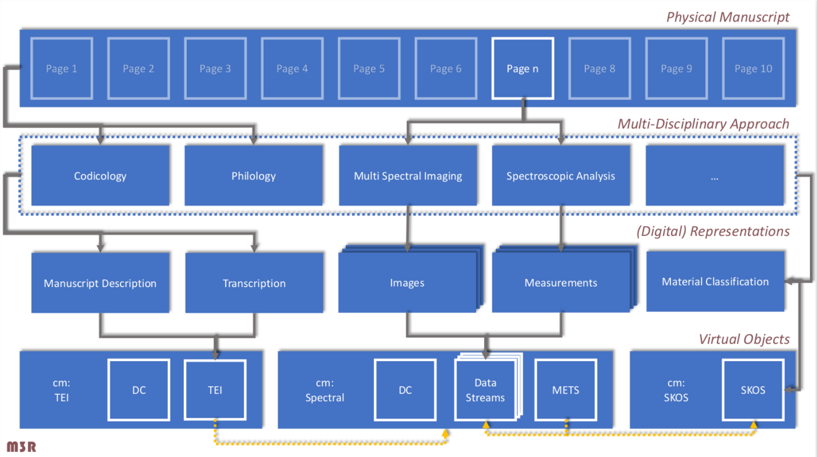
1. **Metadata standards:** DC, METS, IIIF
2. **Semantic Web Stack:** RDF(S), SKOS

How to learn all this?

The slides contain a few suggestions for practice (slides with blue background).

There are Digital Humanities classes, master's degrees & certificates in Graz & Vienna.

Understanding the M3R Content Model (GAMS `cm:Spectral`)



Digitale Transformation in der Handschriftenforschung

Kontext

100200 KO ÄdL: Digitale Transformation in der Handschriftenforschung: Einführung in die Methoden der Computer Vision und der Materialanalyse (2022S)

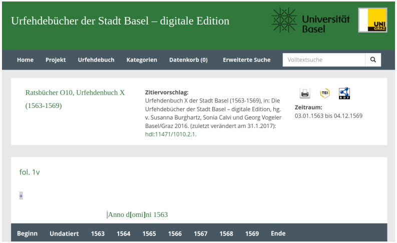
- Robert Sablatnig, Simon Brenner: Grundlagen der digitalen Bildgebung und -Verarbeitung
- Simon Brenner: Multi- und Hyperspektralfotographie
- Robert Sablatnig, Simon Brenner: 3D Rekonstruktion
- Florian Kleber: Dokumentanalyse
- Stiftsbibliothek Klosterneuburg: Restauratorische Sicht auf Handschriften, Tilgungen und Schäden. Demonstration am Original, Hands-On

- Manfred Schreiner, Wilfried Vetter: Grundlagen der zerstörungsfreien Materialbestimmung 1
- Federica Cappa: Grundlagen der zerstörungsfreien Materialbestimmung 2
- Guadalupe Pinar: Biokodikologie
- Sarah Lang: Archivierung und Linked Open Data
- INTK (AkBild), Materialanalyse: Gerätedemonstration / Hands-On

GAMS

Archiving and publishing heritage data... i

GAMS ufbas



Urfehdebücher der Stadt Basel – digitale Edition

Ratsbücher O10, Urfehdenbuch X (1563-1569)

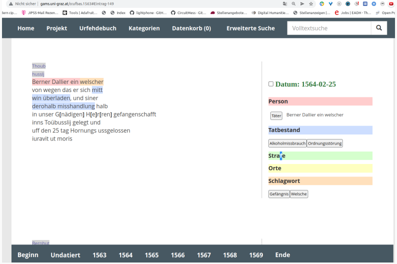
Zitiervorschlag:
Urfehdenbuch X der Stadt Basel (1563-1569), in: Die Urfehdebücher der Stadt Basel – digitale Edition, hg. v. Susanna Burghartz, Sonia Calvi und Georg Vogeler Basel/Graz 2016. (zuletzt verändert am 31.1.2017): hdl:11471/1010.2.1.

Zeitraum:
03.01.1563 bis 04.12.1569

fol. 1v

+

|Anno d[omi]ni 1563



Thoub
husslj
Berner Dallier ein welscher
von wegen das er sich mitt
win überladen, und siner
derohalb misshandlung halb
in unser G[nädigen] H[e]r[ren] gefangenschafft
inns Toübusslij gelegt und
uff den 25 tag Hornungs ussgelossen
iuravit ut moris

Bernhut

☐ **Datum: 1564-02-25**

Person

Täter Berner Dallier ein welscher

Tatbestand

Alkoholmissbrauch Ordnungsstörung

Strafe

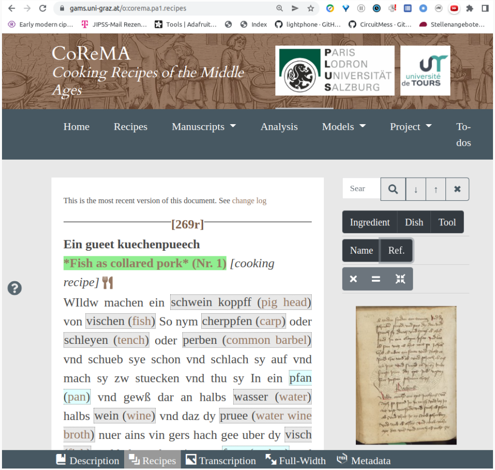
Orte

Schlagwort

Gefängnis Welsche

Archiving and publishing heritage data... ii

CoReMA → Look at: `http://gams.uni-graz.at/corema`



This is the most recent version of this document. See change log

[269r]

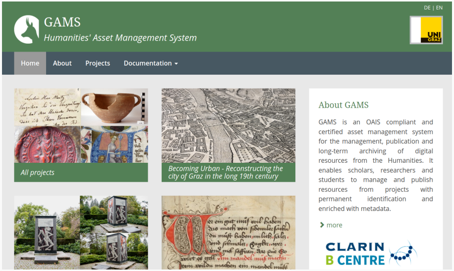
Ein gueet kuechenpueech

***Fish as collared pork* (Nr. 1)** *[cooking recipe]* 🍴

WIldw machen ein schwein koppff (pig head) von vischen (fish) So nym cherppfen (carp) oder schleyen (tench) oder perben (common barbel) vnd schueb sye schon vnd schlach sy auf vnd mach sy zw stuecken vnd thu sy In ein pfan (pan) vnd gewß dar an halbs wasser (water) halbs wein (wine) vnd daz dy pruee (water wine broth) nuer ains vin gers hach gee uber dy visch

GAMS (Humanities Asset Management System) i

- GAMS (*Geisteswissenschaftliches Asset Management System* = AMS for the Humanities)
- based on **FEDORA** (*Flexible Extensible Digital Object Repository Architecture*) = infrastructure dedicated to the persistent archival and management of resources considered to be worthy of long-term preservation.


DE | EN
GAMS
Humanities' Asset Management System
UNI GRAZ
Home
About
Projects
Documentation
All projects
Becoming Urban - Reconstructing the city of Graz in the long 19th century
more
CLARIN B CENTRE

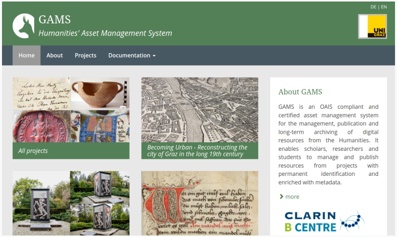
About GAMS

GAMS is an OAIS compliant and certified asset management system for the management, publication and long-term archiving of digital resources from the Humanities. It enables scholars, researchers and students to manage and publish resources from projects with permanent identification and enriched with metadata.

https://gams.uni-graz.at

GAMS (Humanities Asset Management System) ii

- user access provided through the **Cirilo client**
- **functionalities:** object creation and management, versioning, normalization & standards, choice of data formats.
- offers a plethora of **pre-defined content models** for data such as TEI, MEI, LIDO, SKOS, ontologies, R code and story lines
- → offers publication pipelines but is highly customizable
- more info: *`http://gams.uni-graz.at/doku`*



GAMS

Humanities' Asset Management System

All projects

Becoming Urban - Reconstructing the city of Graz in the long 19th century

About GAMS

GAMS is an OAIS compliant and certified asset management system for the management, publication and long-term archiving of digital resources from the Humanities. It enables scholars, researchers and students to manage and publish resources from projects with permanent identification and enriched with metadata.

› more

Long-term archiving heritage data i

Long-term preservation

denotes the process of maintaining, curating and keeping data usable over a long period of time (10+ years).

Key functionalities of long-term archiving architectures include

- persistent identification,
- versioning,
- support of different data formats,
- management of associated metadata,
- data export and retrieval,
- security and scalability.

Special emphasis is placed on

- sustainability,
- citability &
- guarantee of long-term access to the contained resources.

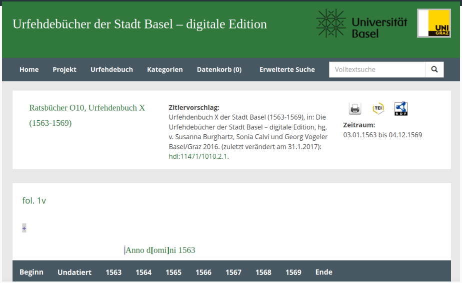
Long-term archiving heritage data ii

XML is great for long-term archiving!

Consequently, data formats and software used for preservation should follow **open source & non-proprietary standards**; data is ideally encoded in an **unicode XML format**:

- plain text files are small
- human- & machine-readable
- recognized standard stable since 1998
- more on XML later...

Examples of data quality in GAMS i



Ratsbücher O10, Urfehdenbuch X (1563-1569)

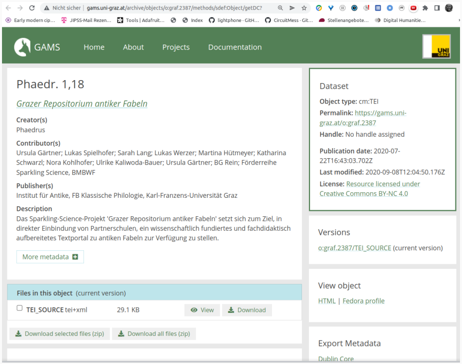
Zitiervorschlag: Urfehdenbuch X der Stadt Basel (1563-1569), in: Die Urfehdebücher der Stadt Basel – digitale Edition, hg. v. Susanna Burghartz, Sonia Calvi und Georg Vogeler Basel/Graz 2016. (zuletzt verändert am 31.1.2017): hdl:11471/1010.2.1.

Zeitraum: 03.01.1563 bis 04.12.1569

fol. 1v

+

|Anno d[omi]ni 1563



Phaedr. 1,18

Grazer Repositorium antiker Fabeln

Creator(s)
Phaedrus

Contributor(s)
Ursula Gärtner; Lukas Spielhofer; Sarah Lang; Lukas Werzer; Martina Hütmeyer; Katharina Schwarzl; Nora Kohlhofer; Ulrike Kaliwoda-Bauer; Ursula Gärtner; BG Rein; Förderreihe Sparkling Science, BMBWF

Publisher(s)
Institut für Antike, FB Klassische Philologie, Karl-Franzens-Universität Graz

Description
Das Sparkling-Science-Projekt 'Grazer Repositorium antiker Fabeln' setzt sich zum Ziel, in direkter Einbindung von Partnerschulen, ein wissenschaftlich fundiertes und fachdidaktisch aufbereitetes Textportal zu antiken Fabeln zur Verfügung zu stellen.

Files in this object (current version)

□ TEI_SOURCE tei+xml 29.1 KB

Dataset

Object type: cm:TEI
Permalink: https://gams.uni-graz.at/o:graf.2387
Handle: No handle assigned

Publication date: 2020-07-22T16:43:03.702Z
Last modified: 2020-09-08T12:04:50.176Z
License: Resource licensed under Creative Commons BY-NC 4.0

Versions

o:graf.2387/TEI_SOURCE (current version)

View object

HTML | Fedora profile

Export Metadata

Dublin Core

← *http://gams.uni-graz.at/o:ufbas.1563*

- **top image:** info on recommended citation, downloadable source data in XML & RDF
- **below:** 'data view' of the Dublin Core (DC) metadata for an XML object in GAMS (ensuring citability, etc.)

← *http://gams.uni-graz.at/archive/objects/o:graf.2387/methods/sdef:Object/getDC?*

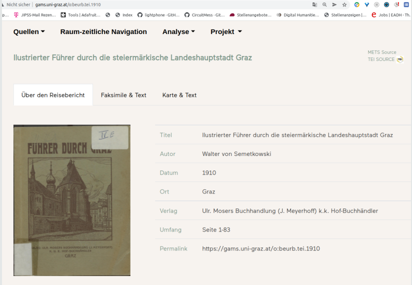
Examples of data quality in GAMS ii

- GAMS isn't an out-of-the-box tool – it's something between a **Content Management System (CMS)**, a **repository** (long-term archiving) and a **publication platform**.
- The XSLT transformations applied to files in GAMS are custom but there are **wippets** (standard javascript functionalities → *widget + snippet*, example here) and **standard templates** which can be reused.
- as a publication platform: presentation and long-term archiving, allowing for persistent access to the data, versioning, etc.

