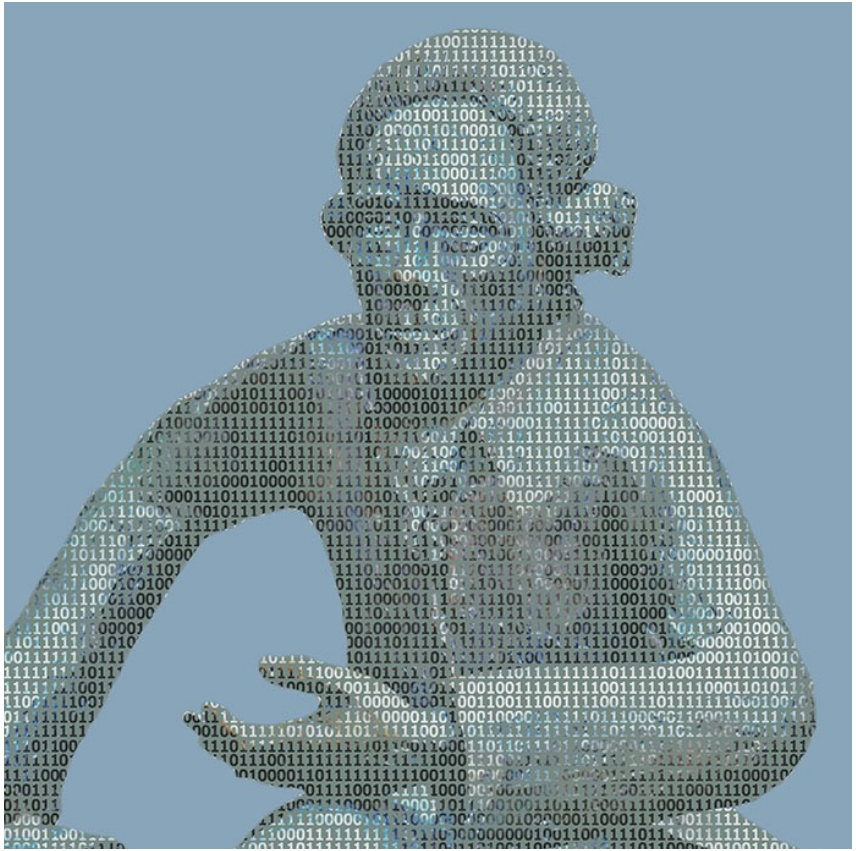
II. Digitisation And Data management

2. Experiencing- Expressing- Understanding: How do you digitise culture?

Prof. Dr. Martin Langner



Schreibman / Siemens / Unsworth (2004) Kap. 13–19;
Schreibman / Siemens / Unsworth (2016) Kap. 11–12;
Jannidis / Kohle / Rehbein (2017) Kap. 4–7. 9. 11. 18. 19

Cultural Heritage

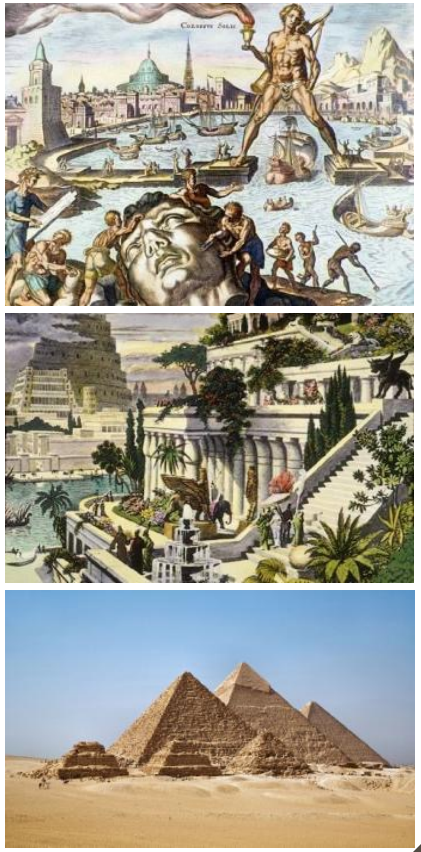
»Damage to cultural property, belonging to any people whatsoever, means damage to the cultural heritage of all mankind, since each person makes its contribution to the culture of the world.«

The Hague Convention, 1954

Nicolas Poussin, Les Bergers d'Arcadie ou Et in Arcadia ego, 1638/40, Paris, Louvre 7300

DOBES – Documentation of endangered languages http://dobes.mpi.nl/

Seven Wonders of the World



‘The Two Cultures’

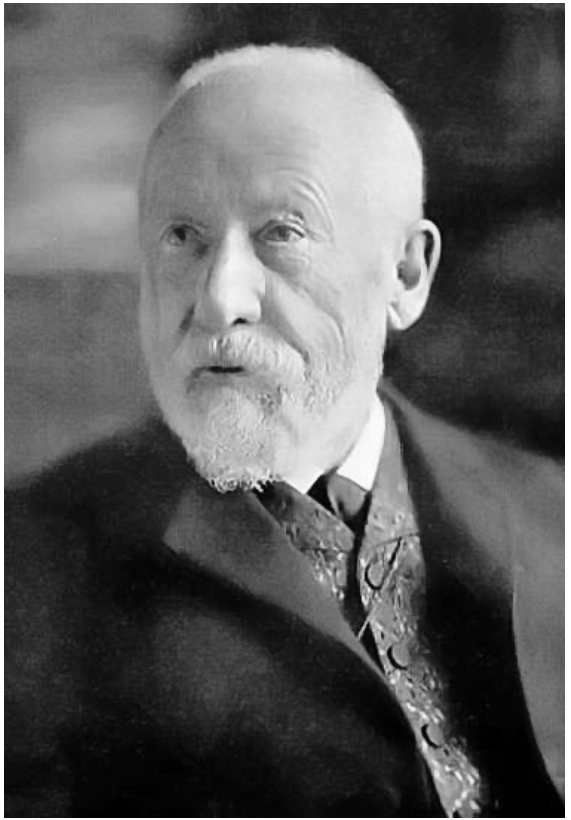
“I believe the intellectual life of the whole of western society is increasingly being split into two polar groups. […] Literary intellectuals at one pole — at the other scientists, and as the most representative, the physical scientists. Between the two a gulf of mutual incomprehension—sometimes (particularly among the young) hostility and dislike, but most of all lack of understanding. They have a curious distorted image of each other. Their attitudes are so different that, even on the level of emotion, they can't find much common ground.“

Charles Percy Snow, *The Two Cultures* (London 1959)
http://s-f-walker.org.uk/pubsebooks/2cultures/Rede-lecture-2-cultures.pdf

Historisch-hermeneutische Ansätze

„Humanity, conceived in terms of perception and cognition, would be a physical fact for us and as such would only be accessible to scientific cognition. As an object of the humanities, however, it arises only insofar as human circumstances are **experienced**, insofar as they are **expressed** in expressions of life, and insofar as these expressions are **understood**.“

Wilhelm Dilthey (1910), *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften*, Ges. Schriften 7, S. 87



Wilhelm Dilthey (1833–1911)

Nomothetic Approaches

„In the knowledge of reality, the empirical sciences seek either the general in the form of natural law or the individual in the historically determined form; they look in one part at the always unchanging form, in the other part at the unique, intrinsically determined content of real events. The one are **sciences of law**, the other **sciences of events**; the former teach what always is, the latter what once was. Scientific thought is - if one may form new terms of art - nomothetic in the one case, idiographic in the other.“

Wilhelm Windelband (1894), *Geschichte und Naturwissenschaft*,

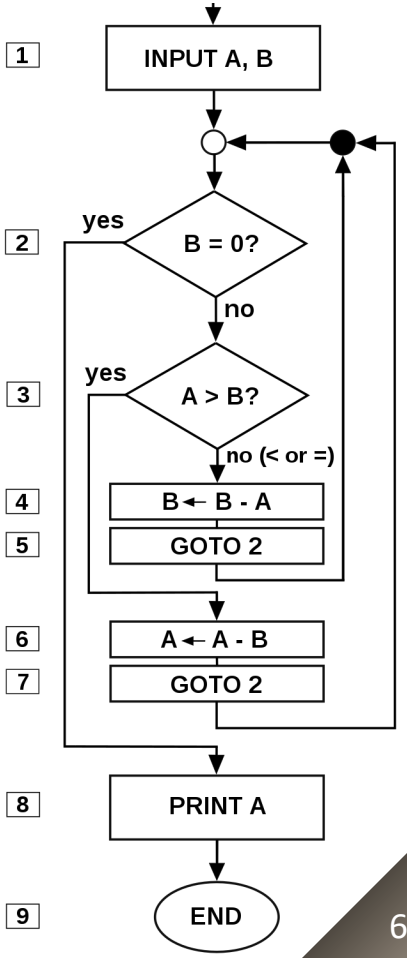
Wilhelm Windelband (1848–1915)

Alan Mathison Turing (1912–1954)

Algorithmic Approaches

"Let us imagine that the operations performed by the computer to be split up into 'simple operations' which are so elementary that it is not easy to imagine them further divided."

Alan M. Turing, On Computable Numbers, With An Application to the Entscheidungs-problem, Proceedings of the London Mathematical Society. Series 2. 42, 1936–37, 230–265

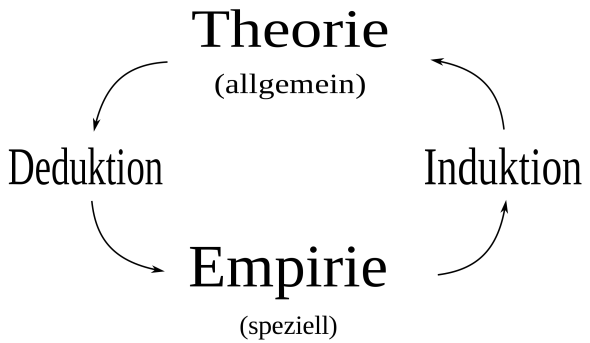


Euklid‘s algorithm to calculate the greatest common divisor

Deduction and Induction

Deduction
All human beings are mortal (premise)
Socrates is a human being. (fact)
→ Socrates is mortal. (consequence)

Induction
Plato, Aristotele and Epicurus were human beings. (facts)
Plato, Aristotele and Epicurus have died. (consequence)
→ All human beings are mortal. (general principle)



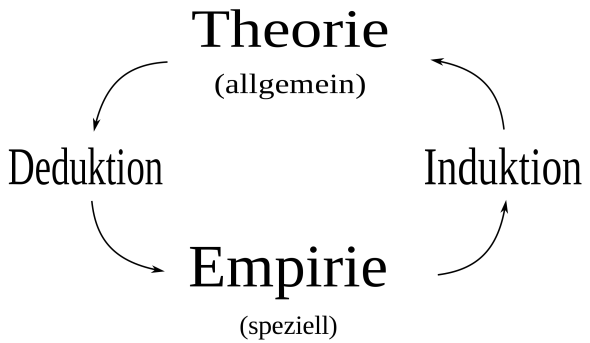
Schematic representation of the relationship between theory, empiricism, induction and deduction, as it has been represented since Aristotle.

Deduction and Induction

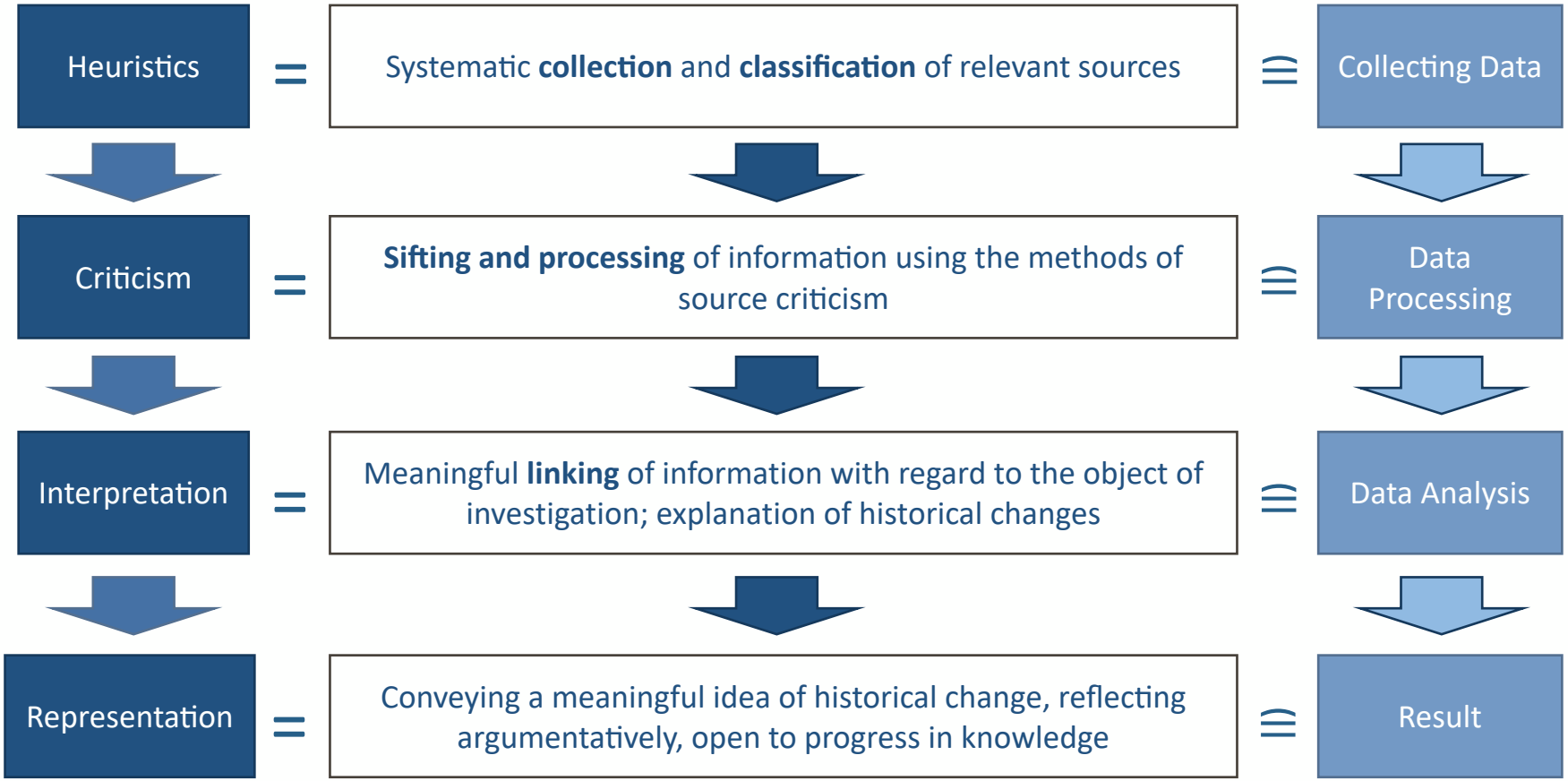
An algorithm is a formalised, **unambiguous** set of instructions for solving a problem or a class of problems. Algorithms consist of a finite number of well-defined individual steps.

