

# OCR at the Bayerische Staatsbibliothek: projects and experiences

based on the German Powerpoint presentation

by

Dr. Markus Brantl,

Dr. Karl Märker

# Agenda

---

- 1) State of the art
- 2) Strategy and history
- 3) Project examples
- 4) future perspectives

# State of production

---

- Freely accessible books (titles) online:  
**1.023.875 (ca. 383.953.125 pages)**
- Among them titles with OCR:  
**900.000**
- PDF downloads of entire books (titles), image files  
per day (2013): **2.300**

# Strategy

---

- Pragmatically oriented approach, driven by user demands or experiences
- First option: machine-produced OCR without any intervention
- Second option: outsourcing – multiple OCR, keyboarding, corrections for projects with a need for a high level of text accuracy

The MDZ

Digital Collections

Highlights

Service

## Explore the collections



### Subjects

Our digitized works by the classic shelf number classification of the Bavarian State Library



### Authors

Our digitized works alphabetically by author names



### Time

Our digitized works sorted by centuries and year by year



### Place of publication

Our digitized works by place of publication



### Publishing houses

Our digitized works by their publishing or printing houses



### Image Similarity

Similarity of motifs by colors, textures, distinctive shapes and contrasts



### Latest Additions

Recently digitized works, by online publication date

## Search the collections



Data basis: 1.049.045 Digitized Works

## What's new

Latest Additions to our Digital Collections >>

Updated daily: [Latest additions](#) to the Digital Collections of the Bavarian State Library. Items online: 1,049,045

[23.10.2014]



Close enough to touch: The Golden Evangeliary of Mainz at the Library of the Aschaffenburg Castle >>

From October 24 to November 21, 2014, the [Library of the Aschaffenburg Castle](#) is presenting the [Golden Evangeliary of Mainz](#). The "BSB Explorer" provided to the exhibition by the MDZ, allows full access in high resolution to the digitized version of this precious manuscript of the 13th century.

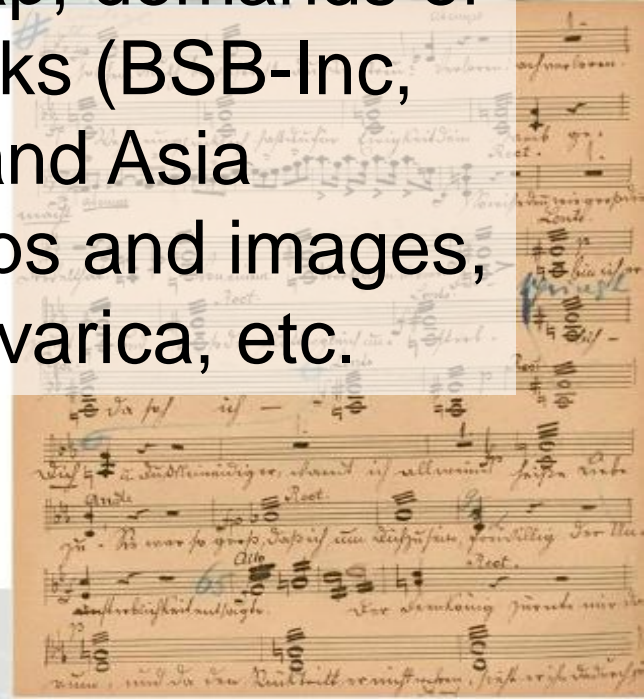
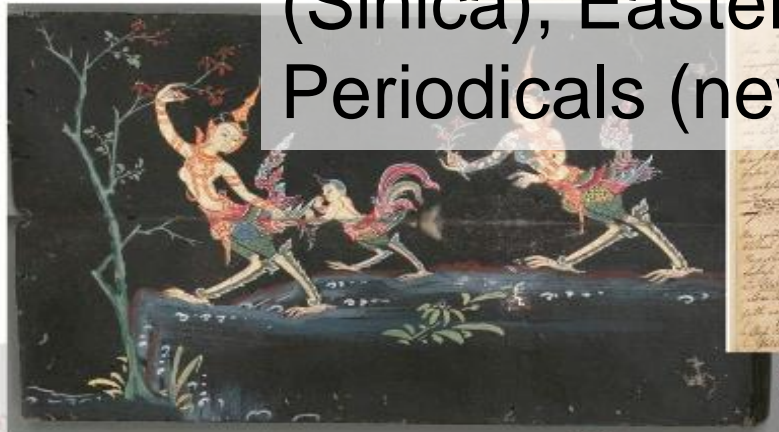
[15.10.2014]



Other recent news items >>

# Content providers

Inhouse departments:  
Manuscripts (national roadmap, demands of full text, T-Pen) and Rare books (BSB-Inc, VD16, VD18), Music, Orient and Asia (Sinica), Eastern Europe, Maps and images, Periodicals (newspapers), Bavarica, etc.