Examples of data quality in GAMS iii

One more example: *`https://gams.uni-graz.at/beurb`*

Metadata view on XML data

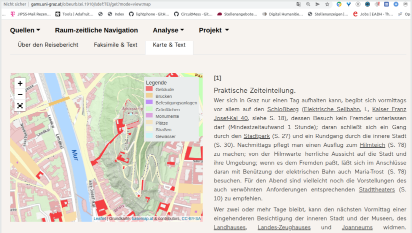


Nicht sicher | gams.uni-graz.at/o:beurb.tei.1910
JIPSS-Mail Rezen... Tools | Adafruit... Index lightphone · GitH... CircuitMess · Git... Stellenangebote... Digital Humanitie... Stellenanzeigen |... Jobs | EADH - Th...
Quellen
Raum-zeitliche Navigation
Analyse
Projekt
METS Source
TEI SOURCE
Über den Reisebericht
Faksimile & Text
Karte & Text
FÜHRER DURCH GRAZ
IV E

Ilustrierter Führer durch die steiermärkische Landeshauptstadt Graz

Titel	Ilustrierter Führer durch die steiermärkische Landeshauptstadt Graz
Autor	Walter von Semetkowski
Datum	1910
Ort	Graz
Verlag	Ulr. Mosers Buchhandlung (J. Meyerhoff) k.k. Hof-Buchhändler
Umfang	Seite 1-83
Permalink	https://gams.uni-graz.at/o:beurb.tei.1910

Map view linked to edited text



Nicht sicher | gams.uni-graz.at/o:beurb.tei.1910/sdef:TEI/get?mode=view:map
JIPSS-Mail Rezen... Tools | Adafruit... Index lightphone · GitH... CircuitMess · Git... Stellenangebote... Digital Humanitie... Stellenanzeigen |... Jobs | EADH - Th... Op
Quellen
Raum-zeitliche Navigation
Analyse
Projekt
Über den Reisebericht
Faksimile & Text
Karte & Text
+
−
Legende
🟥 Gebäude
🟧 Brücken
🟪 Befestigungsanlagen
🟩 Grünflächen
🟪 Monumente
🟧 Plätze
🟨 Straßen
🟦 Gewässer
Mur
Lendkai
Neubaugasse
Schlossberg
anz-Josef-Kai
Leaflet | Grundkarte: basemap.at & contributors, CC-BY-SA

[1]

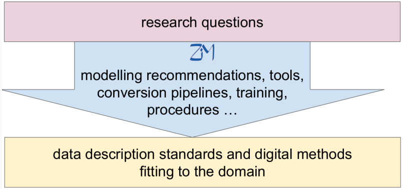
Praktische Zeiteinteilung.

Wer sich in Graz nur einen Tag aufhalten kann, begibt sich vormittags vor allem auf den Schloßberg (Elektrische Seilbahn, I., Kaiser Franz Josef-Kai 40, siehe S. 18), dessen Besuch kein Fremder unterlassen darf (Mindestzeitaufwand 1 Stunde); daran schließt sich ein Gang durch den Stadtpark (S. 27) und ein Rundgang durch die innere Stadt (S. 30). Nachmittags pflegt man einen Ausflug zum Hilmteich (S. 78) zu machen; von der Hilmwarte herrliche Aussicht auf die Stadt und ihre Umgebung; wenn es dem Fremden paßt, läßt sich im Anschlüsse daran mit Benützung der elektrischen Bahn auch Maria-Trost (S. 78) besuchen. Für den Abend sind vielleicht noch die Vorstellungen des auch verwöhnten Anforderungen entsprechenden Stadttheaters (S. 10) zu empfehlen.

Wer zwei oder mehr Tage bleibt, kann den nächsten Vormittag einer eingehenderen Besichtigung der inneren Stadt und der Museen, des Landhauses, Landes-Zeughauses und Joanneums widmen.

Both views are created ‘on-the-fly’ from the same XML file (=*single source principle*, more on that later...).

GAMS i


research questions

modelling recommendations, tools, conversion pipelines, training, procedures …

data description standards and digital methods fitting to the domain

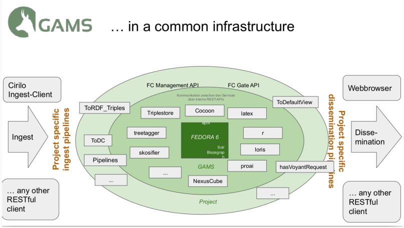
Content Models:
generic object(s)
(data and methods)

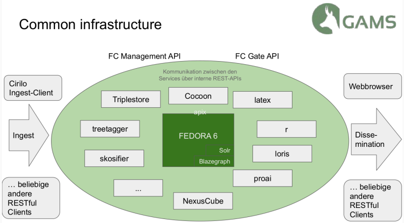
https://gams.uni-graz.at/doku

Content Models
project specific object(s)
(data and methods)

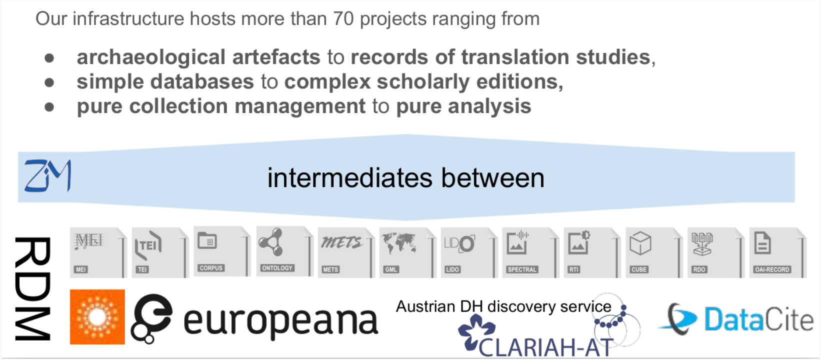
http://gams.uni-graz.at/context:cantus/sdef:Object/getDC

http://gams.uni-graz.at/context:cantus


GAMS … in a common infrastructure


Common infrastructure

GAMS ii



Our infrastructure hosts more than 70 projects ranging from

- **archaeological artefacts** to **records of translation studies**,
- **simple databases** to **complex scholarly editions**,
- **pure collection management** to **pure analysis**

intermediates between

If we want to understand GAMS, we need to understand some terms first...

→ Digital Edition versus Digital Archive/Repository

Digital Scholarly Editing

What is a Digital Scholarly Edition? i

Digital editions (as per Sahle 2016):

- discipline-independent & not exclusive to text
- overcome the limitations of print (financial, inter-/ multimediality, etc.)
- follow a digital paradigm like printed editions follow the print paradigm (“shaped by the technical limitations and cultural practices”)
- “A digitised edition is not a digital edition.” → it’s not about the storage medium (just being on the internet doesn’t make it a digital edition!)
- “A digital edition cannot be given in print without significant loss of content and functionality.”
  - different views, interactivity, searchability
  - facsimiles alongside diplomatic transcription and reading versions
  - all generated from the same source data: **single source principle** → How? Data transformation!
  - ≠ Digital Archive, text corpus or facsimile → has critical engagement

What is a Digital Scholarly Edition? ii

We need to start at the beginning:

What is a scholarly edition?

- originally stems from ecdotics (*ecdotica, ecdotique, Editorik / Editionswissenschaft*) → tradition of Philology
- **goal = textual criticism**: reconstruct the original version of a text transmitted to us via textual witnesses (*Lachmannian paradigm*)

Sahle 2016

> “ A scholarly edition is the critical representation of historic documents.
> *Edition ist die erschließende Wiedergabe historischer Dokumente.* ”

What is a Digital Scholarly Edition? iii

Criteria from Sahle 2016:

1. **Representation:** recoding a document and its transformation (in the same or other media)
   - Visually: image reproduction, facsimile
   - Textually: transcription.
   - varying degrees of closeness/abstraction with regard to the original document
2. representation $\neq$ presentation
3. **Critical engagement** (based on scholarly agenda). “Critical engagement without representation is not an edition–but an examination, a catalogue or a description.”
   - textual criticism
   - historic criticism
   - bibliographic criticism
   - material criticism
   - visual criticism
   - etc.

What is a Digital Scholarly Edition? iv

4. **Documents:** “every non-abstract object that is the subject of an edition can be called a document.”
5. **Historic:** editions “explain what is not evident to the present-day reader. In short, they bridge a distance in time, a historical difference. Texts that are created today do not need to be critically edited. They can speak for themselves.”

Sahle 2016

> A representation without [critical] treatment or the addition of information is not an edition–but a facsimile, a reproduction or–nowadays–a digital archive or library.
>
> **Critical representation** as a compound notion of editing aims at the reconstruction and reproduction of texts and as such addresses their material and visual dimension as well as their abstract and intentional dimension (Sahle 2016).

What is a Digital Scholarly Edition? v

Questions to ask if you want to know if it's a Digital Scholarly Edition:

1. **Is there a full representation** of the subject in question?
2. **Is it critical?** → processing rules stated and applied, scholarly knowledge included to make the document easier to understand, regarding material, document genesis/creation, context and reception?
3. **Is the edition of academic quality?** → transparent and rigorous edition process, responsibilities stated, enables future research on a reliable basis.
4. **Does the edition follow a digital paradigm?** → makes use of the possibilities of the digital, not printable without major loss of content or functionality.

What is a Digital Scholarly Edition? vi

Ideally, a DSE should also implement the FAIR criteria

Findable. in library catalogs, discovery systems or repositories, with a persistent identifier

Accessible. free for any user, no access restrictions? (from open access to usability, language selection, etc.)

Interoperable. data in standardized and widely used format (for example TEI or similar standards), allowing for reuse and data exchange outside of the project.

Reusable. data accessible (individual download, aggregate download, repository, API)? Licenses allowing reuse. Data creation, modelling and processing documented adequately so that others can make sense of it?

→ RIDE (review journal for digital editions and resources)

→ RIDE Criteria for Reviewing Digital Editions and Resources

Edition workflows and schools i

The following slides are adapted from Georg Vogeler's slides on digital editing.

Traditional Workflow of Philological Editing

- **Heuristics:** Find your textual witnesses
- **Transcription:** transfer the text into your prefered alphabet (from the original/from a photo)
- **Collation:** Compare the textual witnesses
- **Recensio:** Evaluate the variants and create a stemma (commenting)
- Write your introduction
- Typesetting
- Create an index (refering to pages)
- Print and distribution

In the digital realm:

- edition
- beta version
- just TEI data publication
- hybrid (web and print) publication
- minimal edition (adaptation from minimal computing)

A digital edition is social, iterative, a process...

Edition workflows and schools ii

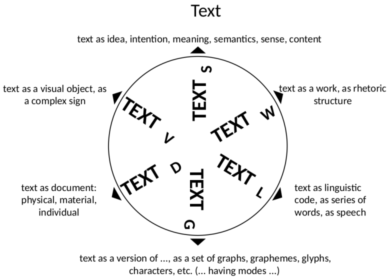
„Edition“

- a particular version of a book
- a particular version of a product
- all the copies of a book that are printed or published at one time

„Historical-critical“ Edition

- Documentation of the history of the text transmission in the “apparatus of variants”
- Critical evaluation of the textual transmission

Patrick Sahle’s text wheel


Text
text as idea, intention, meaning, semantics, sense, content
text as a work, as rhetoric structure
text as linguistic code, as series of words, as speech
text as a version of ..., as a set of graphs, graphemes, glyphs, characters, etc. (... having modes ...)
text as document: physical, material, individual
text as a visual object, as a complex sign
TEXT S
TEXT W
TEXT L
TEXT G
TEXT D
TEXT V

Karl Lachmann (1793–1851)

Editorial intervention: Philological knowledge (linguistics, style) allows to emend fragmentary textual transmission.

Edition workflows and schools iii

Editorial Schools

- Lachmann (‘Historisch-Kritische Ausgabe’, stemmatology)
- Reading text vs. critical edition = reader-oriented pragmatism / modernised edition / ‘reading text’
- Last authorized edition
- First edition / *editio princeps*
- Main manuscript (Bédier 1937, ‘Leithandschrift’)
- ‘diplomatic edition / transcription’
- Typographic/photographic facsimile
- ‘documentary editing’ (Tanselle)
- variorum edition
- genetic edition / *critique génétique* (Gabler)
- Copy-Text-Theory (Greg 1950/3, Bowers 1976, 1978)
- New Philology (Stackmann 1964, 1993, 1994, 1999; Ruh 1978, Cerquiglini 1989)

If you wanted to practice...

Write a RIDE review on a digital edition (suggestions for review) according to their review criteria!

(If you actually plan on submitting, contact the editors first. 2000 words minimum, project from your field. Peer-review can spot things you're unsure about.)

Digital Archives

Digital long-term archiving

- ensuring the authentic and sustainable availability of digital ressources on the level of the bitstream and on a semantic level
- integral principle to every form of sustainable data storage
- begins with the data production in a sustainable data format (and ideally, following a recognized data standard)
- requires standardization of data formats and archiving workflows
- serves both the dissemination as well as the preservation of digital content
- not just about technical solutions but also institutional stability and policies

A digital archive

- is more than a mere collection of scanned book pages or digitized images, etc.
- if offers metadata, norm data and controlled vocabularies

Digital Archives

Digital Archive

- organized collection of digital objects (text, images, audio, video and multimedia streams)
- digital objects are described by standards both in terms of contents (e.g. TEI) and bibliographically (e.g. Dublin Core)
- published sustainably using interfaces, services and APIs (e.g. OAI-PMH)
- digital objects have unique, persistent and citable identifiers (e.g. DOI, URN, PURL, PID)
- authenticity of objects is checked by means of digital signatures or checksums (‘Is the number of bytes in the object still the same it used to be?’)

Trustworthy Digital Archives

as defined by the Research Libraries Group (RLG)

- secure organizational structure and legal status
- financial sustainability
- technological and procedural aptness
- ensuring data and system security
- documentation and transparency
- conformity with the OAIS standard

→ retro-digitizing objects but also standards for new (born digital) resources.