Heuristics, on the other hand, work with limited knowledge (= incomplete information) and try to arrive at probable statements or practicable solutions via a **variety** of analogies and other **conjectural** conclusions.

The exactness of the computer is contrasted with the proof of probability in the humanities.



Schematic representation of the relationship between theory, empiricism, induction and deduction, as it has been represented since Aristotle.


Heuristics
=
Systematic **collection** and **classification** of relevant sources
≡
Collecting Data
Criticism
=
Sifting and processing of information using the methods of source criticism
≡
Data Processing
Interpretation
=
Meaningful **linking** of information with regard to the object of investigation; explanation of historical changes
≡
Data Analysis
Representation
=
Conveying a meaningful idea of historical change, reflecting argumentatively, open to progress in knowledge
≡
Result

HUMANITIES

- Interpretation and evaluation of phenomena
- Methods are less standardised
- Statements more qualitative, less formalised
- Results have to be seen, interpreted and evaluated in the overall context.

Structuring the data is an important hermeneutic act. Processing data with the methods and tools of computer science does not lead to unambiguous and neutral results, but is always a scientific construction of the facts.

COMPUTER SCIENCE

- Analysis of exact, quantitatively measurable statements
- statistically evaluable

1. "EXPERIENCING": ACQUISITION OF THE RESEARCH OBJECTS

a) Scanning

b) OCR

c) Layout

2. "EXPRESSING": FORMALISING THE RESEARCH OBJECTS

a) Structure

b) HTML and XML

c) TEI

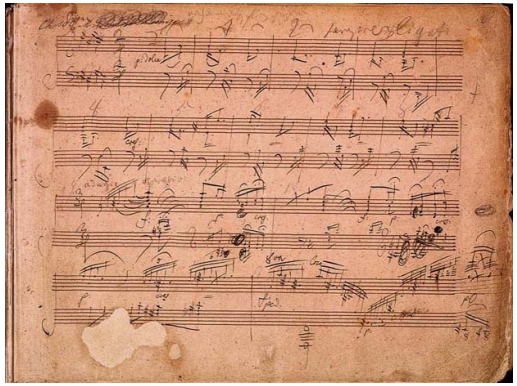
d) Metadata

e) Authority records

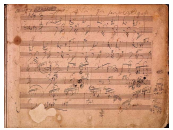
3. "UNDERSTANDING": SEMANTICS OF THE RESEARCH OBJECTS

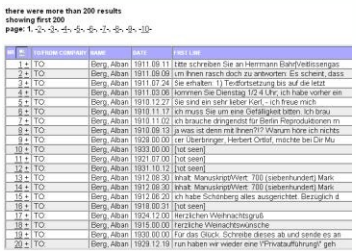
Ontologies, RDF and Linked Open Data

1. "EXPERIENCING": ACQUISITION OF THE RESEARCH TOPICS

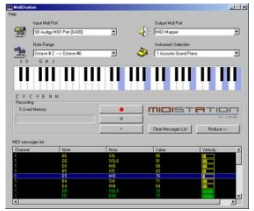


Beethoven, Piano Sonata op. 109





links



midi

pdf

XSTL

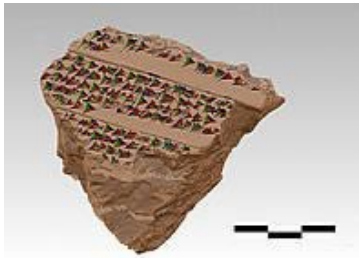
HTML

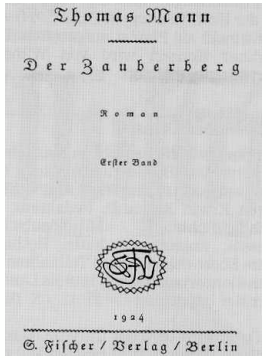
jpg

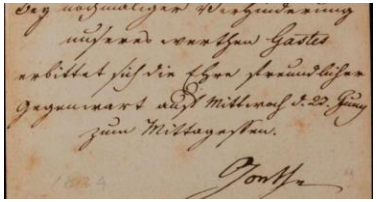
Source **Scan/OCR** **Image / Metadata (XML)** **Database**

Digital Scanning

- creates a digital image using a scanner or a digital camera.
- Camera -> creates a digital image of the text (e.g. in pdf or jpg) by placing a virtual grid over the text and recording the colour or grey value of each cell (= pixel).
- Scans, i.e. the digital images, can be viewed with suitable viewing software (e.g. a PDF viewer).
- but they cannot be read by a computer -> for the computer it is an image, not a text





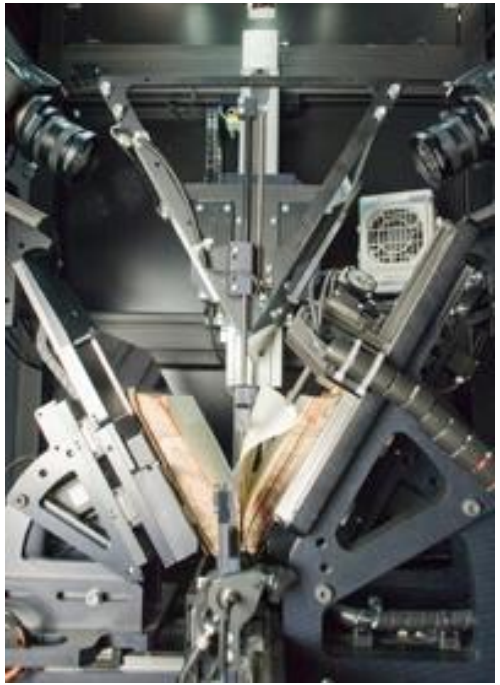


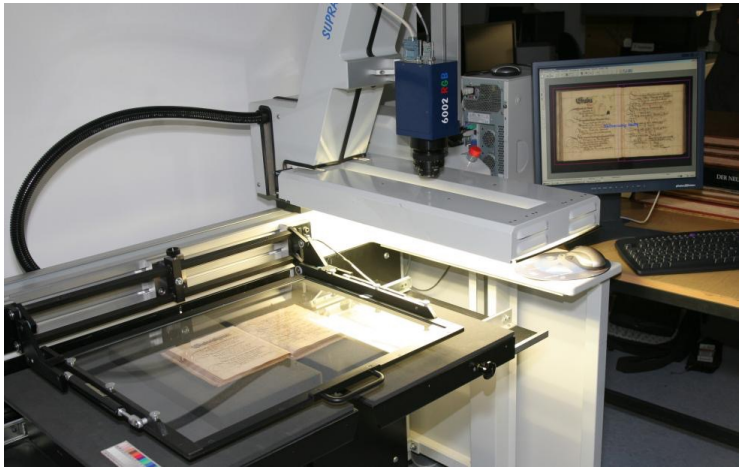
Challenges

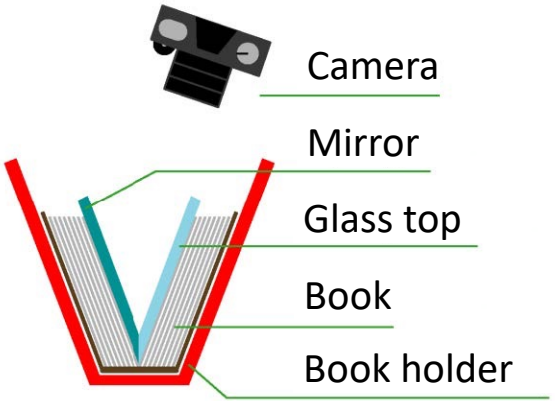
- Speed and cost; how much can be automated (e.g. scrolling)?
- Fragility of the documents
- How much handling is acceptable from a conservation point of view?
- How far can old books be opened without damaging the binding?
- How much exposure to light?
- Quality and reliability of the digital image (lighting, camera / scanner quality, resolution, colour depth, use of colour wedge and tape measure for calibration etc.)
- Which digital format (tiff vs. jpeg etc.)?

Recommended reading: DFG Practical Guidelines on Digitisation [12/16]
https://www.dfg.de/formulare/12_151/12_151_en.pdf

Scanner and Scan Robots







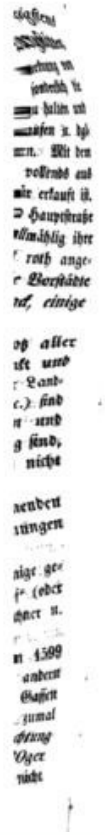
Scanning Problems

nummern geschaffen werden können,
Das Speichern in Xanadu, Nelson nann
die Möglichkeit, eine virtuelle Kopie
eine
Das

nummern geschaffen werden können,
Das Speichern in Xanadu, Nelson nann
die
eine
Das

nummern geschaffen werde
Das
die Ausgangs-
eines Dokumentes zu erstel
Das Ausgangs-Dokument

Making texts machine-readable



Manual Transcription

- Input letter by letter (without language skills), but by two independent transcribers (double key)
- Guidelines, transcription tools
- Comparison of versions and correction of errors
- Decide on the best transcription for these cases -> reduces errors -> Calculate measure of transcription accuracy (percentage of words / letters correct in both transcribers) -> 99.997% accuracy

Automatic Methods

- Optical Character Recognition (OCR)

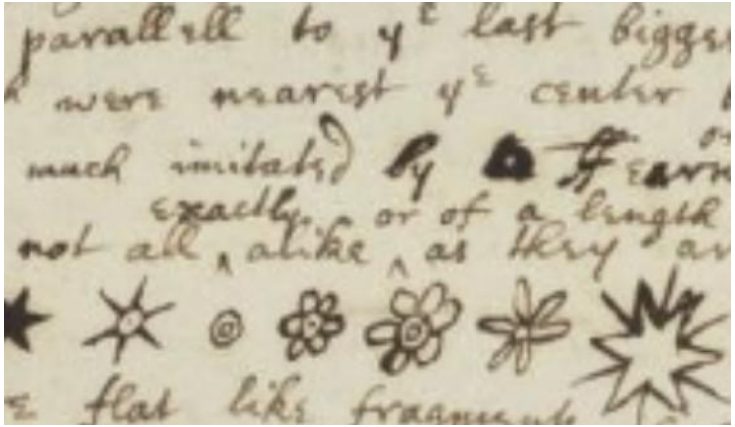
Text structures

Codex Justinianus, 1475 in Mainz gedruckt von Peter Schöffer

Erste Seite des Traktats Bĕrachot der Talmud-Ausgabe von Daniel Bomberg (In der Mitte : Text der Mischna und Gĕmara, rechts: der Kommentar Raschis, links: die Glossen der Tossafisten). Gedruckt 1520 in Venedig; Anordnung in allen folgenden Ausgaben beibehalten.

Manual Transcription

- Old manuscripts are often not written in a completely linear way (margins, crossed-out sections, etc.)
- What to do with non-ASCII characters, non-local characters and symbols that are not Unicode?
- How to avoid intentional or unintentional corrections, i.e. how to ensure faithful transcriptions especially of historical material?
- Manual transcription is extremely time-consuming and therefore very expensive.



Manuscript by Isaac Newton

Optical Character Recognition (OCR)

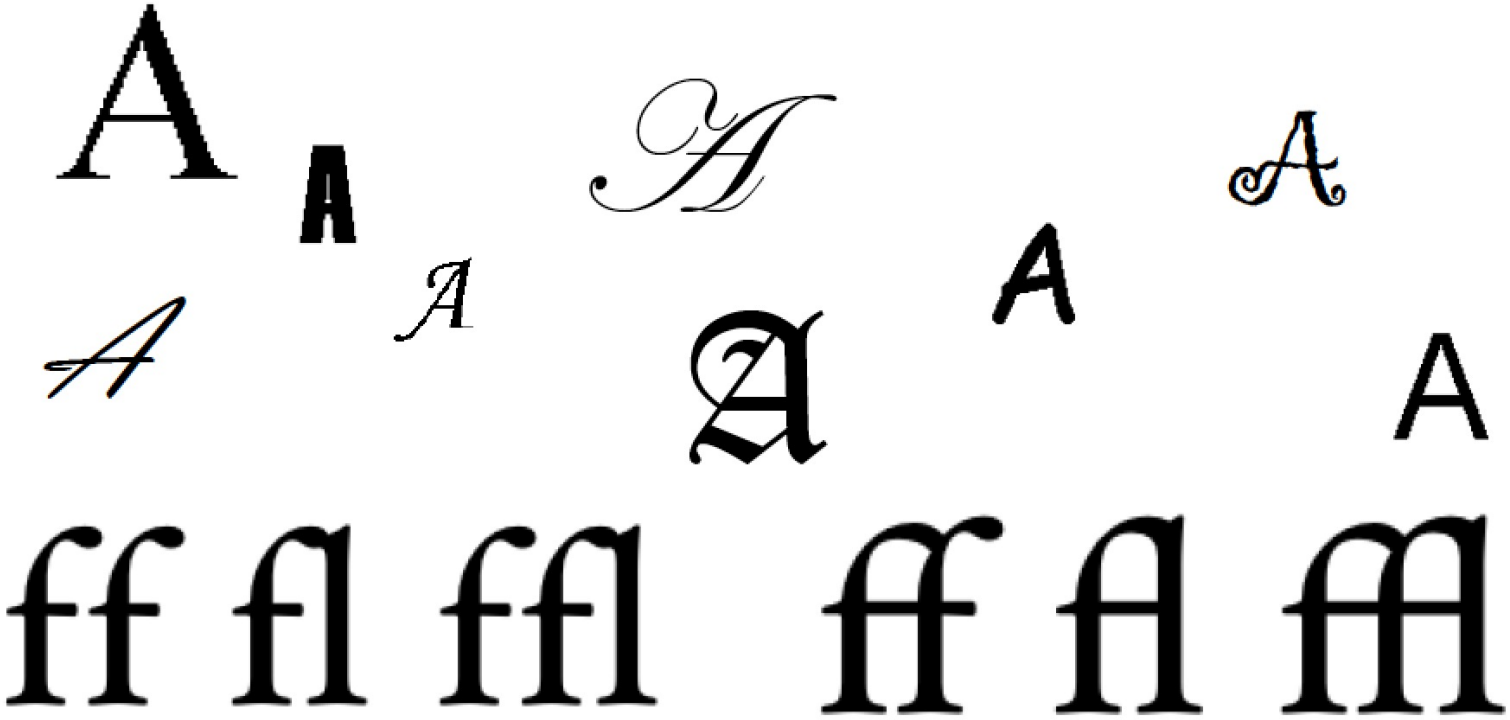
- Software for analysing digital images of text and recognising the characters / letters in the text.
often included with scanners
- Algorithm typically combines two types of information:
– Knowledge of the visual characteristics of different characters, i.e. the letter shape
– Knowledge of typical character sequences in a particular language (= language model)

Optical Layout Recognition

aims to acquire the layout of a page, e.g.

- Recognition of tables, illustrations and photos in the text
- Recognition of captions
- Detection of the number of columns in the text

Optical Character Recognition (OCR)



ohne Ligaturen

mit Ligaturen

Optical Character Recognition (OCR)

Informationsdienst der Christlich Demokratischen Union Deutschlands

1/85

Union in Deutschland

Bonn, den 10. Januar 1985

Heiner Geißler:

1985 – ein Jahr der Offensive

1985 ist für die CDU ein Jahr des programmatischen Fortschritts. Wir haben 1984 bei der Verabschiedung der Stuttgarter Leitsätze gezeigt, daß wir auch als Regierungspartei nicht programmatisch stehen bleiben, sondern als dynamische Kraft die Politik der Bundesregierung tragen, anregen und begleiten, erklärte Generalsekretär Heiner Geißler in Bonn.

In seiner Sitzung am 10. Dezember 1984 hat der Bundesvorstand entschieden, daß auf dem 33. Bundesparteitag in Essen (20. bis 22. März 1985) die Frauenpolitik Schwerpunktthema des Parteitages sein wird. Wie bereits bei dem Jugendparteitag 1981 in Hamburg wird auch der 33. Bundesparteitag ein offener Parteitag sein. Die CDU wird ca. 400 Frauen aus den Bereichen der Gesellschaft (Einrichtungen, Verbände, Fraueninitiativen, Frauenhäuser usw.) als sachverständige Diskussionsteilnehmer einladen. Ziel des Bundesparteitages ist es, vom Bundesvorstand eingebrachte Leitsätze zur Frauenpolitik zu verabschieden.

Themenschwerpunkte werden u. a. sein:

- Wandel der Lebensperspektiven und Lebensplanung von Frauen,
- Frauen in Beruf und Familie,
- Frauen im Berufsleben,
- Frauen in Politik und Gesellschaft.

In dieser Ausgabe lesen Sie:

Die Küche und der Herd.

Die Küche dient einer tüchtigen Hausfrau, welche sich selbst um die Zubereitung der Speisen bekümmert, viele Stunden des Tages als Aufenthaltsort. Auch die Dienstboten, denen in den meisten Fällen kein eigenes geräumiges Zimmer zur Verfügung steht, müssen sich dort den ganzen Tag über aufhalten; und wir können sicher sein, daß sie ihre Arbeit in einer möglichst freundlichen Küche mit größerer Bereitwilligkeit verrichten, als in einem dunkeln, unfreundlichen Raume.

The OCRed Text (1)

Informationsdienst der Christlich Demokratischen Union Deutschlands
Union in Deutschland
Bonn, den 10. Januar 1985
Heiner Geißler:
1985 - ein Jahr der Offensive
1985 ist für die CDU ein Jahr des programmati- schen Fortschritts. Wir haben 1984 bei der Verabschiedung der Stuttgarter Leitsätze ge- zeigt, daß wir auch als Regierungspartei nicht programmatisch stehen bleiben, sondern als dynamische Kraft die Politik der Bundesregie- rung tragen, anregen und begleiten, erklärte Generalsekretär Heiner Geißler in Bonn.
In seiner Sitzung am 10. Dezember 1984 hat der Bun- desvorstand entschieden, daß auf dem 33. Bundes- Parteitag in Essen (20. bis 22. März 1985) die Frauen- Politik Schwerpunktthema des Parteitages sein wird. Wie bereits bei dem Jugendparteitag 1981 in Ham- burg wird auch der 33. Bundesparteitag ein offener Parteitag sein. Die CDU wird ca. 400 Frauen aus den Bereichen der Gesellschaft (Einrichtungen, Verbän- de, Fraueninitiativen, Frauenhäuser usw.) als sach- verständige Diskussionsteilnehmer einladen. Ziel des Bundesparteitages ist es, vom Bundesvorstand ei- 9e brachte Leitsätze zur Frauenpolitik zu verabschie- den.

Unicode

International Standard for encoding as many sense-bearing characters as possible

Different assignment procedures:

- UTF (Unicode Transformation Format), UTF-8 (Standard), UTF-16
- ASCII (128 characters), no umlauts like ä, ö, ü
- ISO 8859-1 (256 characters) umlauts, no Greek (only with change of code)
- Unicode (theoretically over 1 million characters) : Greek, Cyrillic, Han, Thai, etc. Part of newer operating systems; currently approx. 90,000 characters defined.

OCR mistakes

have a negative effect on acquisition and hinder further processing:

- often influence word segmentation (-> tokenisation)
- omit dots (-> sentence segmentation)
- deviate more often from correctly spelled words than from typos
- letter-to-letter mappings are less often one-to-one (e.g. "m" -> "iii")
- are more likely to affect several places in a word
- can be systematically wrong -> methods for correcting spelling errors often do not work well

OCR mistakes

Even if the text recognition is correct, post-processing may be necessary to remove artefacts of the printed text that may affect the readability by the machine:

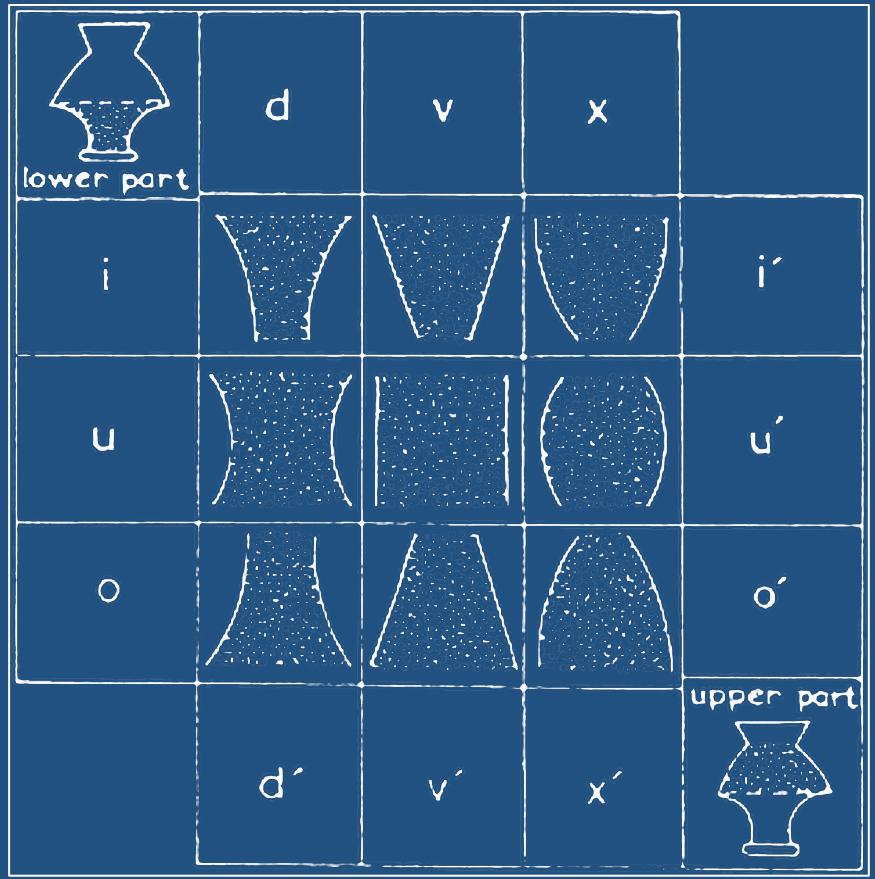
- word separation at the end of the line (e.g. "sun- days") -> not always trivial, e.g. for German spelling (e.g. "Schiff-fahrt"-> "Schiffahrt")
- Page numbers that can occur in the middle of sentences and even words and make further processing difficult (e.g. "recogni- Page 7 tion") -> The pagination information should be removed or stored separately from the plain text

Improving text recognition

Three possibilities:

- Improving the input quality (quality of digital images)
- Improvement of the OCR algorithm
- Improving the output by post-processing the OCR (= pre-processing before applying further NLP tools)

2. "EXPRESSING": FORMALISING THE RESEARCH OBJECTS



Formalisation

Augustus de Morgan (1851), *Briefe*

Statistical mean value of word lengths (measured in syllables) in comparison of the Pauline epistles

D. Holmes, The Evolution of Stylometry in Humanities Scholarship, Literary and Linguistic Computing 13/3, 1998, 111–117

Figure 4 shows how the combinations of the six terms, under two different

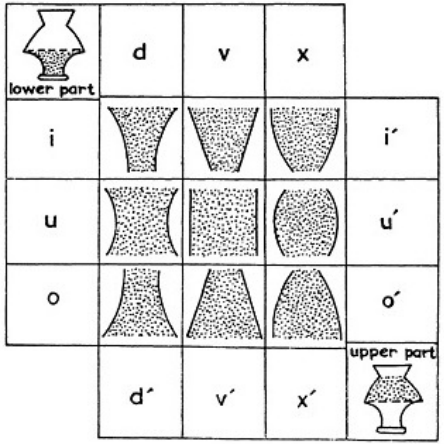


FIGURE 4. Pottery, main body: 9 elementary profiles.

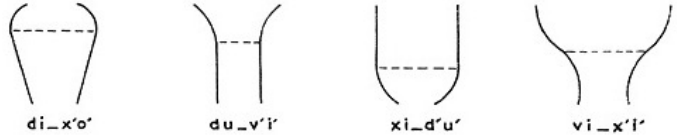


FIGURE 5. Pottery, main body: examples of code words for different shapes.

Jean-Claude Gardin, On the Coding of Geometrical Shapes And Other Representations, with Reference to Archaeological Documents, 1959

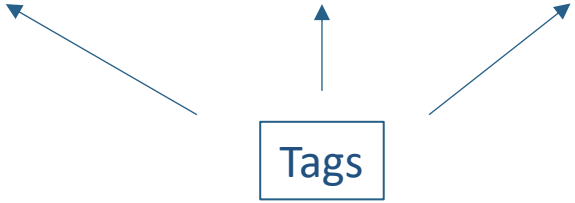
Text- Layout

– Direct typing

```
This is bold.
```

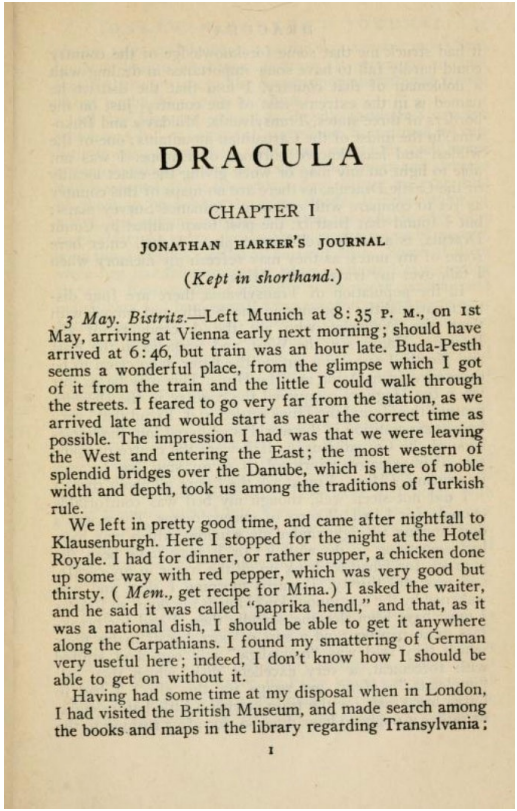
– Digital Markup

```
<p>This is <b>bold.</b>
```



Text – Logical Structure

```
<div1 type="work">
  <head>Dracula</head>
  <div2 type=chapter><head>Chapter 1</head>
    <p> Jonathan Harker's Journal<p>
    <p> (Kept in shorthand)<p>
    <div3 type="journal"><p><i>3 May.
      Bistritz</i> — Left Munich at 8:35 p.m.
          [...]
      </p>
          [...]
    </div3>
  </div2>
</div1>
```



DRACULA

CHAPTER I

JONATHAN HARKER'S JOURNAL

(*Kept in shorthand.*)

3 May. Bistritz.—Left Munich at 8:35 P. M., on 1st May, arriving at Vienna early next morning; should have arrived at 6:46, but train was an hour late. Buda-Pesth seems a wonderful place, from the glimpse which I got of it from the train and the little I could walk through the streets. I feared to go very far from the station, as we arrived late and would start as near the correct time as possible. The impression I had was that we were leaving the West and entering the East; the most western of splendid bridges over the Danube, which is here of noble width and depth, took us among the traditions of Turkish rule.