Open Archival Information System Reference Mode

OAIS

Generic model for the organization of a digital archive → developed 1995–2002 by the Consultative Committee for Space Data Systems (CCSDS)

Tasks for Digital Archives according to OAIS

1. data ingest
2. archival storage
3. data management
4. system administration
5. preservation planning
6. access

What's the difference between a Digital Archive and a Digital Scholarly Edition? (Sahle 2007)

➜

Digital Archives vs. Digital Editions

Digital documents

Material objects are the target of digitization but digitization doesn't reproduce them – it represents them in a digital format. A digital document is a view on the original material object.

Archive

Traditionally an archive is an ordered collection of documents with the goal of documenting them, preserving them in the long-term and making them accessible. → this function isn't only carried out by actual archives but also by museums, libraries and other cultural heritage institutions. Traditional archive material doesn't need to be represented because its physical objects can be accessed directly.

(we have already def'd editions)

Editions are often based in archival material. The edition isn't a storage device, it is a publication of a historical source and the editorial work done on it. While the edition contains lots of editorial work and enrichment, archival documents are usually original and largely unprocessed. A non-digital archive isn't a form of publication in and of itself – a digital archive *is* a form of publication, too, like a digital edition → blurs the lines a bit.
We could say that **the difference lies in the depth of data enrichment and editorial work.** → once a presentation form is provided, data from a digital archive becomes a digital edition.

Digital Archive

A main goal of a digital archive is to preserve and publish a specific choice of documents as a collection (ideally, representative and well-balanced). The digital objects aren't necessarily direct representations but may have undergone editorial intervention (e.g. normalized orthography). Its documents should be uniform in how they are encoded and processed. In the case of the single source principle, the edition is generated dynamically ('on the fly') from the archived source data. (If any edition is provided).

GAMS uses many data standards...

→ XML/TEI is very useful for encoding text-based data.

Annotating with XML markup

TEI, now what?

The **Text Encoding Initiative** (TEI) for XML has become the gold standard for scholarly editions of texts.

Goals for the next part

1. wait, what was…
    - ✖ XML?
    - ✖ TEI?
    - ✖ How do I use the TEI for digital editing?

XML: eXtensible Markup Language

- W3Schools Tutorial
- paradigm of the separation of form and content
- XML is a metalanguage

.XML

- RSS, SOAP, XAML
- MathML, GraphML
- XHTML
- RDF
- KML
- Scalable Vector Graphics (SVG)

> “ **Extensible Markup Language (XML)** is a **markup language** and file format for storing, transmitting, and reconstructing arbitrary data. It defines a **set of rules for encoding documents** in a format that is **both human-readable and machine-readable.** (Wikipedia) ”

XML rules

XML can be checked for **validity** (validation if it complies with a standard) and **well-formedness** (following the rules of XML) → will only be parsed if well-formed. Thus: Heed thy error messages!

There are rules on how elements can be named (you can look them up if relevant or will get informed by an error message).

<key>value</key> . XML as a key value notation

Rules

- Hierarchical nesting below the root
- exactly one root element, i.e. one out-most russion doll
- start and end tag
- tag names are case-sensitive (!)
- empty elements allowed (& can be shortened)

Minimal example

```xml
<?xml version="1.0" ?>
<root>
  <element attribute="value">
    content
  </element>
  <!-- comment -->
</root>
```

XML rules i

Prolog

`<xml version="1.0" encoding="utf-8">`	XML declaration
`<?xsl-stylesheet type="text/xsl" href="my.xsl"?>`	processing instructions (optional)

you can include document models (optional)
DTD, XML Schema, RelaxNG, Schematron

entities ‘protected’ characters that have a meta meaning in XML like:

`&lt;`	<
`&gt;`	>
`&amp;`	&

XML family and vocabularies

XML structured description of data
XPath navigating xml documents
XML Schema strict data model
XSL Extensible Stylesheet Language
XSLT XSL-Transformations, i.e. transforming XML documents
XSL-FO formatted output (e.g. print)
XQuery query language for XML databases
and more

- **(X)HTML** Hypertext Markup Language
- **EAD** Encoded Archival Description
- **TEI** Text Encoding Initiative
- **CEI** Charters Encoding Initiative
- **MEI** Music Encoding Initiative
- **LIDO** Lightweight Information Describing Objects (describing museum or collection objects)
- **SVG** Scalable Vector Graphics
- **KML** Keyhole Markup Language (geography)
- **MathML**
- **CML** Chemical Markup Language, …

Text Encoding Initiative

TEI Primer

Text Encoding Initiative

.XML

XML-Standard, i.e. convention on how to use XML so that resulting data will be interoperable between different projects. (founded in 1987, consortium since 2000)

> The Text Encoding Initiative (TEI) is a text-centric community of practice in the academic field of digital humanities, operating continuously since the 1980s. The community currently runs a mailing list, meetings and conference series, and maintains the TEI technical standard, a journal, a wiki, a GitHub repository and a toolchain. (Wikipedia)

TEI minimal example

```
<TEI> <!-- root element -->
    <teiHeader>
      <!-- author, title, dating,
           sources, edition rules, etc.
    </teiHeader>
    <text> ... </text>
</TEI>
```

Resources

- Learn TEI
- Teach Yourself
- P5 = 5. Proposal
- MEI for music
- CEI for charters
- http://www.tei-c.org/

TEI Header

fileDesc = bibliographical description of the contents of the document

encodingDesc = connection of electronic document to source (i.e. transcription rules, etc.)

```xml
<TEI> <!-- root element -->
    <teiHeader>
        <fileDesc> ... </fileDesc> <!-- obligatory -->
        <encodingDesc> <!-- optional -->
        <profileDesc> <!-- optional -->
        <revisionDesc> <!-- optional -->
    </teiHeader>
    <text> ... </text>
</TEI>
```

profileDesc = decribes all non-bibliogaphical aspects of the text (i.e. creation, languages)

revisionDesc = tracks changes in the digital document

Metadata in the TEI Header i

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>
<!-- title of the resource -->
   </title>
  </titleStmt>
  <publicationStmt>
   <p>
<!-- Information about distribution of the resource -->
   </p>
  </publicationStmt>
  <sourceDesc>
   <p>
<!-- Information about source from which the resource derives -->
   </p>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

Metadata in the TEI Header ii

The title and author in the *<titleStmt>* isn't the bibliographic data from the source! It describes the digital document and its authors or editors.

If you want to desribe your source documents, you need elements like *<sourceDesc>* or *<msDesc>*:

```
<sourceDesc>
 <bibl>
  <title level="a">The Interesting story of the Children in the Wood</title>. I
 <author>Victor E Neuberg</author>, <title>The Penny Histories</title>.
 <publisher>OUP</publisher>
  <date>1968</date>. </bibl>
</sourceDesc>
```

```
<sourceDesc>
 <p>Born digital: no previous source exists.</p>
</sourceDesc>
```

Metadata in the TEI Header iii

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Thomas Paine: Common sense, a
       machine-readable transcript</title>
   <respStmt>
    <resp>compiled by</resp>
    <name>Jon K Adams</name>
   </respStmt>
  </titleStmt>
  <publicationStmt>
   <distributor>Oxford Text Archive</distributor>
  </publicationStmt>
  <sourceDesc>
   <bibl>The complete writings of Thomas Paine, collected and edited
       by Phillip S. Foner (New York, Citadel Press, 1945)</bibl>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

<msDesc>

```
<msDesc>
 <msIdentifier>
  <settlement>Oxford</settlement>
  <repository>Bodleian Library</repository>
  <idno type="Bod">MS Poet. Rawl. D. 169.</idno>
 </msIdentifier>
 <msContents>
  <msItem>
   <author>Geoffrey Chaucer</author>
   <title>The Canterbury Tales</title>
  </msItem>
 </msContents>
 <physDesc>
  <objectDesc>
   <p>A parchment codex of 136 folios, measuring approx
       28 by 19 inches, and containing 24 quires.</p>
   <p>The pages are margined and ruled throughout.</p>
   <p>Four hands have been identified in the manuscript: the first 44
       folios being written in two cursive anglicana scripts, while the
       remainder is for the most part in a mixed secretary hand.</p>
  </objectDesc>
 </physDesc>
</msDesc>
```

<titlePage>

To describe a title page (e.g. early modern print copperplates, etc.), use *<titlePage>*:

```
<titlePage>
 <docTitle>
  <titlePart type="main">THOMAS OF Reading.</titlePart>
  <titlePart type="alt">OR, The sixe worthy yeomen of the West.</titlePart>
 </docTitle>
 <docEdition>Now the fourth time corrected and enlarged</docEdition>
 <byline>By T.D.</byline>
 <figure>
  <head>TP</head>
  <p>Thou shalt labor till thou returne to duste</p>
  <figDesc>Printers Ornament used by TP</figDesc>
 </figure>
 <docImprint>Printed at <name type="place">London</name> for <name>T.P.</name>
  <date>1612.</date>
 </docImprint>
</titlePage>
```

<front>

You might also need *<front>* (front matter): contains any prefatory matter (headers, abstracts, title page, prefaces, dedications, etc.) found at the start of a document, before the main body.

```
<front>
 <epigraph>
  <quote>Nam Sibyllam quidem Cumis ego ipse oculis meis vidi in ampulla
     pendere, et cum illi pueri dicerent: <q xml:lang="grc">Σίβυλλα τί
       θέλεις</q>; respondebat illa: <q xml:lang="grc">ὰποθανεῖν θέλω.</q>
  </quote>
 </epigraph>
 <div type="dedication">
  <p>For Ezra Pound <q xml:lang="it">il miglior fabbro.</q>
  </p>
 </div>
</front>
```

Making the TEI your own i

How to find information on TEI elements

...and teach yourself how to use new elements:

- General TEI guidelines (XML Primer, Learn the TEI page, etc.)
- web-search TEI + (element you want to know about), i.e. “tei teiHeader” and you will get:
  1. definition page
  2. list of all examples for that element → directly over websearch or click ‘show all’ in the examples on the ‘definitons page’
  3. sometimes even an module overview text for things as big as *`<teiHeader>`* (has its own module)

Relevant TEI modules

all	All modules
5	Characters, Glyphs, and Writing Modes,
10	Manuscript Description,
11	Representation of Primary Sources,
12	Critical Apparatus,
13	Names, Dates, People, and Places.

Also: The TEI guidelines are documentation and reference, not necessarily ideal teaching tools → overwhelming. Maybe try other tutorials like the TEI by example page, for example the tutorials on Primary Sources and Critical Editing.

Making the TEI your own ii

Oxygen tricks

- If the TEI schema is linked to your document and you have internet, you can hover over elements and click to be redirected to the relevant info page.
- If you open a tag (by just typing '`<`'), the editor will suggest a list of elements currently allowed where you're standing (for example, *`<teiHeader>`* is very picky about the sequence).

TEI practice!

Fill out the `teiHeader` or `msDesc`.

Use websearch (‘tei msDesc’) to learn how to use new elements (overview plus examples view).

TEI for Digital Editing i

TEI can describe the structure of a text, e.g.

- speaker, verse line, stage directives
- greeting, signature
- Visual aspects of the script
- special characters, new lines

Simple layout markup

- beginning of a new **line**: *<lb/>*
- beginning of a new **page**: *<pb/>* *@n* for an explicit numbering
- beginning of a new **column**: *<cb/>*
- **highlighted** text: *<hi>*
  - Attribute *@rend* to describe the appearance
  - Alternative encoding: , ,
- graphical elements in the text: *<figure>*
- *<fw>* (forme work) contains a running head (e.g. a header, footer), **catchword**, or similar material appearing on the current page.

```
<fw place="top-centre" type="head">Poëms.</fw>
<fw place="top-right" type="page-no">29</fw>
```

TEI for Digital Editing ii

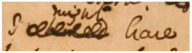
Documenting particularities of the writing surface

<damage> *@agent, @degree, @unit, @quantity, @extent, @precision, @scope*

<unclear>

<gap> any ommission in the transcription – @reason, e.g. sampling, inaudible, irrelevant, cancelled

e.g. unclear passage



Other important attributes

@cert(ainty) how certain you are about the suggested transcription?

@resp(onsibility) who did it?

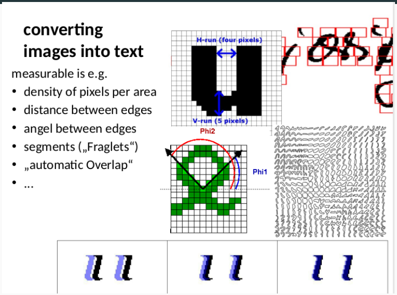
@evidence where you got the clues from (internal, external, conjecture)?

```
I <subst>
 <add place="above">might</add>
  <del>
   <unclear reason="overinking"
       cert="medium" resp="#LDB">
       should</unclear>
  </del> </subst> have
```

```
<gap reason=“wormhole" quantity="5" unit="character"/>
<damage agent=“coffee” quantity=“3” unit=“line”/>
```

Transcription i

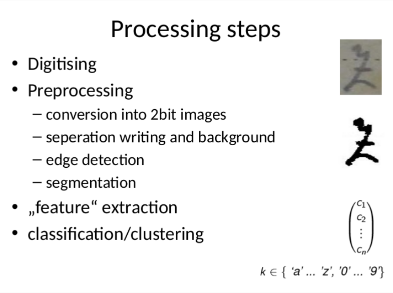
- OCR (Optical Character Recognition) – e.g. Transkribus (transcription support)
- also: Transkribus Keyword Spotting
- also: fuzzy search which should also find the word if it's mistranscribed
- Writer identification



converting images into text

measurable is e.g.

- density of pixels per area
- distance between edges
- angel between edges
- segments („Fraglets")
- „automatic Overlap"
- …



Processing steps

- Digitising
- Preprocessing
  - conversion into 2bit images
  - seperation writing and background
  - edge detection
  - segmentation
- „feature" extraction
- classification/clustering

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

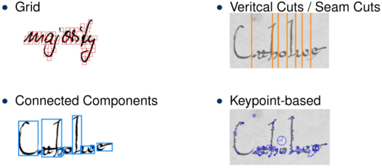
$k \in \{$ 'a' … 'z', '0' … '9'$\}$

Transcription ii

Typical phenomena

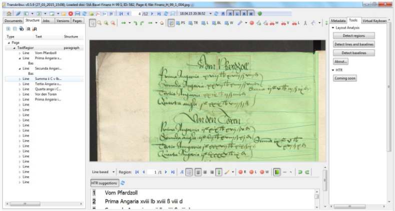
- “Special characters”
- Abbreviations
- damaged or unreadable text
- additions, deletions, substitutions, corrections
- editorial interventions (emendations and conjectures)
- editorial additions or omissions

Preprocessing: e.g. different segmentation methods



- Grid
- Veritcal Cuts / Seam Cuts
- Connected Components
- Keypoint-based

Transkribus



Transkribus v0.5.9 (27_03_2015_15:08), Loaded doc: StA Basel Finanz H 99 1, ID: 582, Page 4, file: Finanz_H_99_1_004.jpg

Type	Text	Structure
Page		
TextRegion		paragraph
Line	Vom Pfardzoll	
Line	Prima Angaria x...	
Bas		
Line	Secunda Angari...	
Bas		
Line	Summa ii C v lb...	
Line	Tertia Angaria x...	
Line	Quarta anga i C...	
Line	Vor den Toren	
Line	Prima Angaria i...	

1 Vom Pfardzoll
2 Prima Angaria xviii lb xviii ß viii d

Transcription iii

Transcriptions contain:

- layout
- additions
- corrections
- modifications
- voids, space, holes, gaps ...
- alternative transcriptions
- editorial interventions
- Enhanced transcription

```
<pb>, <lb>, <cb>, <hi>, <g>,
<handShift/>
<add>, <addSpan>,
<corr>, <del>, <delSpan>, <sic>
<subst>
<gap>, <damage>
<choice>, <alt>
<unclear>, <supplied>, <reg>
<add>, <addSpan>, <corr>, <choice>,
<damage>, <del>, <delSpan>,
<restore>, <gap>, <sic>
```

The script: palaeography

- soft hyphen: *@break="no"*
- *@rend | @rendition*
  - *@rend*: verbal description; each word describes a single facet (*rend="indented:5cm"*)
  - *@rendition*: reference to description of the rendition in the *teiHeader//encodingProfile*
- „special characters":
  - *<g>*
  - Does it exist in Unicode? (*http://www.unicode.org*). As an entity in XML:

```
&#x[hexadecimal code];
&#[decimal code];
```

Unicode for historical texts

- *Combining Diacritical Marks* (0300–036F) and *Supplement* (1DC0–1DFF): Superscripts, Subscripts
- *Latin Extended Additional* (1E00–1EFF): characters with diacritics
- *Latin Extended-D* (A720–A7FF): Ligatures, abbreviations, ...
- *General Punctuation* (2000–206F) and *Supplement* (2E00–2E7F)
- *Ancient Symbols* (10190–101CF): roman measurements, coins...
- Search unicode entities: *https://unicode-table.com/*

Editorial interventions

- expansion of abbreviations
- Conjectures
- Normalisations

- *<abbr>, <expan>* plus *<am>, <ex>*
- *<sic>, <corr>*
- *<orig>, <reg>*

All these can be paired:

- General for all editorial interventions: *<choice>*.
- Explicitly for substitutions: *<subst>*.
- *<supplied>* for additions by the editor
- *<unclear>* for unreadable text (*@reason, @agent, @hand*)

Abbreviations i

In Western MS, we usually distinguish:

- **Suspensions:** the first letter or letters of the word are written, generally followed by a point : for example ‘e.g.’ for ‘exempla gratia’
- **Contractions:** both first and last letters are written, generally with some mark of abbreviation such as superscript strokes, or points : e.g. ‘Mr.’ for ‘Mister’
- **Brevigraphs:** Special signs such as the Tironian nota used for ‘et’, the letter p with a barred tail used for ‘per’, the letter c with a circumflex used for ‘cum’/’con’ etc.
- **Superscripts:** Superscript letters (vowels or consonants) used to indicate various kinds of contraction: e.g. ‘w’ followed by superscript ‘ch’ for ‘which’.

- *<expan>* The element content is considered as the expansion of an abbreviation. In the text: USA → transcription:

```
<expan>United States of America</expan>
```

- *<abbr>* The element content is an abbreviation

```
<abbr>USA</abbr>
```

- *<ex>* (expansion) and *<am>* (abbreviation mark) for the omitted part of the abbreviation, e.g.

```
e<ex>xempla</ex> g<ex>ratia</ex>
e<am>.</am> g<am>.</am>
```

Abbreviations ii

Abbreviations can also be considered as alternatives: *`<choice>`*, e.g. ‘Zum Beispiel’ and ‘z.B.’:

```
<choice>
 <expan>Zum Beispiel</expan>
 <abbr>Z.B.</abbr>
</choice>
```

Or respectively:

```
Z<choice>
   <am>.</am>
   <ex>um</ex>
  </choice>

B<choice>
   <am>.</am>
   <ex>eispiel</ex>
 </choice>
```

Transcription = Interpretation

‘gemination dash’ – possible solutions:

- Uncommented expansion
- Unicode m with ”combining macron”
- Encoding as an XML-Entity
- *<g>* refering to *<charDecl>*
- Only *<am/>* for the stroke
- Only *<ex>* for the expansion
- As a *<choice>* with *<abbr>* and *<expan>*, the first incl. a abbreviation mark *<am>* and the second the expansion *<ex>*

Gemination dash

horizontal, bended or curved stroke above a nasal letter indicating the omission of a further instance of the same letter. (source)

Modifications

- addition, deletion, substitution, transpose, *or:*
- modification (represents any kind of general modification without interpretation)

Changing writer

<handShift />	*@new* : the hand which writes from this place onward
<handDesc>	(part of *msDesc*)
<handNotes>	(part of *profileDesc*)
<handNote>	for a particular description
@xml:id	an identifier for the hand

Text and images i

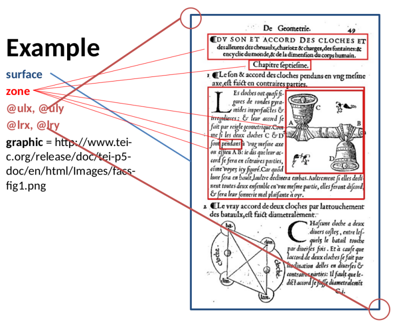
Images of a text are encoded in a *facsimile* – structure parallel to *teiHeader* and *text*:

```
<tei>
  <teiHeader>...</teiHeader>
  <facsimile> ...</facsimile>
  <text>...</text>
</tei>
```

<facsimile>

- *<surface>* = something meant to be seen
  - *@uly, @ulx; @lrx, @lry* =upper left x/y- and lower right y/x coordinates
  - coordinates form a grid, which can be referred → *@ulx* and *@uly* are usually 0
- *<graphic>*: image, *@url* : image file
- *<zone>* = an area on the surface. Coordinates refer to the grid defined in *@uly*, *@ulx*; *@lrx, @lry* of the *<surface>*.

Text and images ii



Example

surface

zone

@ulx, @uly

@lrx, @lry

graphic = http://www.tei-c.org/release/doc/tei-p5-doc/en/html/Images/facs-fig1.png

De Geometrie. 49

¶DV SON ET ACCORD DES CLOCHES ET desalleures des cheuaulx, chariotz & charges, des fontaines: & encyclie du monde, & de la dimenſion du corps humain.

Chapitre ſeptieſme.

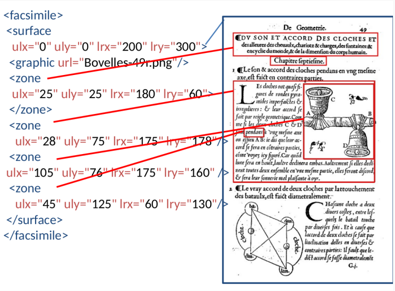
1 ¶Le ſon & accord des cloches pendans en vng meſme axe, eſt faict en contraires parties.

Les cloches ont quaſi figures de rondes pyramides imperfaictes & irregulieres: & leur accord ſe fait par reigle geometrique. Comme ſi les deux cloches C & D ſont pendans à vng meſme axe ou eſſieu A B: ie dis que leur accord ſe fera en cõtraires parties, cõme voyez icy figuré. Car quãd lune ſera en hault, laultre declinera embas. Aultrement ſi elles declinent toutes deux enſemble en vne meſme partie, elles feront diſcord, & ſera leur ſonnerie mal plaiſante à oyr.

2 ¶Le vray accord de deux cloches par lattouchement des bataulx, eſt faict diametralement.

Haſcune cloche a deux diuers coſtez, entre leſquelz le batail touche par diuerſes fois. Et à cauſe que laccord de deux cloches ſe fait par inclination delles en diuerſes & contraires parties: Il fault que ledict accord ſe faſſe diametralemẽt

G.i.



```
<facsimile>
 <surface
  ulx="0" uly="0" lrx="200" lry="300">
  <graphic url="Bovelles-49r.png"/>
  <zone
  ulx="25" uly="25" lrx="180" lry="60">
  </zone>
  <zone
   ulx="28" uly="75" lrx="175" lry="178" />
  <zone
 ulx="105" uly="76" lrx="175" lry="160" />
  <zone
   ulx="45" uly="125" lrx="60" lry="130"/>
 </surface>
</facsimile>
```

De Geometrie. 49

¶DV SON ET ACCORD DES CLOCHES ET desalleures des cheuaulx, chariotz & charges, des fontaines: & encyclie du monde, & de la dimenſion du corps humain.

Chapitre ſeptieſme.

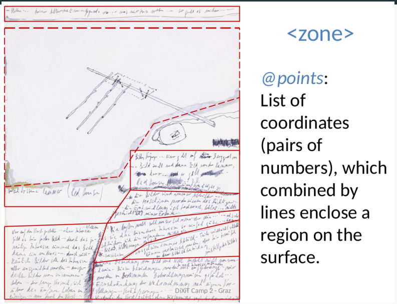
1 ¶Le ſon & accord des cloches pendans en vng meſme axe, eſt faict en contraires parties.

Les cloches ont quaſi figures de rondes pyramides imperfaictes & irregulieres: & leur accord ſe fait par reigle geometrique. Comme ſi les deux cloches C & D ſont pendans à vng meſme axe ou eſſieu A B: ie dis que leur accord ſe fera en cõtraires parties, cõme voyez icy figuré. Car quãd lune ſera en hault, laultre declinera embas. Aultrement ſi elles declinent toutes deux enſemble en vne meſme partie, elles feront diſcord, & ſera leur ſonnerie mal plaiſante à oyr.

2 ¶Le vray accord de deux cloches par lattouchement des bataulx, eſt faict diametralement.

Haſcune cloche a deux diuers coſtez, entre leſquelz le batail touche par diuerſes fois. Et à cauſe que laccord de deux cloches ſe fait par inclination delles en diuerſes & contraires parties: Il fault que ledict accord ſe faſſe diametralemẽt

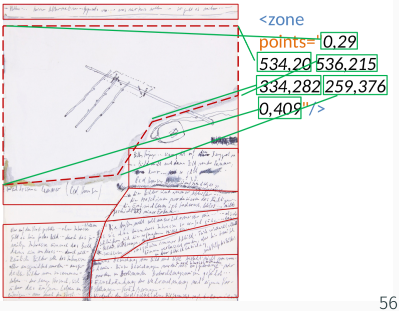
G.i.



<zone>

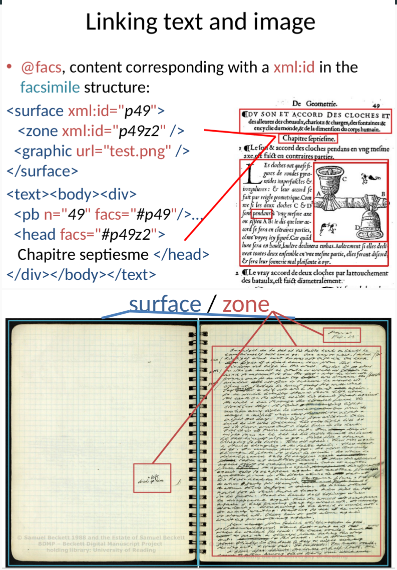
@points: List of coordinates (pairs of numbers), which combined by lines enclose a region on the surface.

DiXiT Camp 2 - Graz



```
<zone
points="0,29
534,20 536,215
334,282 259,376
0,409"/>
```

Text and images iii



Linking text and image

- @facs, content corresponding with a xml:id in the facsimile structure:

```
<surface xml:id="p49">
  <zone xml:id="p49z2" />
 <graphic url="test.png" />
</surface>
<text><body><div>
 <pb n="49" facs="#p49"/>...
 <head facs="#p49z2">
  Chapitre septiesme </head>
</div></body></text>
```

surface / zone

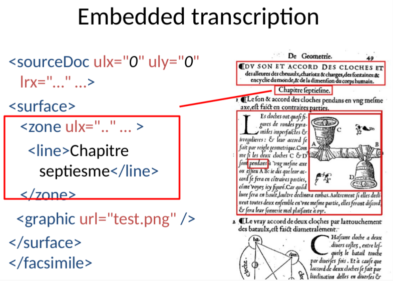
Tools for text-image linking

1. Image markup tool (Martin Holmes, *http://www.tapor.uvic.ca/~mholmes/image_markup/index.php*)
2. TextGridLab: http://www.textgridlab.de
3. T-PEN (*http://www.t-pen.org*)
4. *http://imagecoordinates.com*

Text and images iv

Embedded transcription

- "Embedded transcription": Text directly in *<surface>*
- *Relevant elements: <sourceDoc>*, *<surface>*, *<zone>*, *<line>*, *@rotate*



Embedded transcription

```
<sourceDoc ulx="0" uly="0"
  lrx="..." ...>
<surface>
  <zone ulx=".." ... >
    <line>Chapitre
      septiesme</line>
  </zone>
  <graphic url="test.png" />
</surface>
</facsimile>
```

De Geometrie. 49

DV SON ET ACCORD DES CLOCHES ET desalleures des cheuaulx, chariotz & charges, des fontaines: & encyclie du monde, & de la dimenſion du corps humain.