We left in pretty good time, and came after nightfall to Klausenburgh. Here I stopped for the night at the Hotel Royale. I had for dinner, or rather supper, a chicken done up some way with red pepper, which was very good but thirsty. (*Mem.*, get recipe for Mina.) I asked the waiter, and he said it was called "paprika hendl," and that, as it was a national dish, I should be able to get it anywhere along the Carpathians. I found my smattering of German very useful here; indeed, I don't know how I should be able to get on without it.

Having had some time at my disposal when in London, I had visited the British Museum, and made search among the books and maps in the library regarding Transylvania;

1

Text – Logical Structure

Chapter 1. Jonathan Harker's Journal

- In the text:

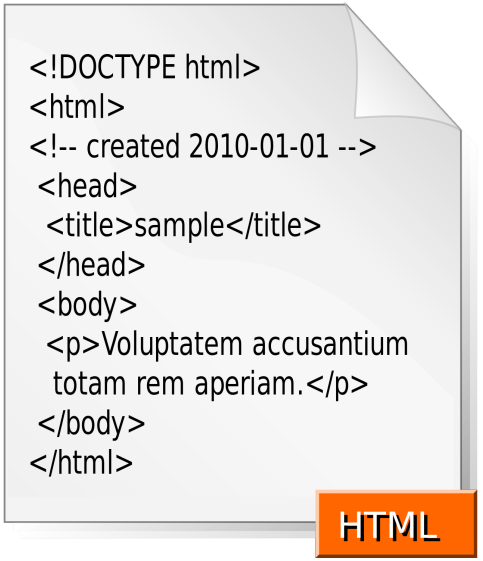
```
<h1 class=„Chapter“> Chapter 1. Jonathan Harker's Journal</h1>
```

- In stylesheet

```
.Chapter { font-family:Blackadder,sans-serif; font-size:24pt;
color:blue; }
```

Hypertext Markup Language (HTML)

HTML serves as a markup language to structure a text semantically, but not to format it. The visual presentation is not part of the HTML specifications and is determined by the web browser and design templates such as CSS.



Advantages

- Procedural (HTML) vs. descriptive (Word) markup
- Machine-readable markup vs. machine languages
- Structuring of the entire document

Disadvantages

- Its own system: property of a company
- Not platform independent

XML – eXtensible Markup Language

is a simple, highly flexible text format derived from the Standard Generalised Markup Language SGML (ISO 8879). Originally developed to meet the challenges of large-format electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the web and elsewhere.

- Meta-language for markup languages used to define data.
- Defines the overall structure of textual data according to document type definitions
- Search and Point of Speech (POS) is possible
- W3C standard, relies on SGML like HTML

http://www.w3.org/XML/

XML

Tag

```
<heading>1. Kapitel</heading>
```

Blank tags

```
<newLine/>
```

Attributs

```
<div type="chapter">
<div type="chapter" nr="1">
```

XML

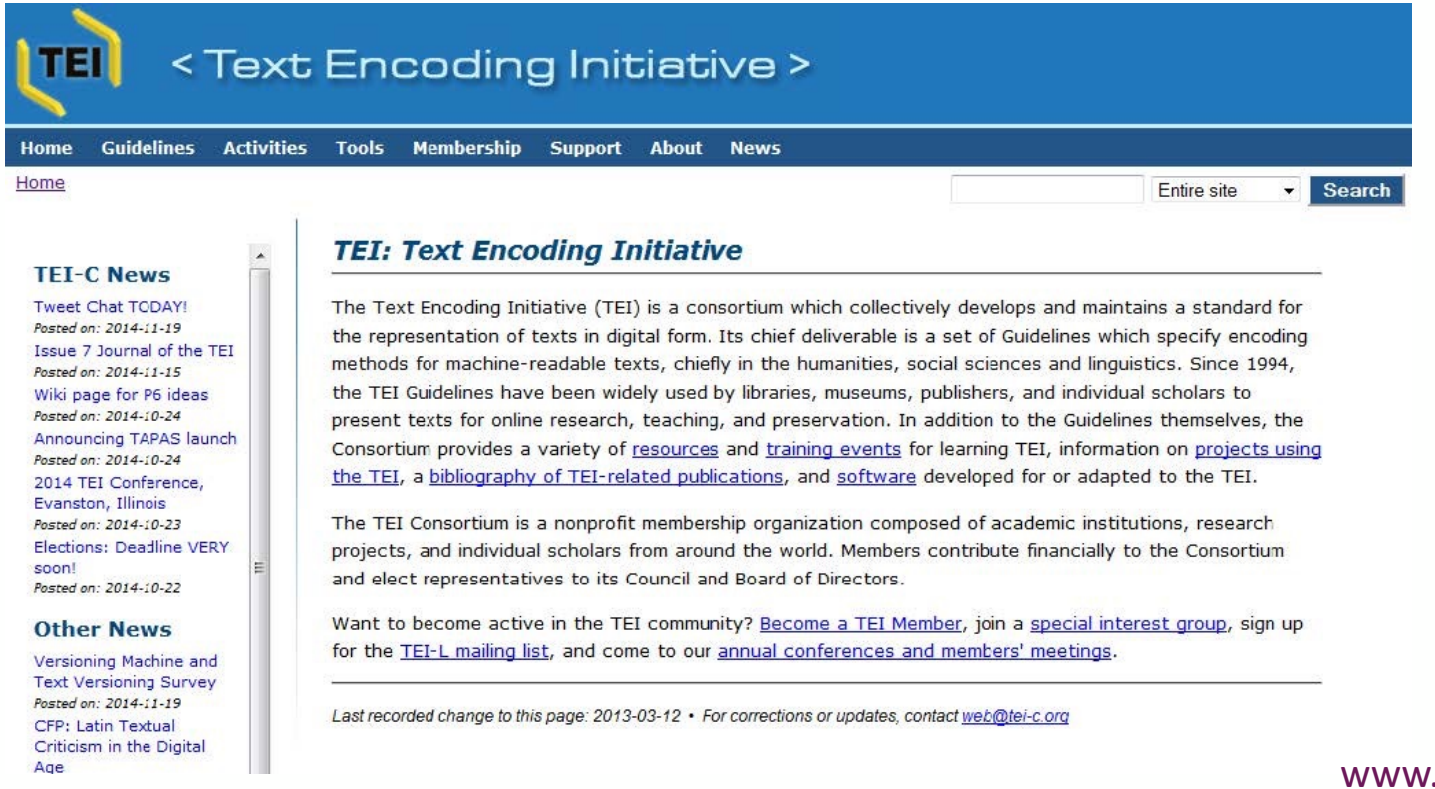
Commentary

```
<!—Is the markup correct here? -->
```

Structure of XML documents

```
<person>
  <firstname>Alan</firstname>
  <lastname>Turing</lastname>
  <profession>Computer Scientist</profession>
  <profession>Mathematician</profession>
  <profession>Cryptograph</profession>
</person>
```

Text Encoding Initiative (TEI)



www.tei-c.org

TEI

ca. 450 elements or tags are defined and categorised into 3 classes:

- Core Tag Set
  for every TEI document
- Base Tag Sets
  choose only one Base Tag Set
- Additional Tag Sets
  as much as you like

https://de.dariah.eu/tei-tutorial

TEI – Core Tag Set (Beispiel)

Foreign language
John eats a <foreign lang="fr">croissant</foreign> every morning.

Mentioned
<mentioned lang="fr">Croissant</mentioned> is difficult to pronounce with your mouth full.

Term
A <term lang="fr">croissant</term> is a crescent-shaped piece of light, buttery, pastry that is usually eaten for breakfast, especially in France.

TEI – Base Tag Sets

- TEI.prose: markup for prose
- TEI.verse: for poetry
- TEI.drama: for drama
- TEI.spoken: for spoken language
- TEI.dictionaries: for dictionaries
- TEI.terminology: for terms
- TEI.mixed: for documents of different types and genres
- TEI.general: as TEI.mixed, but only one base tag set

TEI – Drama

```
<sp>
<speaker>Erwin</speaker>
<l>Elmire, </l>
<stage>(he leaps forth)</stage>
</sp>
<sp>
<speaker>Elmire</speaker>
<l>Woe is me!</l>
</sp>
<sp>
<speaker>Erwin</speaker>
<stage>(at her feet)</stage>
<l>It's me.</l>
</sp>
```

TEI – Verse

Lines, verses, stanza…

```
<div1 Type=poem><head>God’s plan</head>
  <lg Type=stanza>
  <l n=1>God’s plan made a hopeful beginning.</l>
  <l n=2>But man spoiled his chances by sinning.</l>
  <l n=3>We trust that the story</l>
  <l n=4>Will end in God’s glory,</l>
  <l n=5>But at present the other side’s winning</l>
  </lg>
```

TEI – Transcription of sources

Elements for

- deletion,
- insert,
- replacement,
- Missing words usw.

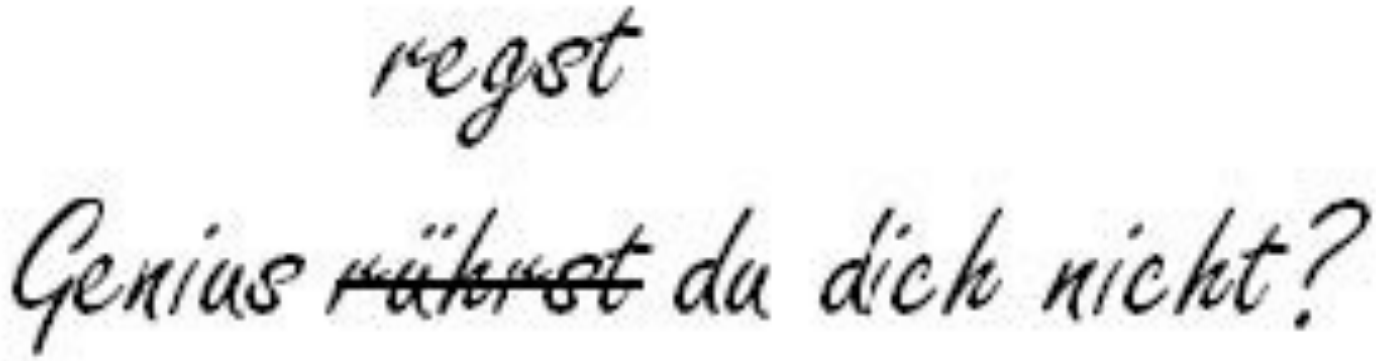
TEI – deletion <del>


```
<p>Er konnte es nicht
<del type="immediate correction"
rend="overstrike" resp="FJ"
hand="JWG">fas</del>glauben!</p>
```

TEI – addition <add>


```
<l>[...] Genius <add
type="supralinear" hand="JWG"
resp="FJ">regst</add> du dich nicht?</l>
```

TEI – correction <corr>




```
<l>[...] Genius <corr sic="rührst">regst</corr>
du dich nicht?</l>
```

TEI – DTD to the document

- Prolog
  - 1. What type of DTD
  - 2. What kind of tag sets
  - 3. Which code for the letters
- Header
  - Author, content, editor etc.
- Text
  - The text itself

TEI – Prolog – Tagset

```
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main Document Type//EN"
  "tei2.dtd" [
    <!ENTITY % TEI.XML 'INCLUDE' >

<!-- TEI base tag set specified here: ... -->
    <!ENTITY % TEI.prose 'INCLUDE' >

<!-- TEI additional tag sets optionally specified here: ... -->
    <!ENTITY % TEI.textcrit 'INCLUDE' >
    <!ENTITY % TEI.linking 'INCLUDE' >
]>
```

Metadata

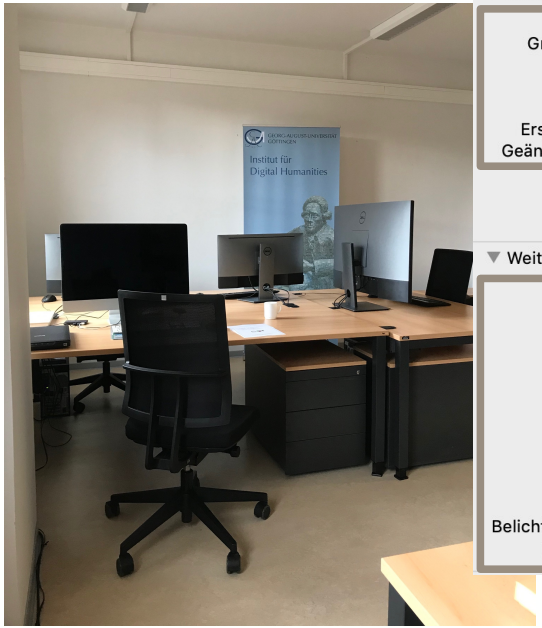
- Standardised description of metadata for documents
- Dublin Core Metadata Initiative (DCMI) since 1994

„The Dublin Core Metadata Initiative, or ‘DCMI’, is an open organization engaged in the development of interoperable metadata standards that support a broad range of purposes and business models.” (http://dublincore.org)

Metadata

Class

Property



Infos zu: IMG_1403.JPG

IMG_1403.JPG 2,2 MB
Geändert: 10. Januar 2020 um 13:34

Tags ...