Chapitre ſeptieſme.

1 Le ſon & accord des cloches pendans en vng meſme axe, eſt faict en contraires parties.

Es cloches ont quaſi figures de rondes pyramides imperfaictes & irregulieres: & leur accord ſe fait par reigle geometrique. Comme ſi les deux cloches C & D ſont pendans à vng meſme axe ou eſſieu A B: ie dis que leur accord ſe fera en cõtraires parties, cõme voyez icy figuré. Car quãd lune ſera en hault, laultre declinera embas. Aultrement ſi elles declinent toutes deux enſemble en vne meſme partie, elles feront diſcord, & ſera leur ſonnerie mal plaiſante à oyr.

2 Le vray accord de deux cloches par lattouchement des bataulx, eſt faict diametralement.

Haſcune cloche a deux diuers coſtez, entre leſquelz le batail touche par diuerſes fois. Et à cauſe que laccord de deux cloches ſe fait par linclination delles en diuerſes &

```
<sourceDoc>
 <surfaceGrp n="leaf1">
  <surface facs="page1.png"> <zone>All the writing on page 1</zone> </surface>
  <surface>
   <graphic url="page2-highRes.png"/>
   <zone> <line>A line of writing on page 2</line> </zone>
  </surface>
 </surfaceGrp>
</sourceDoc>
```

Further suggestions

Stemmatology in TEI

<eTree> each part of the tree which can have descendants

<eLeaf> each part of the tree, which has only ancestors

@type e.g. hypothetical, extant, lost ...

<label> for the short names („Sigla“)

<ptr> for „contaminations“ i.e. texts influenced by other manuscript traditions

Tools for collation

Juxta Commons Texts are reduced to flat text. Variants are encoded in the parallel segmentation method.

CollateX Creates a graph. Compares every version with the existing graph and searches for gaps.

TEI Critical Apparatus Toolbox

- *http://teicat.huma-num.fr/*
- by Marjorie Burghart
- **Check encoding:** consistency etc.
- **Display parallel versions.**
- **Print an edition of a TEI XML edition,** with a TEI-to-LaTeX and PDF transformation (*reledmac*! → XSL is here).
- **Annotate images:** lets you easily trace zones on an image to prepare a documentary edition (sometimes kind of buggy) → create your *<facsimile>*.
- **Get statistics** on the XML tags used in different parts of your edition plus word counts.

But we need more metadata standards...

→ both for GAMS but also to represent our Heritage Imaging data!

Primer on metadata formats

Digital data representation

→ i.e. machine processable

Digital representations

- Images
  - raster graphics (*.png*, *.jpeg*)
  - vector graphics (*.svg*)
- Text
  - plain text (`.txt`)
  - formatted text (`.docx`, `.rtf`, `.xml`, `.tex`)
- Lists & tables (*.csv*, `.xlsx`)
- Sound (*.wav*, `.midi`)
- **Objects:** (Simulated) 3D view, abstracted representation by description and images

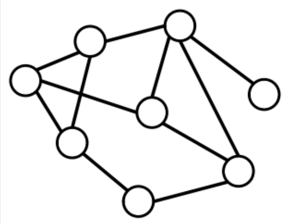
Further types

- markup languages (`.xml`, `.html`, etc.)
- data objects (*.json*, etc.)
- graphs / graph databases (`.rdf`, etc.)
- relational databases → SQL

Data structures: Graphs

Applications

- **Resource Description Framework (RDF):**
  - e.g. Blazegraph (Graph-DB)
  - Query: SPARQL
- Labeled Property Graphs
  - e.g. Neo4j (Graph-DB)
  - Query: Cypher



RDF/Turtle-Notation (`.ttl`)

```
@prefix ex: <http://example.com/#> .

ex:Graz a ex:city;
ex:name "Graz" ;
ex:inhabitants 288806 ;
ex:location [ ex:lat 47.4; ex:long 5.26 ] .

ex:Wien a ex:city;
ex:name "Wien" ;
ex:inhabitants 1897491 ;
ex:location [ ex:lat 47.12; ex:long 16.22 ] .
```

SPARQL query

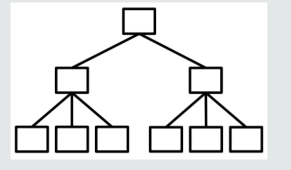
```
@prefix ex: <http://example.com/#> .

SELECT ?name, ?population
WHERE {
    ?city ex:inhabitants ?population .
}
```

Data structures: tree hierarchy 1 / XML

Applications:

eXtensible Markup Language (XML):

- DBs: eXist, BaseX
- Query: XPath (w3s) and XQuery (w3s)

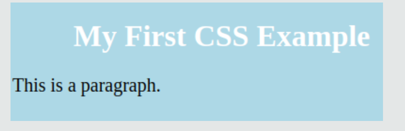


```
<!-- books.xml -->
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
  <book>
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book>
    <title lang="en">Learning XML</title>
    <price>39.95</price>
  </book>
</bookstore>

<!-- XQuery -->
for $x in doc("books.xml")/bookstore/book
where $x/price>30
order by $x/title
return $x/title

<!-- XPath -->
//title[@lang='en']
/bookstore/book[price>35.00]
```

Data structures: tree hierarchy 2 / web pages (HTML)

HTML (w3s) – structure

```
<!DOCTYPE html>
<html>
 <head>
   <title>Page Title</title>
 </head>
 <body>
  <h1>This is a Heading</h1>
  <p>This is a paragraph.</p>
 </body>
</html>
```



My First CSS Example

This is a paragraph.

CSS in HTML (w3s) – rendering

```
<!DOCTYPE html>
<html>
  <head>
    <style>
body {
  background-color: lightblue;
}
h1 {
  color: white;
  text-align: center;
}
p {
  font-family: verdana;
  font-size: 20px;
}
    </style>
  </head>
  <body>
    <h1>My First CSS Example</h1>
    <p>This is a paragraph.</p>
  </body>
</html>
```

Why so many data formats?

Different data formats (& standards) focus on different aspects & have different goals:

1. text-based
   - Text Encoding Initiative (**TEI**)
   - Extensible Hypertext Markup Language (**XHTML** = XML-compliant HTML)
   - Open Document Format for Office Applications (**ODF**)
2. page-based
   - $\TeX$ / $\LaTeX$
   - **XSL-FO** (XSL Formatting Objects, discontinued)
3. ontology-based
   - Resource Description Framework (**RDF**) & RDF Schema (**RDFS**)
   - Web Ontologie Language (**OWL**)
   - Simple Knowledge Organisation System (**SKOS**)
   - Conceptual Reference Model (**CIDOC-CRM**)
4. digital archiving / digital objects
   - Dublin Core Metadata Initiative (**DCMI**), known as Dublin Core (**DC**)
   - Metadata Encoding and Transmission Standard (**METS**)
   - Metadata Object Description Schema (**MODS**)
   - Encoded Archival Description (**EAD**)
   - Charters Encoding Initiative (**CEI**)

A primer on metadata

What are metadata?

- „data about data“
  1. data about containers of data = **structural metadata**
  2. data about the content represented by data = **descriptive metadata**
- functions:
  1. descriptive
  2. administrative
  3. technical
  4. use

There are standards for the description of metadata (and many are XML-based), e.g.

- Machine-Readable Cataloging (**MARC**)
- Metadata Object Description Schema (**MODS**)
- Encoded Archival Description (**EAD**)
- Lightweight Information Describing Objects (**LIDO**)
- Collective Description of Works of Art (**CDWA**) / Visual Research Association (**VRA**)
- Europeana Metadata Model (**EDM**)
- Resource Description Framework (**RDF**)
- Metadata Encoding & Transmission Standard (**METS**)
- Dublin Core (**DC**)
- Functional Requirements for Bibliographic Records (**FRBR**)
- the *`<teiHeader>`* has metadata…

Slides on metadata in DH with more information

Dublin Core (DC) i

What is the DC?

- founded in Dublin (Ohio) in 1995
- **two levels:** simple (15 elements) & qualified (additional *`Audience`*, *`Provenance`* and *`RightsHolder`*)
- **classes of terms:** elements (nouns) & qualifiers (adjectives).
- can be expressed in RDF/XML
- each element is optional & can be repeated
- also: dc:terms

> **The Dublin Core™ metadata standard** is a simple yet effective **element set for describing a wide range of networked resources.** […] Another way to look at Dublin Core™ is as a "small language for making a particular class of statements about resources". In this language, there are two classes of terms – *elements* (nouns) and *qualifiers* (adjectives) – which can be arranged into a simple pattern of statements.
> (source)

Dublin Core (DC) ii

The Core

i.e. the elements:

1. title
2. subject
3. description
4. type
5. source
6. relation
7. coverage
8. creator
9. publisher
10. contributor
11. rights
12. date
13. format
14. identifier
15. language

Qualified

1. (audience)
2. (provenance)
3. (rights holder)

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

   <rdf:Description rdf:about="http://media.example.com
                            /audio/guide.ra">
      <dc:creator>Rose Bush</dc:creator>
      <dc:title>A Guide to Growing Roses</dc:title>
      <dc:description>Describes process for
        planting and nurturing different kinds
        of rose bushes.</dc:description>
      <dc:date>2001-01-20</dc:date>
   </rdf:Description>
</rdf:RDF>
```

Note the two namespaces *rdf:* and *dc:*.

Metadata Encoding and Transmission Standard (METS)

METS

- tool for encoding digital library objects
- container format for documents in which contents of different formats can be integrated
- also describes relationships between objects
- describes logical and physical structure of an object
- also contains descriptive (bibliographical) and administrative metadata
- relatively simple and straightforward
- supports a wide range of materials
- <website> & more info here
- DFG-Viewer: *`http://dfg-viewer.de/`*

Only structural map is required.

```
<mets>
  <metsHdr/>
  <dmdSec/>
  <amdSec/>
  <fileSec/>
  <structMap/>
  <structLink/>
  <behaviorSec/>
</mets>
```

Goals

- link/summarize related metadata
- e.g. link related images to text
- organize data
- provide usage metadata

Resource Description Framework (RDF)

RDF

- framework for describing resources in the World Wide Web
- can contain metadata
- language of the 'Semantic Web' (web 3.0) – makes things machine-processable
- RDF Schema (RDFS) offers Classes and Properties

RDF/ turtle notation (`.ttl`) example from before

```
@prefix ex: <http://example.com/#> .

ex:Graz a ex:city;
ex:name "Graz" ;
ex:inhabitants 288806 ;
ex:location [ ex:lat 47.4; ex:long 5.26 ] .
```

Simple Knowledge Organization System (SKOS)

SKOS

RDF vocabulary for representing semi-formal *knowledge organization systems* (KOSs), such as thesauri, taxonomies, classification schemes and subject heading lists. → less rigorous than the logical formalism of ontology languages such as OWL

(SKOS primer)

```
@prefix skos:
  <http://www.w3.org/2004/02/skos/core#> .
@prefix rdf:
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:
  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix ex: <http://www.example.com/> .
@prefix ex1: <http://www.example.com/1/> .
@prefix ex2: <http://www.example.com/2/> .

ex:animals rdf:type skos:Concept;
  skos:prefLabel "animals"@en;
  skos:narrower ex:mammals.

ex:mammals rdf:type skos:Concept;
  skos:prefLabel "mammals"@en;
  skos:broader ex:animals.
```

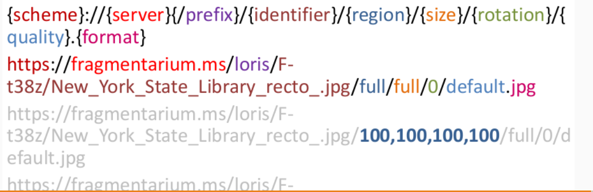
Internat. Image Interoperability Framework (read: *Triple-I-F*)

IIIF

> The International Image Interoperability Framework (`https://iiif.io/`) defines several application programming interfaces that provide a standardised method of describing and delivering images over the web, as well as "presentation based metadata" about structured sequences of images. (Wikipedia)

The standard

- proposed in 2011
- 2012: Version 1.0
- **Image API:** URL for viewing the images
- **Presentation API:** standard for describing a sequence of canvases and the images they are represented by as a representation of an object (*manifest*, `manifest.json`)



{scheme}://{server}{/prefix}/{identifier}/{region}/{size}/{rotation}/{quality}.{format}

https://fragmentarium.ms/loris/F-t38z/New_York_State_Library_recto_.jpg/full/full/0/default.jpg

https://fragmentarium.ms/loris/F-t38z/New_York_State_Library_recto_.jpg/**100,100,100,100**/full/0/default.jpg

https://fragmentarium.ms/loris/F-

IIIF Image API

region

defines the rectangular portion of the underlying image content to be returned

size

determines the dimensions to which the extracted region is to be scaled

rotation

otation parameter specifies mirroring and rotation

IIIF Presentation API

example 1, example 2

quality

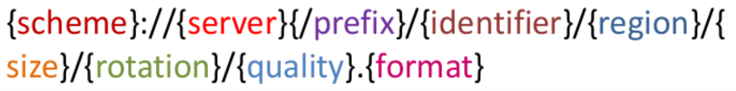
determines whether the image is delivered in color, grayscale or black and white (values: color, gray, bitonal, default)

format

format of the returned image (e.g. jpg, tif, gif, jp2, pdf, webp)

Mirador Web Viewer

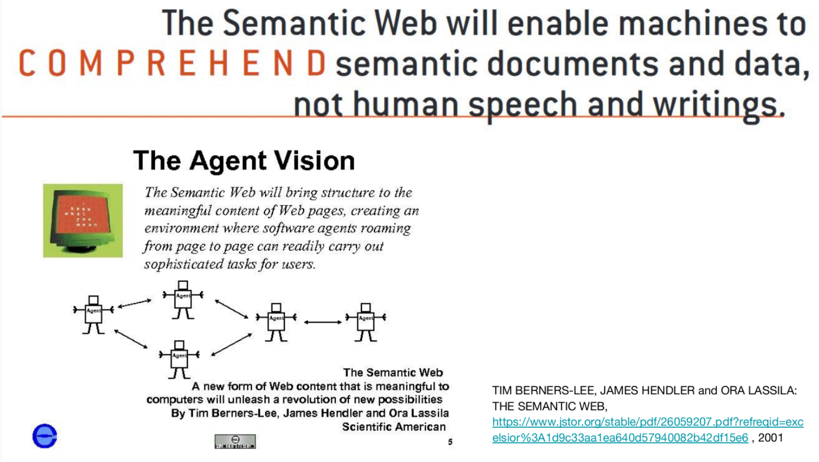
Fully featured IIIF Viewer: *`https://projectmirador.org`*



{scheme}://{server}{/prefix}/{identifier}/{region}/{size}/{rotation}/{quality}.{format}

Linked Open Data (LOD) in the Semantic Web

The Semantic Web (of Data) i



The Semantic Web will enable machines to COMPREHEND semantic documents and data, not human speech and writings.

The Agent Vision

The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.

know what you mean
Agent
Agent
Agent
Agent
Agent

The Semantic Web
A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities
By Tim Berners-Lee, James Hendler and Ora Lassila
Scientific American

5

TIM BERNERS-LEE, JAMES HENDLER and ORA LASSILA: THE SEMANTIC WEB, https://www.jstor.org/stable/pdf/26059207.pdf?refreqid=excelsior%3A1d9c33aa1ea640d57940082b42df15e6 , 2001

The Semantic Web (of Data) ii

Semantic Web

- **Semantic Web = Web of Data**
- Semantic Web technologies to query
- draw inferences using vocabularies
- DBpedia
- **Inference** = discovering new relationships.
- **Data:** modeled as a set of (named) relationships between resources.
- **Tim Berners Lee (2001):** WWW markup should encode function not content: semantic web practices to make meaning explicit for machine-processing
- **Web technologies:** RDF, OWL, SKOS, SPARQL, etc.

Goals of Linked Open Data (LOD)

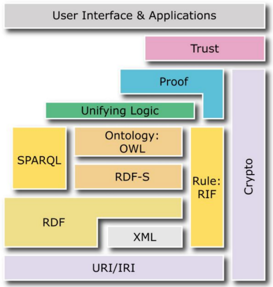
public data on the web following the same set of standards

- can be combined
- distributed, decentral 'database'
- complex queries over all this data
- world-wide knowledge and query space
- from a web of web pages to a web of data
- LOD = first step towards a semantic web
- formal modelling in the form of tripes (statements in subject, verb, object form)

The Semantic Web (of Data) iii

Ontologies

concentrate on classification methods (defining ‘classes’, ‘subclasses’, associations, relationships among classes and their instances).

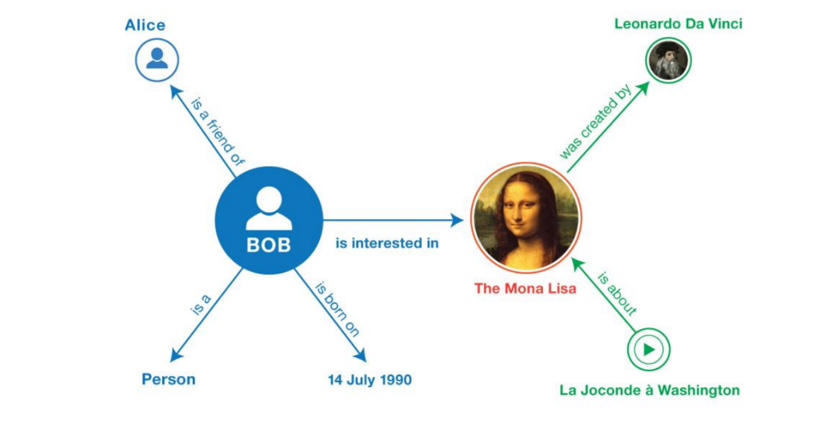


Web of Data Stack

LOD Principles

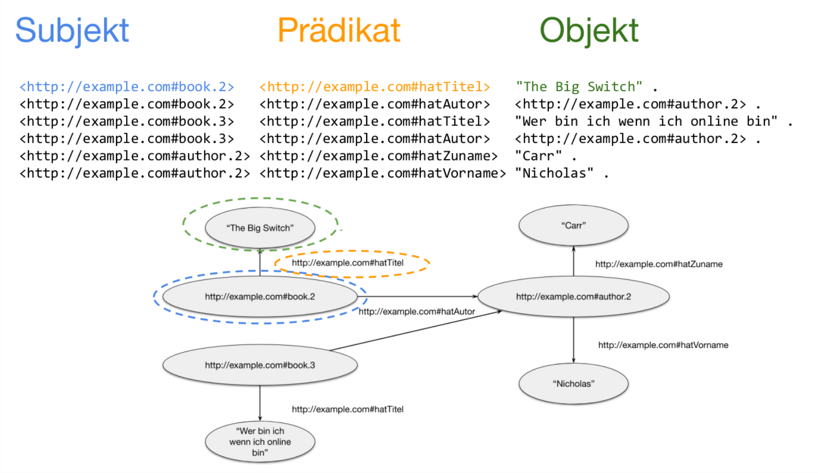
1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs. So that they can discover more things.
5. W3 LOD info

RDF, URIs, SPARQL i



https://www.w3.org/TR/rdf11-primer/

RDF, URIs, SPARQL ii

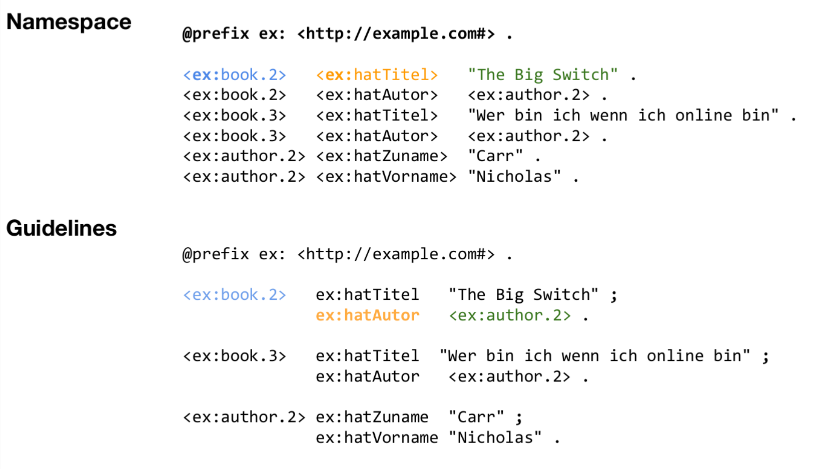


Subjekt Prädikat Objekt

```
<http://example.com#book.2>      <http://example.com#hatTitel>    "The Big Switch" .
<http://example.com#book.2>      <http://example.com#hatAutor>    <http://example.com#author.2> .
<http://example.com#book.3>      <http://example.com#hatTitel>    "Wer bin ich wenn ich online bin" .
<http://example.com#book.3>      <http://example.com#hatAutor>    <http://example.com#author.2> .
<http://example.com#author.2> <http://example.com#hatZuname>   "Carr" .
<http://example.com#author.2> <http://example.com#hatVorname> "Nicholas" .
```

“The Big Switch”
http://example.com#hatTitel
http://example.com#book.2
http://example.com#hatAutor
http://example.com#book.3
http://example.com#hatTitel
“Wer bin ich wenn ich online bin”
“Carr”
http://example.com#hatZuname
http://example.com#author.2
http://example.com#hatVorname
“Nicholas”

RDF, URIs, SPARQL iii



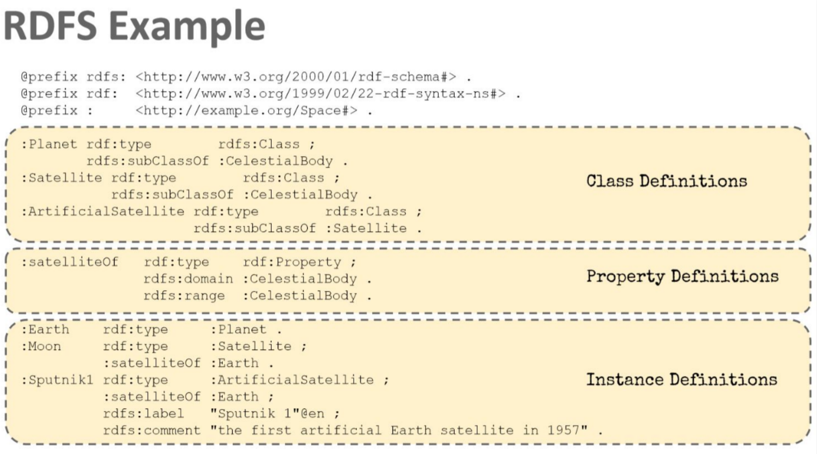
Namespace

```
@prefix ex: <http://example.com#> .

<ex:book.2>   <ex:hatTitel>   "The Big Switch" .
<ex:book.2>   <ex:hatAutor>   <ex:author.2> .
<ex:book.3>   <ex:hatTitel>   "Wer bin ich wenn ich online bin" .
<ex:book.3>   <ex:hatAutor>   <ex:author.2> .
<ex:author.2> <ex:hatZuname>  "Carr" .
<ex:author.2> <ex:hatVorname> "Nicholas" .
```

Guidelines

```
@prefix ex: <http://example.com#> .

<ex:book.2>   ex:hatTitel   "The Big Switch" ;
              ex:hatAutor   <ex:author.2> .

<ex:book.3>   ex:hatTitel  "Wer bin ich wenn ich online bin" ;
              ex:hatAutor   <ex:author.2> .

<ex:author.2> ex:hatZuname  "Carr" ;
              ex:hatVorname "Nicholas" .
```

RDF, URIs, SPARQL iv



RDFS Example

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix :     <http://example.org/Space#> .
```

Class Definitions

```
:Planet rdf:type        rdfs:Class ;
        rdfs:subClassOf :CelestialBody .
:Satellite rdf:type        rdfs:Class ;
           rdfs:subClassOf :CelestialBody .
:ArtificialSatellite rdf:type        rdfs:Class ;
                     rdfs:subClassOf :Satellite .
```

Property Definitions

```
:satelliteOf   rdf:type    rdf:Property ;
               rdfs:domain :CelestialBody .
               rdfs:range  :CelestialBody .
```

Instance Definitions

```
:Earth    rdf:type     :Planet .
:Moon     rdf:type     :Satellite ;
          :satelliteOf :Earth .
:Sputnik1 rdf:type     :ArtificialSatellite ;
          :satelliteOf :Earth ;
          rdfs:label   "Sputnik 1"@en ;
          rdfs:comment "the first artificial Earth satellite in 1957" .
```

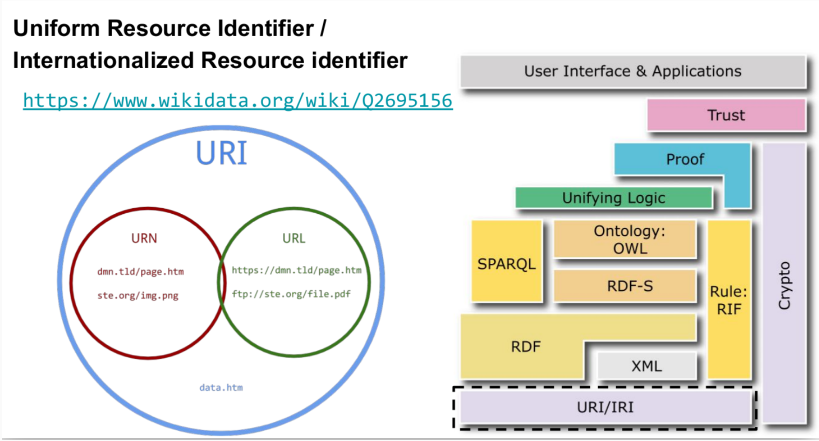
RDF, URIs, SPARQL v

URI für Graz auf

http://dbpedia.org/resource/Graz
http://dbpedia.org/data/Graz.rdf
https://www.wikidata.org/wiki/Q13298

Uniform Resource Identifier / Internationalized Resource identifier
https://www.wikidata.org/wiki/Q2695156