▼ Allgemein:

Art:	JPEG-Bild
Größe:	2.239.704 Byte (2,2 MB auf dem Volume)
Ort:	Macintosh HD ▸ Benutzer ▸ mlangne ▸ Bilder ▸ Raum 1.213
Erstellt:	10. Januar 2020 um 13:34
Geändert:	10. Januar 2020 um 13:34

☐ Formularblock
☐ Gesperrt

▼ Weitere Informationen:

Zul. geöffnet:	26.04.2020, 10:11
Bildgröße:	3024×4032
Gerätemarke:	Apple
Gerätemodell:	iPhone 7
Farbraum:	RGB
Farbprofil:	Display P3
Brennweite:	3,99 mm
Alpha-Kanal:	Nein
Rote Augen:	Nein
Messmethode:	Muster
Blendenzahl:	f/1,8
Belichtungsprogramm:	Normal
Belichtungszeit:	1/17

Value

Infos zu: 5V_1024.JPG

15V_1024.JPG 3,7 MB
Geändert: Freitag, 7. Oktober 2016 um 14:10

Tags ...

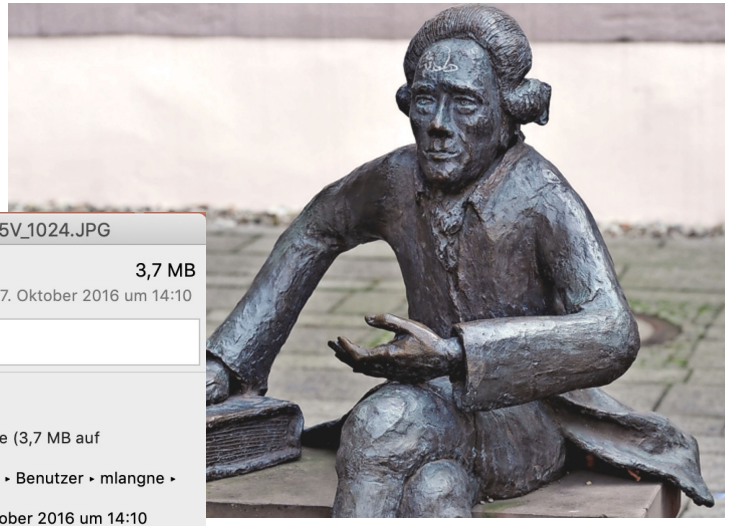
▼ Allgemein:

Art:	JPEG-Bild
Größe:	3.703.725 Byte (3,7 MB auf dem Volume)
Ort:	Macintosh HD ▸ Benutzer ▸ mlangne ▸ Institut
Erstellt:	Freitag, 7. Oktober 2016 um 14:10
Geändert:	Freitag, 7. Oktober 2016 um 14:10

☐ Formularblock
☐ Gesperrt

▼ Weitere Informationen:

Zul. geöffnet:	17. März 2018 um 19:43
Bildgröße:	3264×4928
Gerätemarke:	NIKON CORPORATION
Gerätemodell:	NIKON D7000
Farbraum:	RGB
Farbprofil:	sRGB IEC61966-2.1
Brennweite:	35 mm
Alpha-Kanal:	Nein
Rote Augen:	Nein
Messmethode:	Muster
Blendenzahl:	f/2,5
Belichtungsprogramm:	0
Belichtungszeit:	1/25



Metadata

- describe objects in a **structured** and **uniform** or **standardised** form.
- serve to **select** (search/find/select) and **identify** resources or describe their appropriate use.
- can describe **anything**: Literature, paintings, films, people, fossils, clothing, places, ...
- can be associated with objects (i.e. **embedded** in them) or provided **separately** (in their own records)

Types of Metadata

- **Descriptive Metadata**
  Information needed to search, find and identify relevant objects, such as title, author, date of publication, etc.
- **Administrative Metadata**
  Information on origin, archiving, access rights and other things that serve to manage the objects, such as licence, producer, rights holder etc.).
- **Technical Metadata**
  Information necessary for the appropriate use of the resource, such as file format, image resolution, etc.)
- **Structure Metadata**
  Information about the composition of a resource (e.g. whether a digitised book consists of many individual files) and the relationship of the parts to each other.
- **Linkage/Relationship Metadata**
  Information about the relationships that exist between objects (higher-level entireties, other versions, etc.)
- **Content Ratings Metadata**
  Information about possible users of objects (the target audience)
- **Meta-Metadata**
  which models, syntax and formats underlie the metadata, who created it and when, etc.

Metadaten – Dublin Core (DCMES)

Dublin Core® Metadata Initiative
Making it easier to find information.

- International standard for describing all types of resources
- Origin: OCLC/NCSA Metadata Workshop, March 1995, Dublin/Ohio
- Maintenance and further development by Dublin Core Metadata Initiative (DCMI)
- Goals
  - Simplicity of semantics and application
  - Provide a basis for semantic interoperability

Dublin Core Metadata Element Set (DC Simple): ISO 15836

Contributor	Coverage	Creator	Date
Description	Format	Identifier	Language
Publisher	Relation	Rights	Source
Subject	Title	Type	

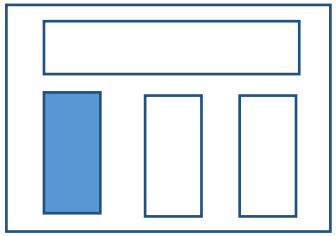
Example

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

Réponse n° 1

Domaine	peinture
Type d'objet	tableau
Titre	PORTRAIT DE MONA LISA (1479-1528) ; DITE LA JOCONDE
Auteur/exécutant	LEONARDO DI SER PIERO DA VINCI ; VINCI Léonard de (dit)
Précision auteur/exécutant	Vinci, 1452 ; Amboise, 1519
Ecole	Italie
Période création/exécution	1er quart 16e siècle
Millésime création/exécution	1503 entre ; 1506 et
Genèse	oeuvre en rapport ; reproduit en gravure
Historique	commandé par le florentin Francesco del Giocondo, époux de Mona Lisa entre 1503 et 1506 ; nombreuses copies dont une conservée au Louvre ; gravé par Fauchery, par Filhol, par Landon
Matériaux/techniques	peinture à l'huile ; bois
Mesures	77 H ; 53 L
Sujet représenté	portrait (Mona Lisa, femme, à mi-corps, de trois-quarts, assis, accoudé, loggia, Italien) ; fond de paysage (montagne, rocher, cours d'eau, pont, plaine, route)
Date sujet représenté	1479-1528
Lieu de conservation	Paris ; musée du Louvre département des Peintures
	Musée de France *au sens de la loi n°2002-5 du 4 janvier 2002*
Statut juridique	propriété de l'Etat ; musée du Louvre département des Peintures
Anciennes appartenances	François Ier ; Couronne de France

Digital Editions



Layout

Sequence of alphanumeric characters

Encoding

Active references within the text/image, to other texts and other media

Hyperlinks

Title, heading, (chapter, paragraph, verse line, act), (colour, shape and size), hyperlinks.

Structure

Document
Sequence of Bytes

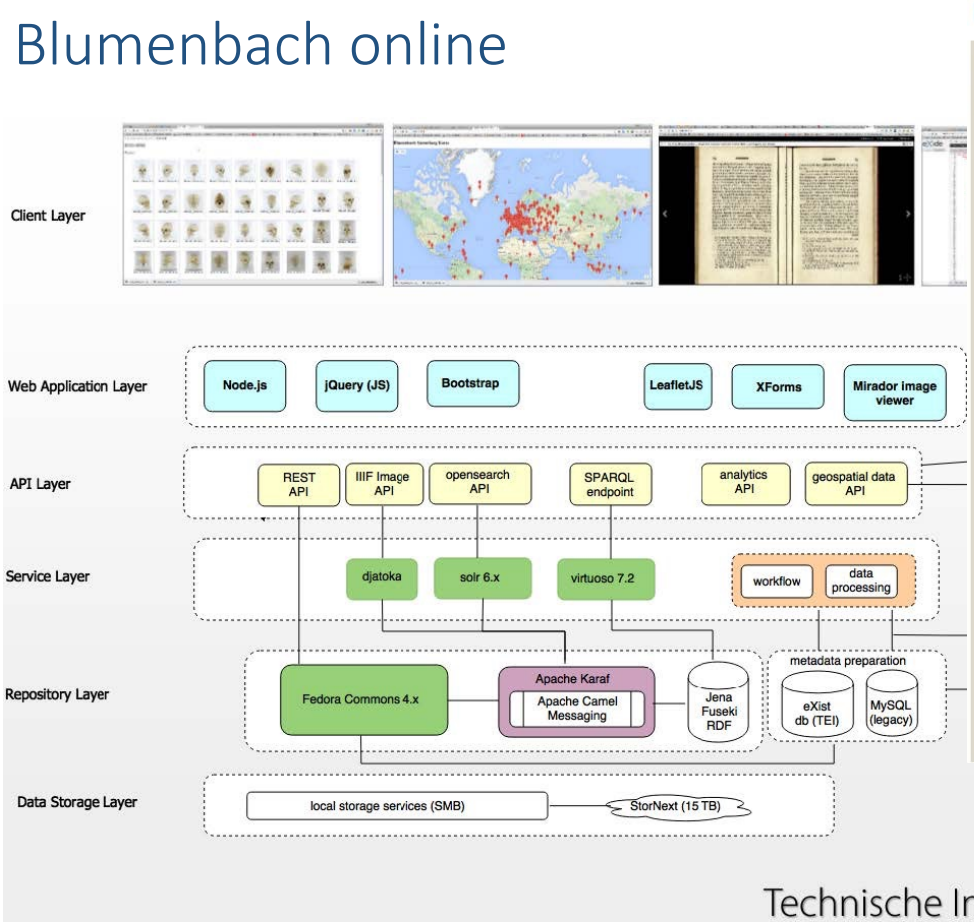
Genre, Concepts, Fictional Worlds etc.

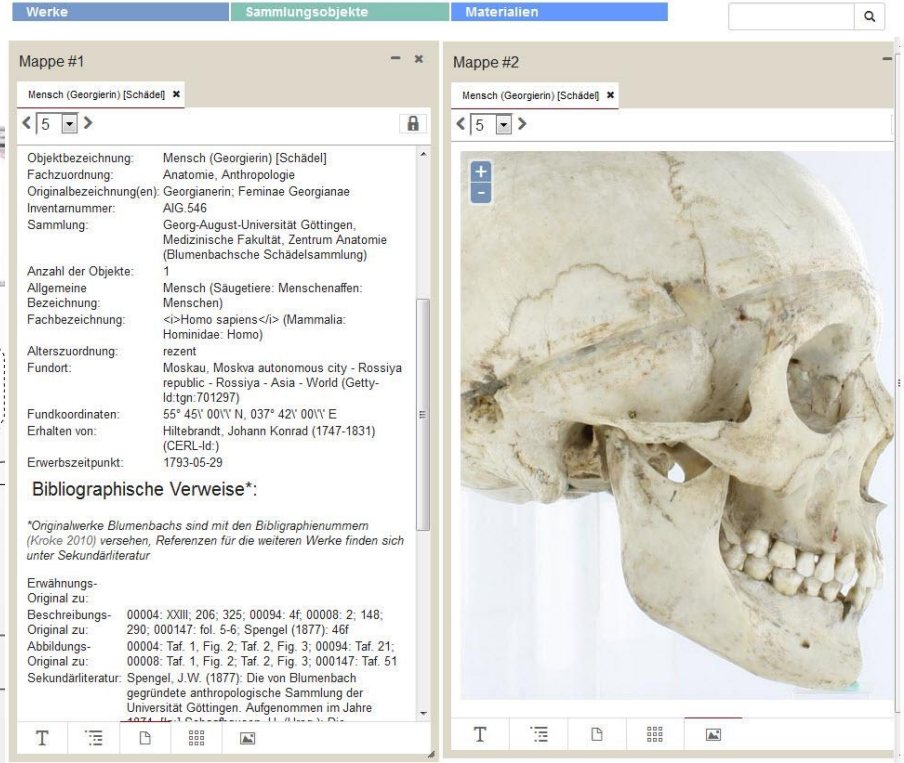
Content

Metadata

author/ artist, date of creation, access rights, last change, place and time of publication