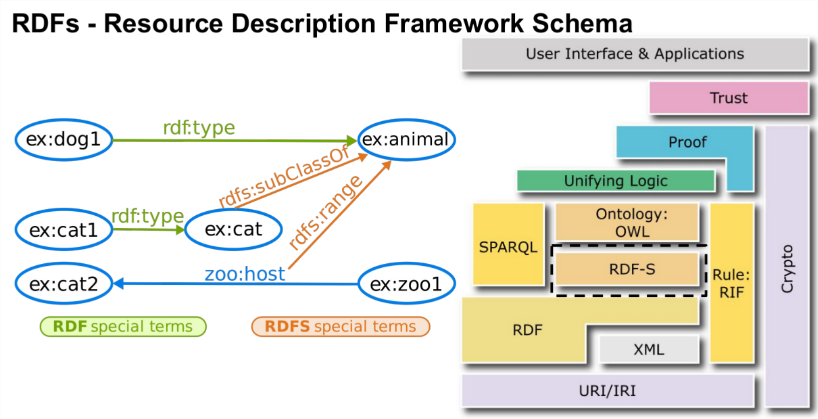
RDF, URIs, SPARQL vi



Uniform Resource Identifier / Internationalized Resource identifier

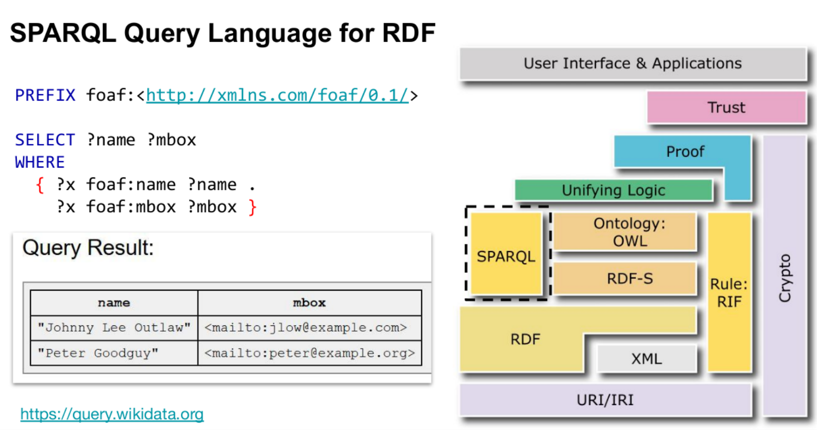
https://www.wikidata.org/wiki/Q2695156

RDF, URIs, SPARQL vii



RDFs - Resource Description Framework Schema

RDF, URIs, SPARQL viii



SPARQL Query Language for RDF

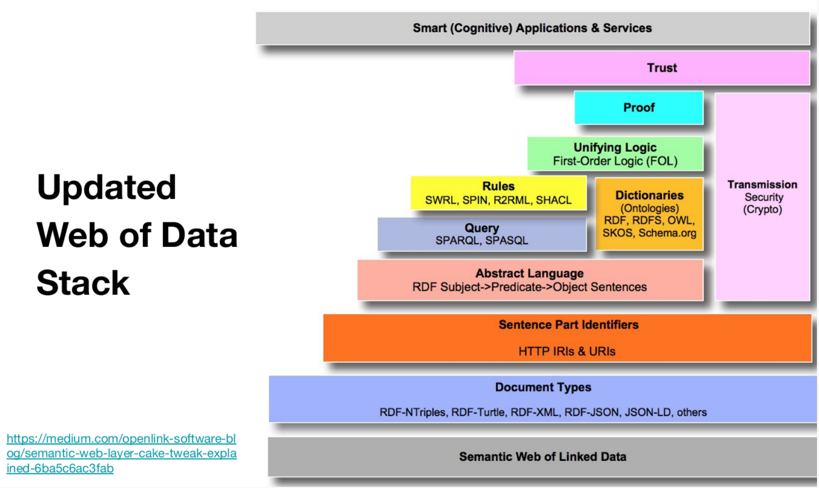
```
PREFIX foaf:<http://xmlns.com/foaf/0.1/>

SELECT ?name ?mbox
WHERE
  { ?x foaf:name ?name .
    ?x foaf:mbox ?mbox }
```

Query Result:

name	mbox
"Johnny Lee Outlaw"	<mailto:jlow@example.com>
"Peter Goodguy"	<mailto:peter@example.org>

https://query.wikidata.org

RDF, URIs, SPARQL ix



Updated Web of Data Stack

Smart (Cognitive) Applications & Services

Trust

Proof

Unifying Logic
First-Order Logic (FOL)

Rules
SWRL, SPIN, R2RML, SHACL

Query
SPARQL, SPASQL

Dictionaries
(Ontologies)
RDF, RDFS, OWL, SKOS, Schema.org

Transmission
Security
(Crypto)

Abstract Language
RDF Subject->Predicate->Object Sentences

Sentence Part Identifiers
HTTP IRIs & URIs

Document Types
RDF-NTriples, RDF-Turtle, RDF-XML, RDF-JSON, JSON-LD, others

Semantic Web of Linked Data

https://medium.com/openlink-software-blog/semantic-web-layer-cake-tweak-explained-6ba5c6ac3fab

RDF, URIs, SPARQL x

SPARQL

The SPARQL Protocol and RDF Query Language 1.1 is…

- serialized in RDF
- a query language for RDF graph traversal
- an http protocol via ‘SPARQL endpoints’
- W3C standard, XML output format
- inspired by SQL
- **is the query language of the Semantic Web.**

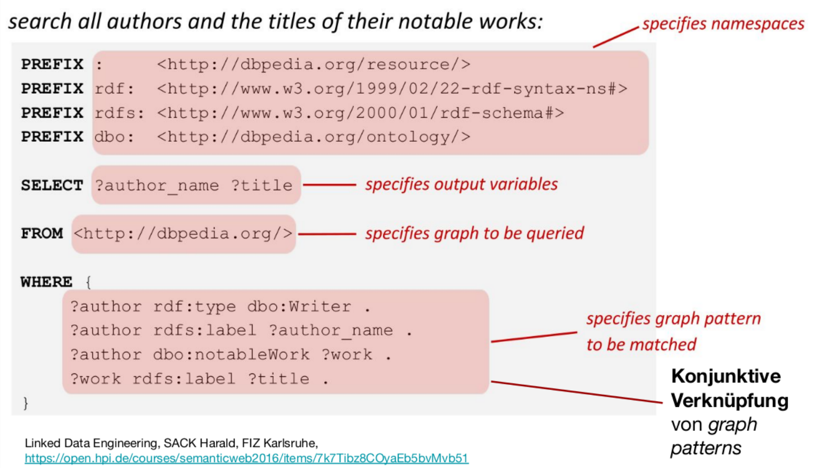
Basic graph pattern matching

A **triple pattern** is an RDF triple with variables, such as:

```
?country dbo:capital ?capital
```

- Results are returned based on whether a specific graph pattern was found in the data.
- Pull values from structured and semi-structured data
- Explore data by querying unknown relationships
- Perform complex joins of disparate databases in a single, simple query
- Transform RDF data from one vocabulary to another

RDF, URIs, SPARQL xi



search all authors and the titles of their notable works:

```
PREFIX :     <http://dbpedia.org/resource/>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo:  <http://dbpedia.org/ontology/>

SELECT ?author_name ?title

FROM <http://dbpedia.org/>

WHERE {
    ?author rdf:type dbo:Writer .
    ?author rdfs:label ?author_name .
    ?author dbo:notableWork ?work .
    ?work rdfs:label ?title .
}
```

specifies namespaces

specifies output variables

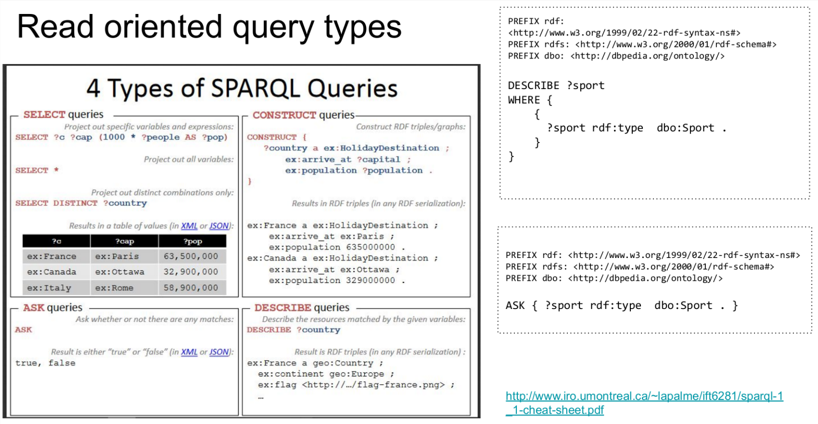
specifies graph to be queried

specifies graph pattern to be matched

Konjunktive Verknüpfung von *graph patterns*

Linked Data Engineering, SACK Harald, FIZ Karlsruhe, https://open.hpi.de/courses/semanticweb2016/items/7k7Tibz8COyaEb5bvMvb51

RDF, URIs, SPARQL xii



Read oriented query types

4 Types of SPARQL Queries

SELECT queries

Project out specific variables and expressions:

```
SELECT ?c ?cap (1000 * ?people AS ?pop)
```

Project out all variables:

```
SELECT *
```

Project out distinct combinations only:

```
SELECT DISTINCT ?country
```

Results in a table of values (in XML or JSON):

?c	?cap	?pop
ex:France	ex:Paris	63,500,000
ex:Canada	ex:Ottawa	32,900,000
ex:Italy	ex:Rome	58,900,000

CONSTRUCT queries

Construct RDF triples/graphs:

```
CONSTRUCT {
   ?country a ex:HolidayDestination ;
        ex:arrive_at ?capital ;
        ex:population ?population .
}
```

Results in RDF triples (in any RDF serialization):

```
ex:France a ex:HolidayDestination ;
    ex:arrive_at ex:Paris ;
    ex:population 635000000 .
ex:Canada a ex:HolidayDestination ;
    ex:arrive_at ex:Ottawa ;
    ex:population 329000000 .
```

ASK queries

Ask whether or not there are any matches:

```
ASK
```

Result is either "true" or "false" (in XML or JSON):

```
true, false
```

DESCRIBE queries

Describe the resources matched by the given variables:

```
DESCRIBE ?country
```

Result is RDF triples (in any RDF serialization) :

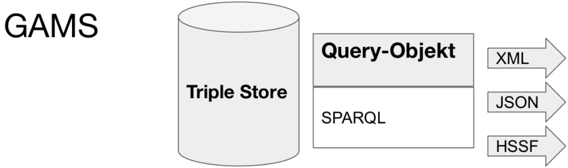
```
ex:France a geo:Country ;
  ex:continent geo:Europe ;
  ex:flag <http://…/flag-france.png> ;
  …
```

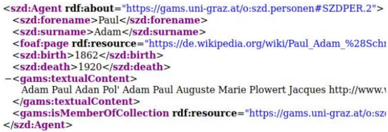
```
PREFIX rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>

DESCRIBE ?sport
WHERE {
     {
       ?sport rdf:type  dbo:Sport .
     }
}
```

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>

ASK { ?sport rdf:type  dbo:Sport . }
```

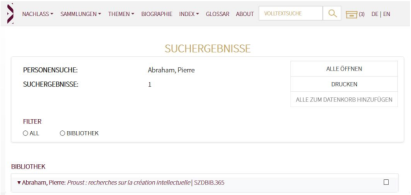
http://www.iro.umontreal.ca/~lapalme/ift6281/sparql-1_1-cheat-sheet.pdf

RDF, URIs, SPARQL xiii





```
<szd:Agent rdf:about="https://gams.uni-graz.at/o:szd.personen#SZDPER.2">
  <szd:forename>Paul</szd:forename>
  <szd:surname>Adam</szd:surname>
  <foaf:page rdf:resource="https://de.wikipedia.org/wiki/Paul_Adam_%28Schı
  <szd:birth>1862</szd:birth>
  <szd:death>1920</szd:death>
 –<gams:textualContent>
    Adam Paul Adan Pol' Adam Paul Auguste Marie Plowert Jacques http://www.
  </gams:textualContent>
  <gams:isMemberOfCollection rdf:resource="https://gams.uni-graz.at/o:sz
</szd:Agent>
```



SUCHERGEBNISSE

PERSONENSUCHE: Abraham, Pierre

SUCHERGEBNISSE: 1

FILTER

○ ALL ○ BIBLIOTHEK

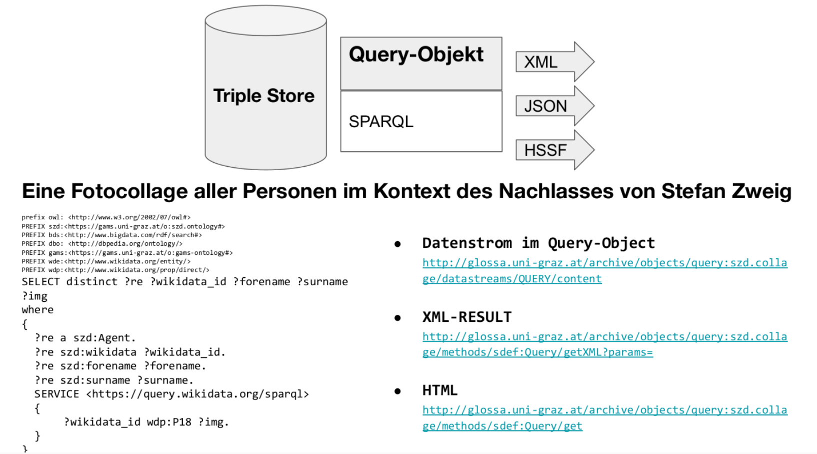
BIBLIOTHEK

▾ Abraham, Pierre: *Proust : recherches sur la création intellectuelle* | SZDBIB.365 □

http://stefanzweig.digital/o:szd.bibliothek/RDF
http://stefanzweig.digital/o:szd.personen/RDF

http://stefanzweig.digital/archive/objects/query:szd.person_search/methods/sdef:Query/get?params=%241%7C%3Chttps%3A%2F%2Fgams.uni-graz.at%2Fo%3Aszd.personen%23SZDPER.1%3E%3B%242%7Cde&locale=de

RDF, URIs, SPARQL xiv



Triple Store

Query-Objekt

SPARQL

XML

JSON

HSSF

Eine Fotocollage aller Personen im Kontext des Nachlasses von Stefan Zweig