Blumenbach online

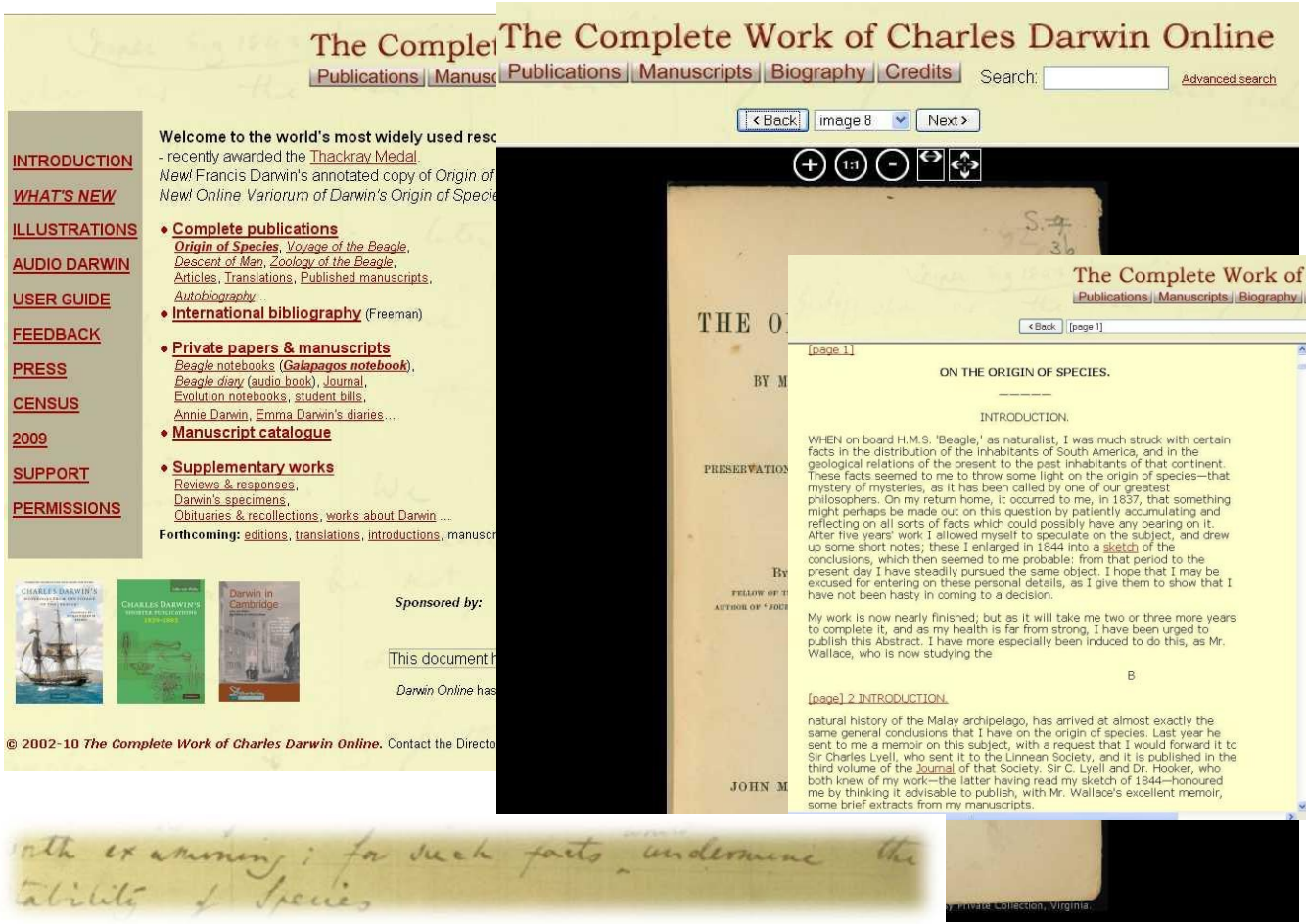




Technische Infrastruktur

Darwin online

http://darwin-online.org.uk/



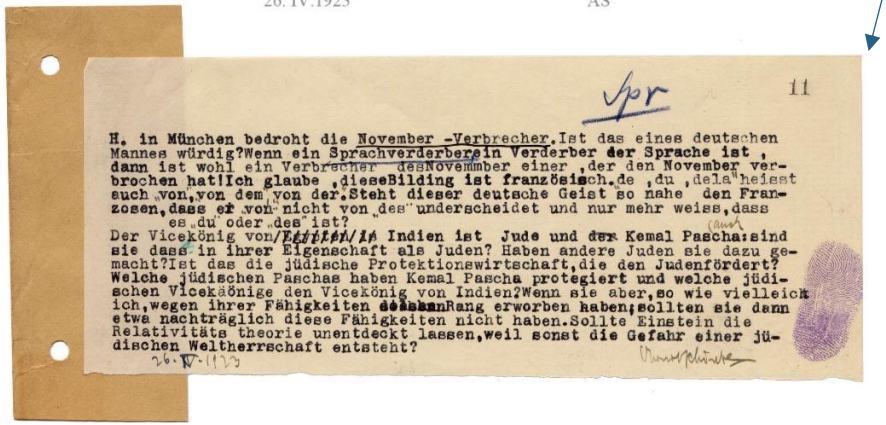
Schoenberg.at

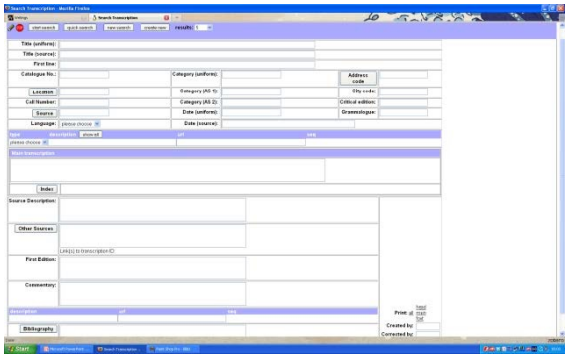
Spr 11

H. in München bedroht die November-Verbrecher. Ist das eines deutschen Mannes würdig? Wenn ein Sprachverderber ein Verderber der Sprache ist, dann ist wohl ein Verbrecher desNovemmber einer, der den November ver= brochen hat! Ich glaube, dieseBilding ist französisch. „de, du, dela“[1] heisst auch „von“, „von dem“, „von der“. Steht dieser deutsche Geist so nahe den Fran= zosen, dass er „von“ nicht von „des“ underscheidet und nur mehr weiss, dass es „du“ oder „des“ ist?

Der Vicekönig von ~~Egypten is~~ Indien ist Jude und ~~der~~ Kemal Pascha ⟨auch⟩: sind sie dass[2] in ihrer Eigenschaft als Juden? Haben andere Juden sie dazu ge= macht? Ist das die jüdische Protektionswirtschaft, die den Judenfördert? Welche jüdischen Paschas haben Kemal Pascha protegiert und welche jüdi= schen Vicekäönige den Vicekönig von Indien? Wenn sie aber, so wie vielleicht ich, wegen ihrer Fähigkeiten /diesen→/solchen Rang erworben haben; sollten sie dann etwa nachträglich diese Fähigkeiten nicht haben. Sollte Einstein die Relativitäts theorie unentdeckt lassen, weil sonst die Gefahr einer jü= dischen Weltherrschaft entsteht?

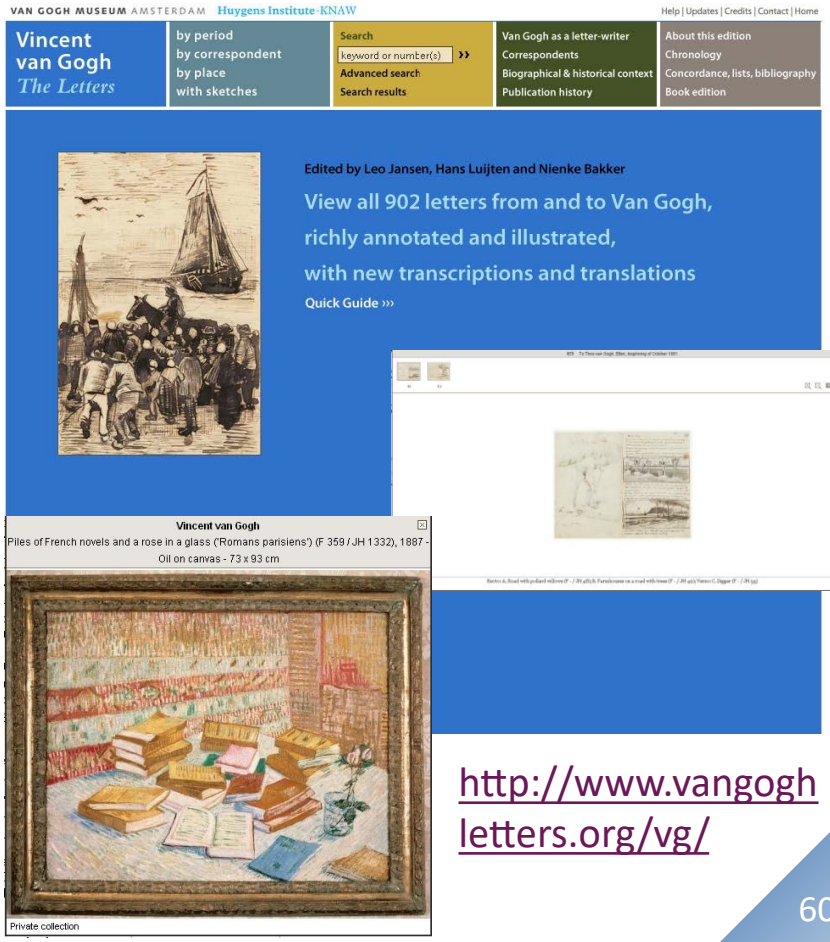
26. IV.1923 AS[3]





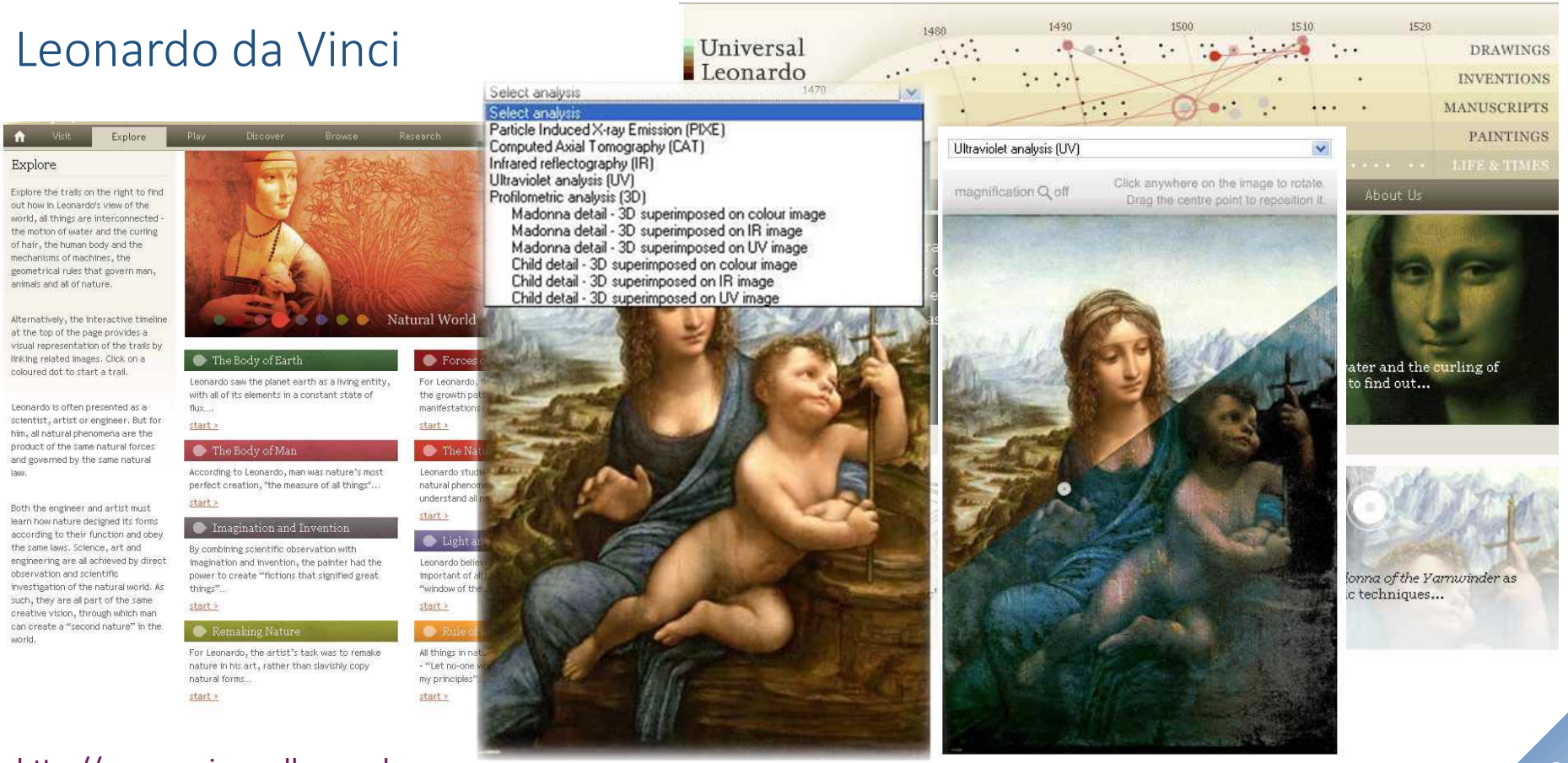
Vincent van Gogh

- Index of original works (by title, by type of work [paintings, X-rays and photographs of paintings, works on paper and sketches]).
- Index of works by other artists,
- Index of sketches by Gaugin
- Index of photographs, documents, journals and literary works
- Index of biblical quotations (divided into relevant books and sections)
- Index of persons in general and by correspondent (the majority of letters are addressed to van Gogh's brother Theo)
- Index of places.



http://www.vangoghletters.org/vg/

Leonardo da Vinci



http://www.universalleonardo.org

van Gogh

van Gogh ist der Familienname folgender Personen:

- Johanna van Gogh-Bonger (1862–1925), niederländische Kunstsammlerin
- Johannes van Gogh (1817–1885), niederländischer Vizeadmiral und Offizier des Militär-Wilhelms-Ordens
- Lothar van Gogh (1888–1945), niederländischer Fußballspieler
- Niels van Gogh (* 1977), deutscher DJ und Musiker
- Theo van Gogh (Kunsthändler) (1857–1891), niederländischer Kunsthändler und -sammler
- Theo van Gogh (Regisseur) (1957–2004), niederländischer Regisseur
- Vincent van Gogh (1853–1890), niederländischer Maler
- Vincent Willem van Gogh (1890–1978), niederländischer Ingenieur und Mäzen

Van Gogh bezeichnet:

- (4457) van Gogh, ein nach Vincent van Gogh benannter Asteroid
- Van Gogh (Band), eine jugoslawische/serbische Rockband
- Van Gogh (1948), ein französischer Kurzfilm von Alain Resnais aus dem Jahr 1948
- Van Gogh (1991), ein französischer Spielfilm von Maurice Pialat aus dem Jahr 1991

Dies ist eine **Begriffsklärungsseite** zur Unterscheidung mehrerer mit demselben Wort bezeichneter Begriffe.