```
prefix owl: <http://www.w3.org/2002/07/owl#>
PREFIX szd:<https://gams.uni-graz.at/o:szd.ontology#>
PREFIX bds:<http://www.bigdata.com/rdf/search#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX gams:<https://gams.uni-graz.at/o:gams-ontology#>
PREFIX wde:<http://www.wikidata.org/entity/>
PREFIX wdp:<http://www.wikidata.org/prop/direct/>
SELECT distinct ?re ?wikidata_id ?forename ?surname
?img
where
{
  ?re a szd:Agent.
  ?re szd:wikidata ?wikidata_id.
  ?re szd:forename ?forename.
  ?re szd:surname ?surname.
  SERVICE <https://query.wikidata.org/sparql>
  {
       ?wikidata_id wdp:P18 ?img.
  }
}
```

- **Datenstrom im Query-Object**
  http://glossa.uni-graz.at/archive/objects/query:szd.collage/datastreams/QUERY/content
- **XML-RESULT**
  http://glossa.uni-graz.at/archive/objects/query:szd.collage/methods/sdef:Query/getXML?params=
- **HTML**
  http://glossa.uni-graz.at/archive/objects/query:szd.collage/methods/sdef:Query/get

RDF, URIs, SPARQL xv

Linked Data Triplestores

- Blazegraph
- Stardog
- Virtuoso
- GraphDB
- AnzoGraph
- AllegroGraph
- MarkLogic
- Apache Rya

→ Addlesee Angus: *Comparing Linked Data Triplestores*

Blazegraph

- Download
- Quick Start Guide
- run the jar file: *`java -server -Xmx4g -jar blazegraph.jar`*
- open in browser: *`http://localhost:9999/blazegraph/#splash`*
- Blazegraph Wiki

Archiving Image Data

Imaging Techniques for Heritage Management

Archiving the Image Data

- Persistent Identifer (e.g Handle Server)
- Metadata (e.g., DC, OAI-PMH)

Disseminating the Image Data

- Making the data accessible
- Presenting the Data
- Web-Views (e.g. Mirador Web Viewer)

Multi-Modal Manuscript Representations (M3R) i

Resources

- Part of the DITAH project: *https://www.ditah.at/projects/m3r.html*
- **Contact:** Simon Brenner (*sbrenner@cvl.tuwien.ac.at*)
- (more information)

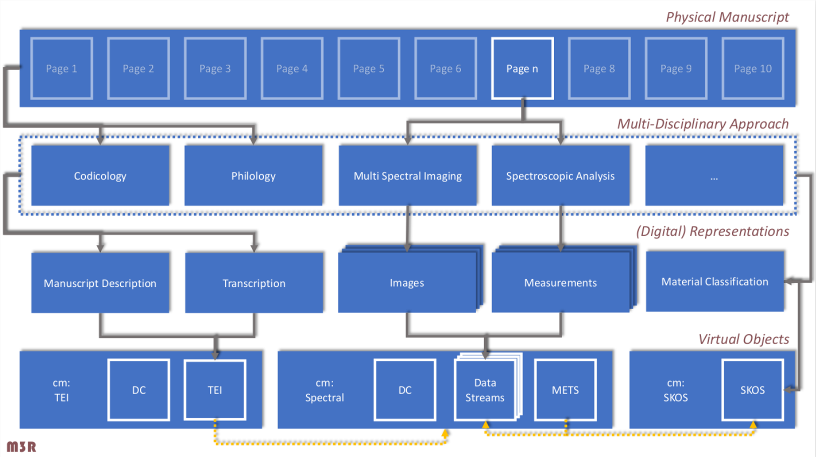
Motivation

The interdisciplinary analysis of historical manuscripts generates a variety of measuring and descriptive data: these range from multi- and hyperspectral images, spectroscopic and microbiological material analyses, codicological and restorative descriptions, to transcriptions and philological editions.

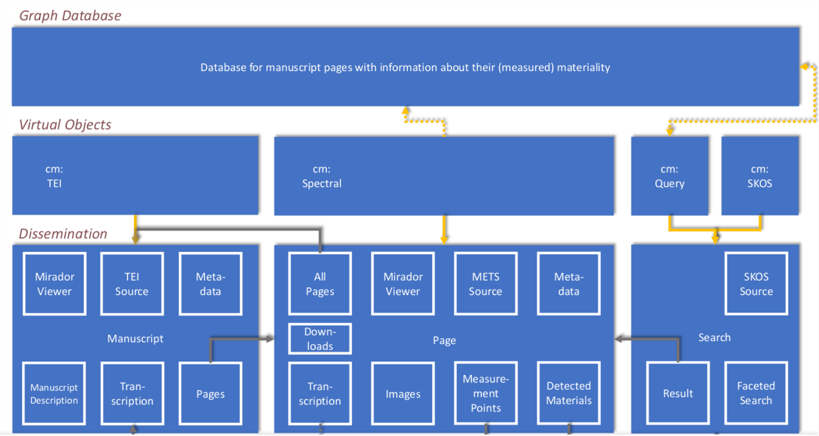
Objectives

The various data streams and metadata are spatially and logically related and combined to digital objects using established standards (IIIF, METS, TEI, SKOS).
The objects are disseminated via a web viewer as well as via technical interfaces.

Multi-Modal Manuscript Representations (M3R) ii



Multi-Modal Manuscript Representations (M3R) iii



The content model: idea i

Solution

Creation of a Multispectal Content Model for the Graz GAMS infrastructure (*`cm:Spectral`*):

- represents image and analysis data in the METS format
- implements the Mirador Viewer (via IIIF)

The goal

> “ […] the creation of a repository for the long-term archiving and dissemination of manuscript research data in the form of Multi-Modal Manuscript Representations is described, which aims at spatially and logically relating the various digital artifacts. With respect to long-term preservation and linked open data, special emphasis is put on the use of established and open standards (e.g. IIIF, METS, TEI, SKOS). The resulting virtual objects are disseminated via a web viewer as well as via technical interfaces. (Brenner, Clausen, and Schneider 2016) ”

The content model: idea ii

Natural color images

> As the most intuitive digital representation of a manuscript page, full-page natural color images serve as the basic reference in order to locate any subsequent investigations, such as material measurements or the annotations of codicological features. For philologists, they are the preferred representation for reading the text, provided the necessary legibility. Reference objects (rulers, color charts) should always be included in the images, such that physical sizes can be inferred. (Brenner, Clausen, and Schneider 2016)

Multispectral images

> Multispectral images and their processed derivatives are used by paleographers to decipher heavily degraded texts. For this task, a high spatial resolution is paramount, and (pseudo-)color images are generally preferred over single-channel (grayscale) images. Typically, multiple images of a given manuscript region are utilized to decipher a difficult portion of text. (Brenner, Clausen, and Schneider 2016)

The content model: idea iii

Spectroscopic measurements

> “Spectroscopic analysis methods such as XRF, FTIR and Raman spectroscopy produce 1-dimensional signals (spectra) for a given measurement point. While the chemical elements present are directly inferred from XRF, FTIR and Raman spectra are interpreted via comparison with databases of reference materials. Material scientists typically use the spectra resulting from multiple complementary methods to deduce material information about a given point.(Brenner, Clausen, and Schneider 2016) ”

The content model: technical implementation i

Annotations

> Under the term annotations we summarize high-level information, mostly derived from the primary data described above, with spatial reference to the manuscript surface:
>
> - **Transcriptions** are typically a result of philological and paleographic activities. When dealing with degraded texts, it is necessary to model uncertainties about transcribed letters or passages.
> - **Material information** is derived from spectroscopic measurements and expressed in terms of chemical elements and known compounds present in a given measurement spot.
> - **Codicological features** such as deletions, damages or production traits are documented by conservators. (Brenner, Clausen, and Schneider 2016)

The content model: technical implementation ii

Possible research questions

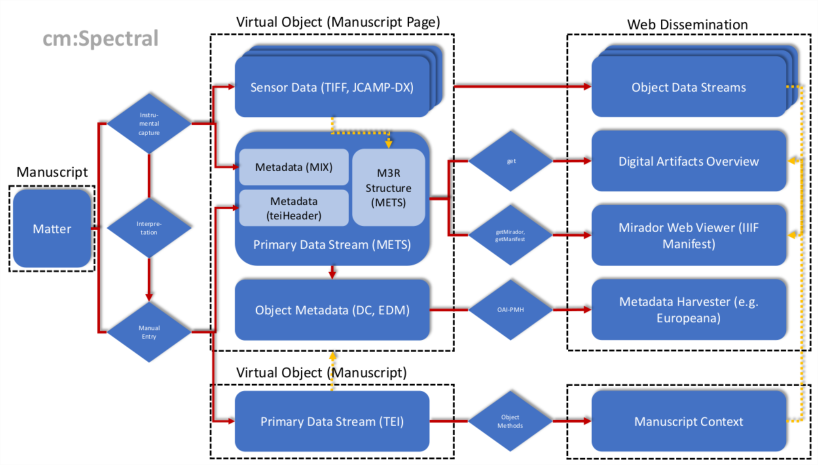
> 1. Does the ink composition within a manuscript change gradually, or are there abrupt changes?
> 2. Is the ink composition identical for all elements, or are there, e.g., notes in a different ink? These questions provide insights about the genesis of a given manuscript.
> 3. Are there similarities between different manuscripts based on the compositions of black inks, red inks and illuminations? Such findings could hint to different having collaborated or worked in the same atelier.
> 4. Can we draw general conclusion about the temporal and geographical influences on ink compositions and employed materials? (Brenner, Clausen, and Schneider 2016)

The content model: technical implementation iii

Integrating different research data

> When material analysis data are not linked to the source manuscript and their metadata in a suitable data structure, any comparative study of this kind must be conducted ’by hand’ and cannot harness the possibilities of automated data analysis and exploration. M3R should enable the dynamic querying of material measurements within all manuscripts in the repository in two ways:
>
> 1. Based on annotations of arbitrary granularity, such as ’red’, ’carbon ink’, ’vermillion’, ’water damage’ ’zoomorphic initial’, etc.
> 2. Content-based retrieval, either based on the visual appearance of the measuring spot (in natural color images or multispectral images), or on the actual spectra measured. (Brenner, Clausen, and Schneider 2016)

The content model: technical implementation iv


cm:Spectral
Virtual Object (Manuscript Page)
Web Dissemination
Manuscript
Matter
Instru-mental capture
Interpre-tation
Manual Entry
Sensor Data (TIFF, JCAMP-DX)
Metadata (MIX)
Metadata (teiHeader)
M3R Structure (METS)
Primary Data Stream (METS)
Object Metadata (DC, EDM)
Virtual Object (Manuscript)
Primary Data Stream (TEI)
get
getMirador, getManifest
OAI-PMH
Object Methods
Object Data Streams
Digital Artifacts Overview
Mirador Web Viewer (IIIF Manifest)
Metadata Harvester (e.g. Europeana)
Manuscript Context

The content model: GAMS `cm:Spectral` i

GAMS `cm:Spectral`

> “ Each object of the type unites the digital artifacts referring to a page as object data streams and is understood as a **virtual proxy for a page** (of a manuscript).
> During the automated creation of such an object (*ingest process*), **a further primary data stream is also generated** and stored in the object.
> **This primary data stream is a METS document** that assembles the individual digital artifacts of a page into a structure consistent with the M3R logic of the project. (Brenner, Clausen, and Schneider 2016) ”

The content model: GAMS `cm:Spectral` i

A `cm:Spectral` object

> ...unites the following types of data streams:
>
> 1. **Primary data stream:** Document describing the virtual object as a multimodal structure in the format: METS (text/xml).
> 2. **Digital artifacts of the manuscript page**
>   - Document fragment (as a part of METS document) about the metadata of the manuscript in the format: TEI (text/xml).
>   - Natural color and Multispectral images in format: TIFF (image/tiff).
>   - Spectroscopic measurements in format: JCAMP-DX (text/plain).
>   - Annotations of different disciplines in the format: RDF (text/xml).
> 3. **Metadata** about the virtual object in format: DC (text/xml).
>
> (Brenner, Clausen, and Schneider 2016)

These slides are partially based on (teaching) materials provided by: Georg Vogeler, Christopher Pollin, Johannes Stiegler, Sean Winslow, Hans Clausen, Gerlinde Schneider and Simon Brenner.

References

[1] Simon Brenner, Hans Clausen, and Gerlinde Schneider. “Multi-Modal Manuscript Representations – towards an interdisciplinary open resource”. In: *42nd Conference on Very Important Topics (CVIT 2016)*. Ed. by John Q. Open and Joan R. Access. OpenAccess Series in Informatics. Dagstuhl: Dagstuhl Publishing, 2016, 23:1–23:9.

[2] Patrick Sahle. “Digitales Archiv und Digitale Edition. Anmerkungen zur Begriffsklärung”. In: *Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien*. Ed. by Michael Stolz. Zürich, 2007, pp. 64–84.

[3] Patrick Sahle. “What is a Scholarly Digital Edition?” In: *Digital Scholarly Editing: Theories and Practices*. Ed. by Matthew J. Driscoll and Elena Pierazzo. Cambridge: Open Book Publisher, 2016, pp. 19–39. URL: *https://books.openedition.org/obp/3381*.