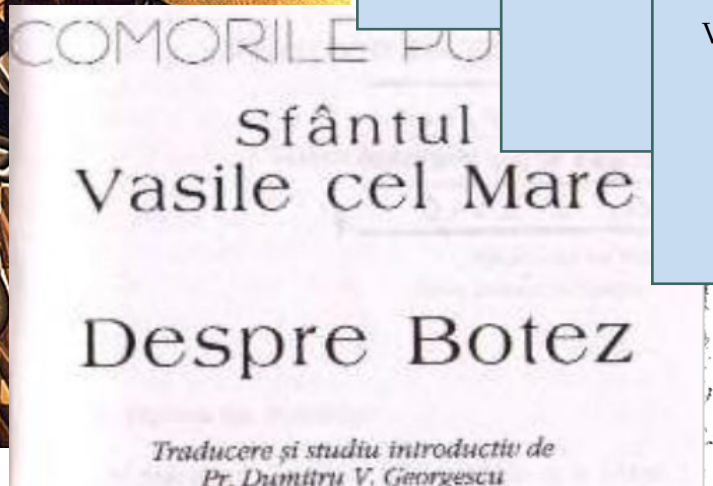
Authority control

Basilius von Caesarea (ca. 330 – 379)

Basilius der Große
s. Basilius <Caesariensis>

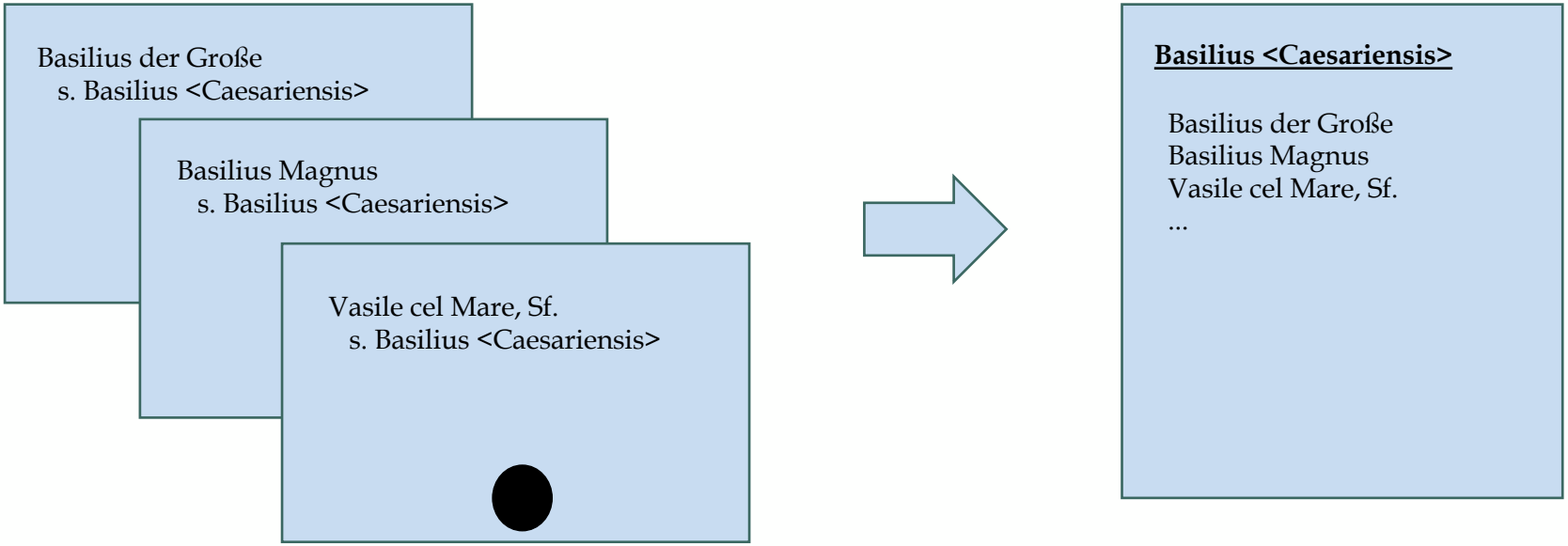
Basilius Magnus
s. Basilius <Caesariensis>

Vasile cel Mare, Sf.
s. Basilius <Caesariensis>



Authority control

The totality of references to the same target form an equivalence class, which can be summarised in a data set called authority record.



Controlled Vocabularies

provide uniqueness for the reference name.

They concretise the value range (Range) of a metadata element (Properties) by listing the permitted values.

This avoids

- synonymous names for identical values
- homonymous names for different values

Authority control: Individualisierung

Each entity should be described **unambiguously** by an authority record in such a way that **confusion** is **excluded**.

Individualisation features may be:

- life and existence data,
- geographical coordinates,
- superordinate terms or entities,
- professions,
- titles of nobility,
- places of activity
- relations to other entities

Types of authority records

In use of libraries

- persons
- corporate bodies,
- keywords,
- classification(s),
- works or unit titles,
- printers/publishers,
- generic terms,
- geographies,
- provenance characteristics

The use of authority records is conceivable wherever one has to deal with different designations for the same entity.

Common authority file of the German-speaking world (Gemeinsame Normdatei: GND)

integrates since April 2012

- Personennamendatei (PND)
- Gemeinsame Körperschaftsdatei (GKD)
- Schlagwortnormdatei (SWD)
- and some local authority records.

coordinated by the German National Library (DNB)
edited by the larger academic libraries in Germany and Austria

Used by libraries, heritage institutions, various DH projects and others (e.g. https://de.wikipedia.org/wiki/Gemeinsame_Normdatei).
Size > 10 million records.

International Standard Name Identifier (ISNI)

International standard since 2012 (ISO 27729)
16-digit number
supported by a consortium consisting of the
- CENL (Conference of European National Librarians)
- OCLC (Online Computer Library Center, Dublin/Ohio)
- ProQuest
- CISAC (Confédération Internationale des Sociétés d'Auteurs et Compositeurs)
- IFFRO (International Federation of Reproduction Rights Organisations)
- IPDA (International Performers Database Association)

currently > 8 Mio assigned identifiers
Subset: ORCID (Open Researcher and Contributor ID)

Thesaurus

a *controlled vocabulary with (hierarchical) relations*
standardised as ISO 2788 / DIN 1463-1

- Equivalence relations
  - **USE** / **UF** (used synonym/used for)
  - multilingual variants
  - descriptors and non-descriptors
- Hierarchical relations
  - **BT** / **NT** / **TT** (broader term/narrower term/technical term)
  - superordinate
  - subordinate
- Associative relations
  - **RT** (related term)
  - like „see also … “)

Thesaurus

Example:
Eurovoc

rail transport

Non-preferred terms

UF ↖ *rail connection*
UF ↖ *rail traffic*
UF ↖ *railway*
UF ↖ *transport by railway*

Broader Terms

BT ↑ land transport

More specific terms

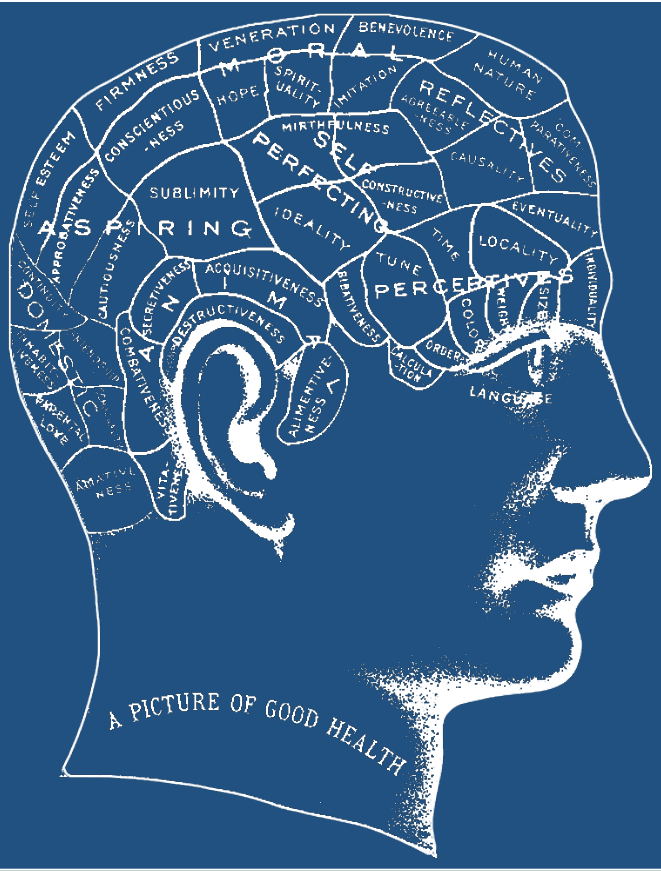
NT4 ↓ CIV Convention
NT4 ↓ rail network ▶
NT4 ↓ rolling stock
NT4 ↓ vehicle on rails

Related terms

RT ⇆ air-cushion vehicle
RT ⇆ European Railway Agency
RT ⇆ high-speed transport
RT ⇆ railway industry
RT ⇆ railway tariff
RT ⇆ transport staff

Bulgare железопътен транспорт
Czech železniční doprava
German Schienentransport
Greek σιδηροδρομικές μεταφορές
Español transporte ferroviario
Français transport ferroviaire
Italien trasporto ferroviario
Dutch vervoer per spoor
Polish transport kolejowy
Portugais transporte ferroviário
Swedish järnvägstransport

https://www.vocabularyserver.com/eurovoc



3. "UNDERSTANDING":

SEMANTICS OF THE RESEARCH OBJECTS

Ontologies

controlled vocabularies with their own domain model

describe
- – Classes of terms
- – Relationships of these classes to each other
- – instances of these classes

and thus allows conclusions to be drawn about the instances (inference).

Description language:
RDF Scheme (RDFS) for simpler domain ontologies
Web Ontology Language (OWL) for complex knowledge representation

Resource Description Framework (RDF)

- elementary statements about things in the form:

  Subject - Predicate - Object (so-called triple).

- Subject and predicate are always a resource (document, entity or property) designated by a URI.

- Object can be a resource or a literal (character string).

- If the object is a literal, statements can be made about
- the data type (e.g. date, numerical value, text, etc.)
- The language

RDF Schema

Classes
- rdfs:Class
- rdfs:Property
- rdfs:Literal
- rdfs:Resource

Eigenschaften
- rdfs:range
- rdfs:domain
- rdfs:type
- rdfs:subClassOf
- rdfs:subProperyOf
- rdfs:label
- rdfs:comment

Semantic Web

„traditional“ Web

- Linked documents
- Documents are identified by URIs (URLs)
- Links are not standardised
- IdR coded in HTML/XHTML
- Document content usually only understandable for humans
- Full text search

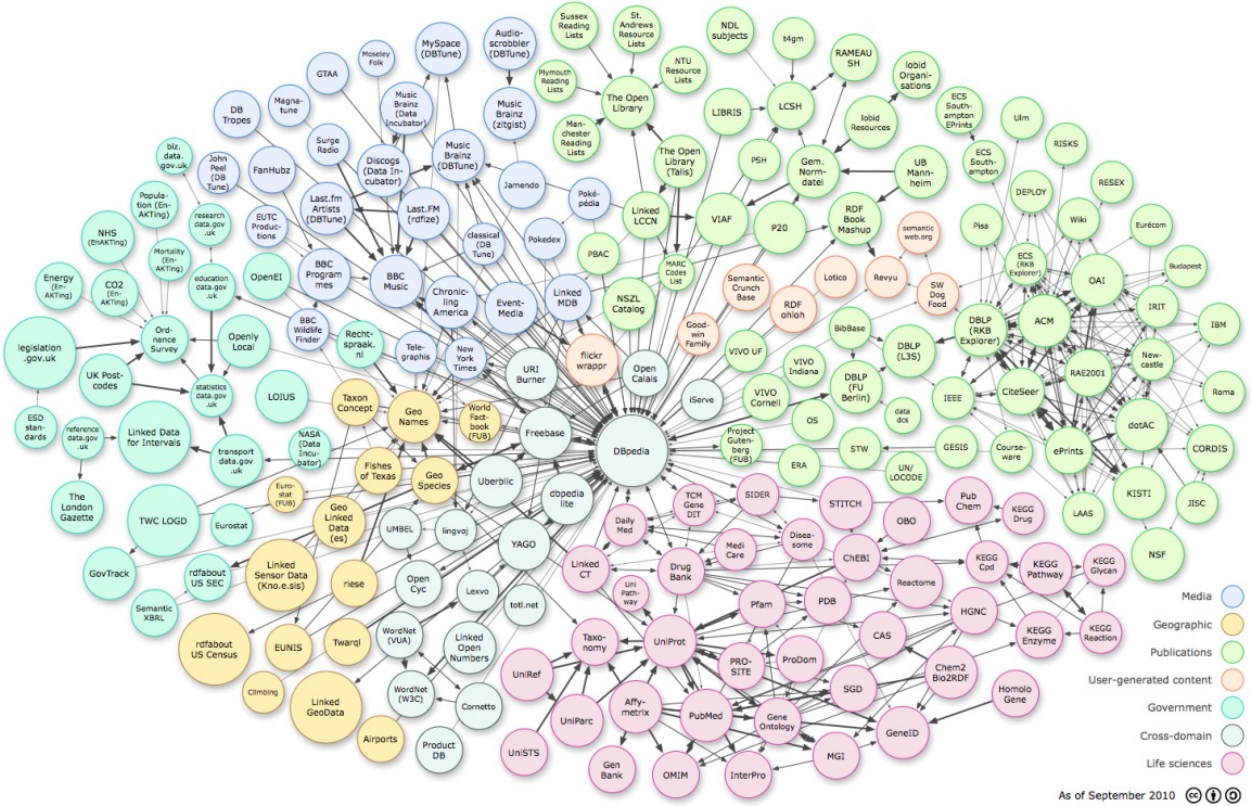
Semantic Web

- linked information (data)
- Entities are designated by URIs (URLs)
- Relationships between entities are also denoted by URIs
- Coding in RDF (mostly RDF/XML)
- Statements about entities are machine-understandable
- Semantic search possible (SPARQL)

Linked Open Data

Linking of different data sets through the use of common identifiers (URIs) for entities.

Licensing of the data under a licence that allows uncomplicated subsequent use of the data by others



http://lod-cloud.net/

Linked Open Data: Design Rules

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.

Tim Berners-Lee, 2006. s. http://www.w3.org/DesignIssues/LinkedData.html

Semantic Search

- Triple-Store:

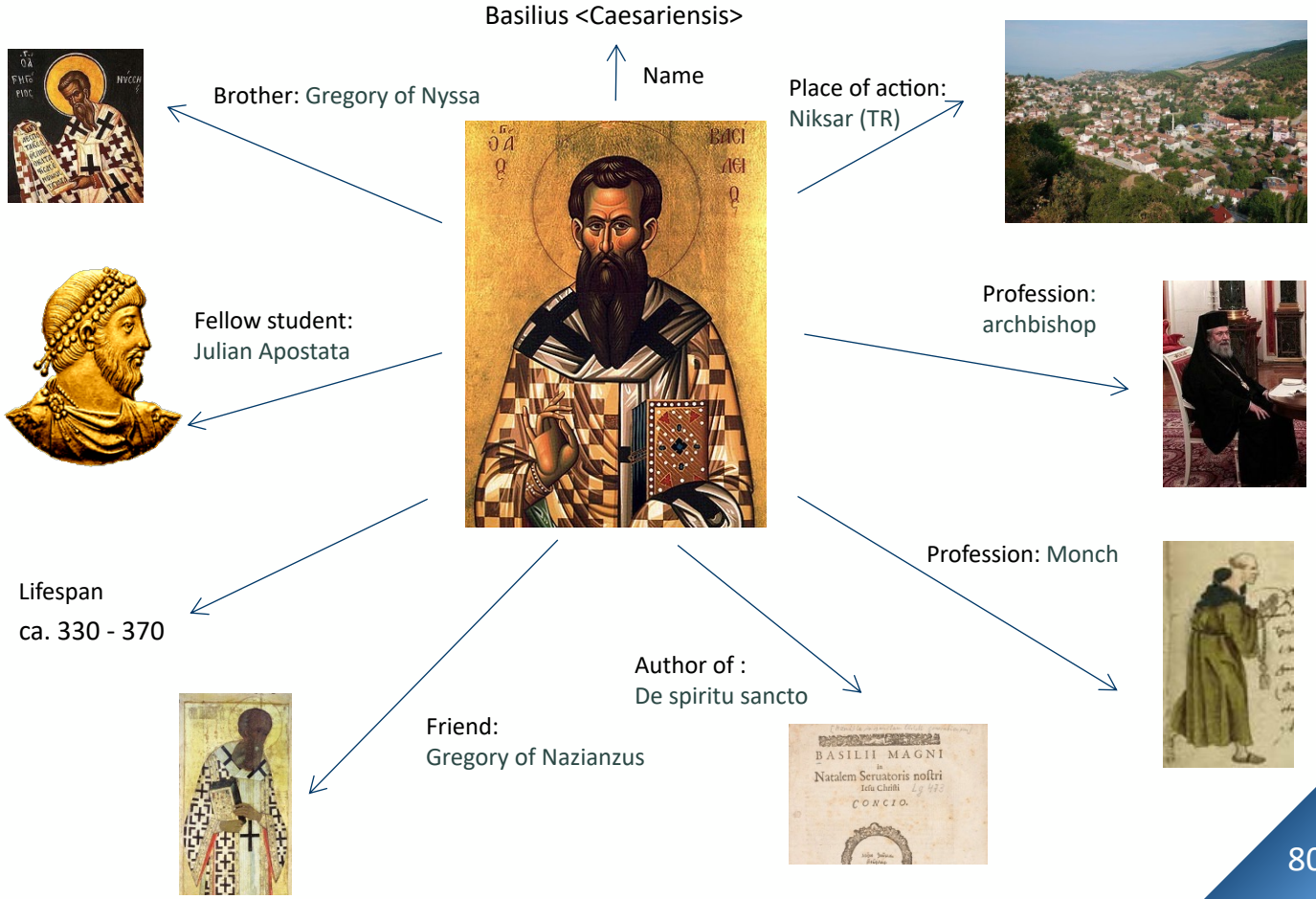
  Database application optimised to handle RDF triples

- SPARQL:

  SPARQL **P**rotocol **A**nd **R**DF **Q**uery **L**anguage = Graph-based query language for RDF.

```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
     abc:isCapitalOf ?y .
  ?y abc:countryname ?country ;
     abc:isInContinent abc:Africa .
}
```

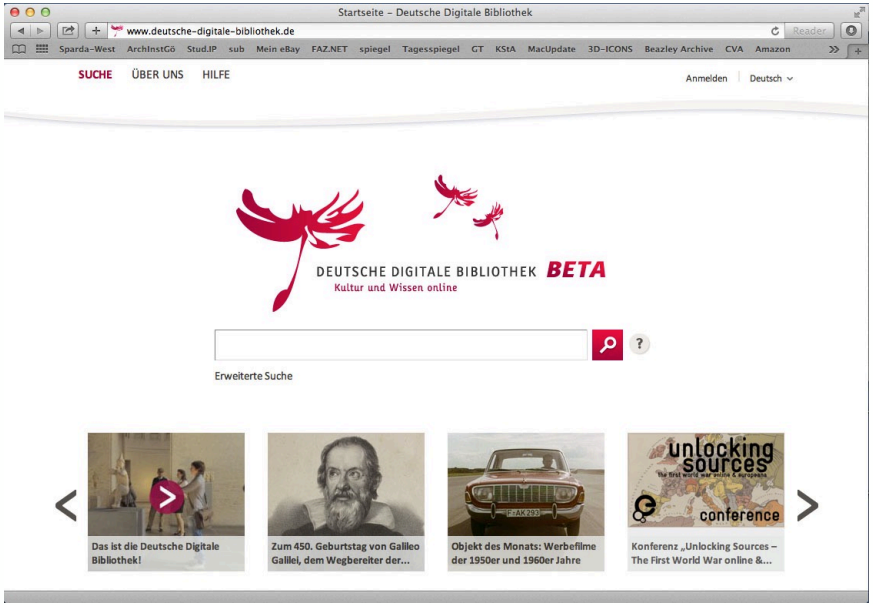
Authority records as information nodes



Ontologies: CIDOC CRM

The CIDOC Conceptual Reference Model (CRM) provides an extensible ontology for concepts and information in the field of cultural heritage and museum documentation. As an international standard (ISO 21127:2014) for the controlled exchange of cultural heritage information, it is used by libraries, archives, museums and other cultural institutions to improve access to museum-related information and knowledge.

Event orientation



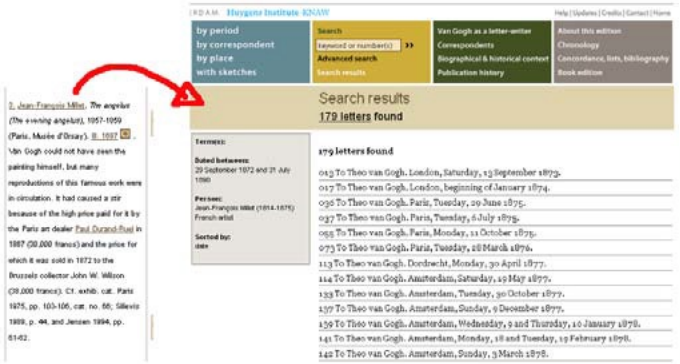
www.deutsche-digitale-bibliothek.de

see:
http://terminology.lido-schema.org/eventType

Event types according to CIDOC CRM Entitiy E5:

- Execution
- Execution
- Excavation
- Exhibition
- Processing
- Extension
- Find
- Use (primary / secondary function)
- Spiritual creation (original)
- Manufacture
- Idea (design / shaping)
- Planning (design)
- Publication
- Reproduction
- Collecting event
- Part removal
- Type definition (?)
- Type assignment (?)
- Transformation
- Completion

Digital Editions and Corpora



Acquisition with common text programmes

Conversion to XML

along enriched TEI guidelines
(http://www.kantl.be/ctb/project/dalf/)

Linking to a database (open source: Ruby, Lucene)

Output in XHTML, image viewer

http://www.digitalhumanities.org/dhq/vol/4/2/000088/000088.html

CHALLENGES OF DIGITALISATION

- Materiality and sensory quality of the objects of investigation ("experience")
- Historical conditionality and perspectivity of data models ("expressions")
- Thesis formation on incomplete, fuzzy and heterogeneous data ("understanding")

- Properties of the presentation of texts – Desktop Publishing (font type & size, typesetting, layout, integration of media etc.)
- Separation of outline / structure, content / information, presentation / formatting
- Relevant procedures and good practices for documenting digital (research) data
- Relevant metadata standards (e. g. ACDM, ADEX, CARARE, CIDOC-CRM, DC, LIDO, EDM, METS, MIDAS)
- Syntax examples (z. B. (X)HTML, XML, TEI, KML, Markdown etc.)
- relevance of controlled vocabularies / authority records

- selecting appropriate text formats and mark-up languages
- making use of basic metadata to exchange and describe text and image sources
- assessing the consequences of a chosen documentation method on the re-usability of the data

What is an algorithm? What is the hermeneutic method? What are the fundamental differences between the two concepts?

Slides 4–10

What is an ontology? Give an internationally used example and its application!

Slides 73. 81

Which peculiarities of the humanities (= problems of computer science) are particularly relevant for the Digital Humanities?

Slide 84

Which metadata standards do you know for marking up texts and images?

Slide 54

How do sources become data, and what is the process of creating digital editions?

Slide 13. 56. 83

What problems does automated text recognition cause?

Slides 26–28

Gerald Futschek, „Algorithmic thinking: The key for understanding computer science“, in: *Lecture Notes in Computer Science 4226* (2006), 159–168

Julia Flanders and Fotis Jannidis, „Data Modeling in a Digital Humanities Context“, in: dies. (ed.): *The Shape of Data in Digital Humanities. Modeling Texts and Text-based Resources* (London 2017).

Helmut Vonhoegen, *Handbuch: Einstieg in XML: Grundlagen, Praxis, Referenz* (Bonn ^{8}2015).

Tim Berners-Lee, James Hendler, and Ora Lassila, „The Semantic Web“, *Scientific American* 284/5 (2005), 34–45.

Martin de la Iglesia, Nicolas Moretto, and Maximilian Brodhun, „Metadaten, LOD und der Mehrwert standardisierter und vernetzter Daten“, in: Heike Neuroth, Andrea Rapp, and Sibylle Söring (eds.), *TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften* (Göttingen 2015), 91–102. DOI: http://dx.doi.org/10.3249/webdoc-3947 (20.04.2020).

Heike Zinsmeister, „Chancen und Grenzen von automatischer Annotation“. Themenheft zu Automatisierte Textanalyse für Sozial- und Kulturwissenschaften, *Zeitschrift für Germanistische Linguistik* 43/1 (2015), 84–110.

IMPRESSUM

Inhaltlich verantwortlich:

Prof. Dr. Martin Langner
Institut für Digital Humanities
Georg-August-Universität Göttingen
37073 Göttingen
Martin.Langner@uni-goettingen.de
Telefon 0551 / 3926790
https://www.uni-goettingen.de/digitalhumanities

Nutzungsbedingungen:

Bitte informieren Sie uns, wenn Sie der Ansicht sind, dass ein hier veröffentlichter Scan eine Urheberrechtsverletzung darstellt; wir werden diesen Fall umgehend lösen.

Haftungshinweis:

Trotz sorgfältiger inhaltlicher Kontrolle übernehmen wir keine Haftung für die Inhalte externer Links. Für den Inhalt der verlinkten Seiten sind ausschließlich deren Betreiber verantwortlich.

ABBILDUNGSNACHWEIS

Folie 2: http://dobes.mpi.nl/
Folie 3:
Folie 4: https://upload.wikimedia.org/wikipedia/commons/d/df/Dilthey1-4.jpg
Folie 5:
https://upload.wikimedia.org/wikipedia/commons/b/b8/Wilhelm_Windelband.jpg
Folie 6:
https://upload.wikimedia.org/wikipedia/commons/7/79/Alan_Turing_az_1930-as_%C3%A9vekben.jpg
Folie 7. 8:
https://upload.wikimedia.org/wikipedia/commons/thumb/f/f2/Induktion-Deduktion.svg/2560px-Induktion-Deduktion.svg.png
Folie 12: https://www.pexels.com/photo/bunch-of-photo-print-1908066/
Folie 13:
https://upload.wikimedia.org/wikipedia/commons/e/e7/Beethoven_Klaviersonate_Nr_30.jpg
Folie 16: Digitalisierungszentrum SUB Göttingen
Folie 19:
Folie 30:
https://upload.wikimedia.org/wikipedia/commons/2/2c/De_Morgan_Augustus.jpg

Die Rechte aller anderen Abbildungen liegen beim Autor, Institut für Digital Humanities der Georg-August-Universität Göttingen.