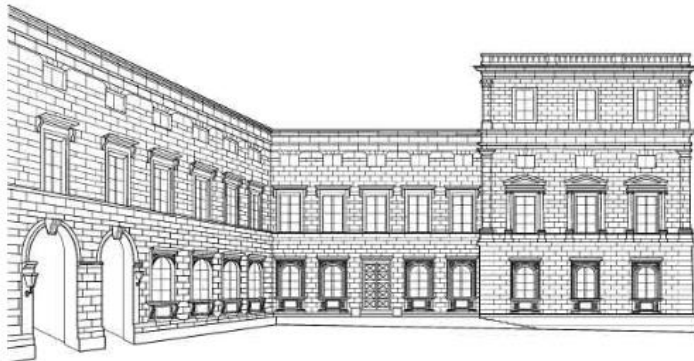
# Content providers



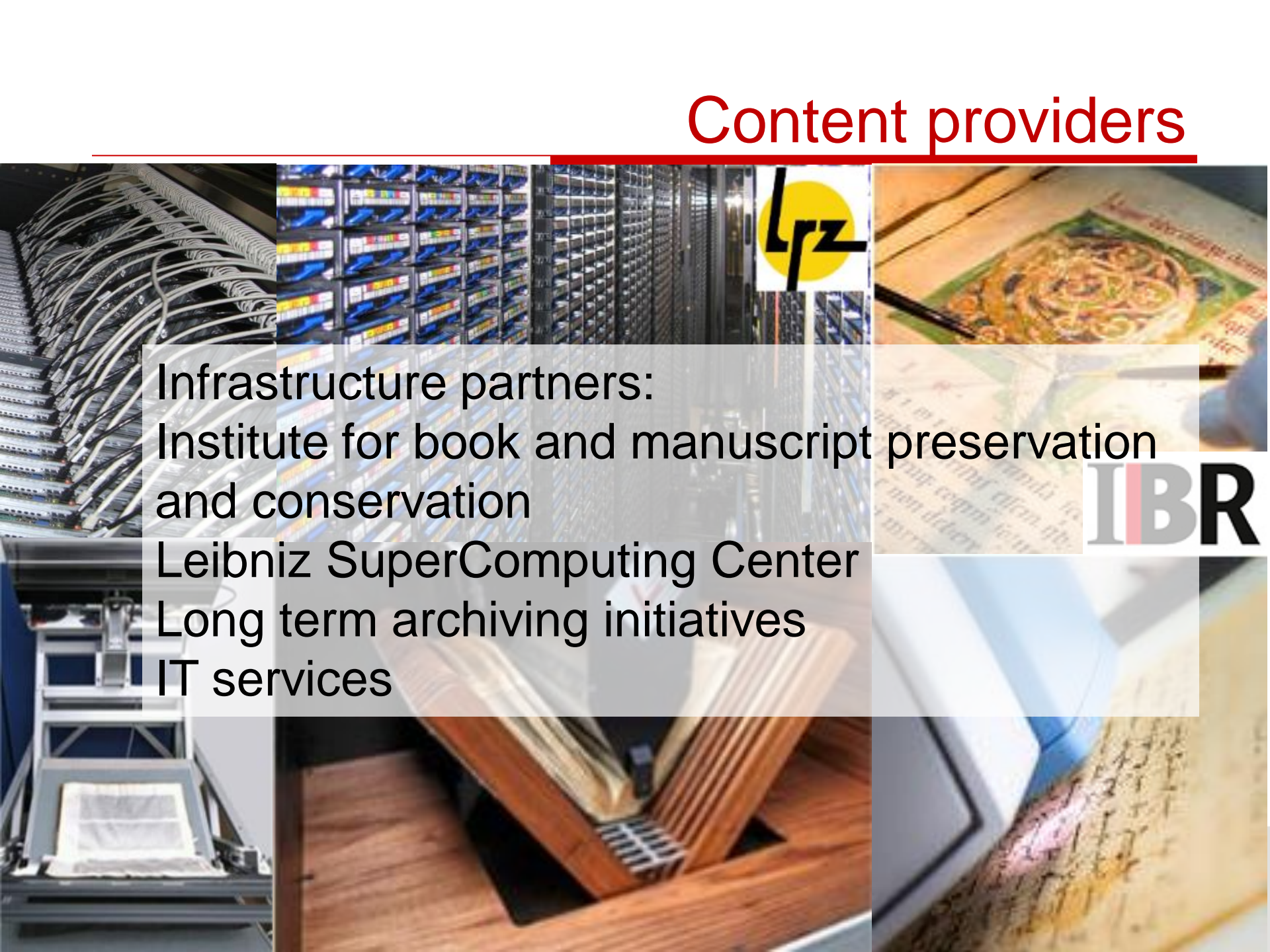
Cooperation partners:  
Monumenta Germaniae Historica,  
Bavarian Academy of Sciences ...

Tendit ad aequum

Sie strebt nach dem Angemessenen



# Content providers



Infrastructure partners:  
Institute for book and manuscript preservation  
and conservation  
Leibniz SuperComputing Center  
Long term archiving initiatives  
IT services



# Munich Digitization Center (MDZ): innovative projects with different partners



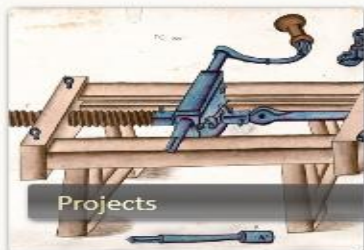
The MDZ



Digital Collections



Highlights



Projects



Explore



Service



Mobile Apps



Long-term preservation

## Welcome to the Munich Digitization Center

Munich Digitization Center (MDZ) handles the [digitization](#) and online publication of the cultural heritage preserved by the Bavarian State Library and by other institutions. It provides one of the largest and fastest growing digital collections in Germany, now comprising more than 900,000 titles available online. Access is free of charge!

The digitization policy reflects the traditional special collection fields of the library: History, Classical Antiquity, Eastern Europe, Musicology. It comprises manuscripts, early prints, modern books, maps and photographic collections as well as journals and newspapers.

The MDZ frames the *Digital Library department* of the Bavarian State Library. It provides the [Digital Collections](#) and is also responsible for the creation of subject portals and for the [long-term preservation](#) of all the library's digital holdings.

### Search the collections

Data basis: 1.049.045 Digitized Works

### What's new

Latest Additions to our Digital Collections >>

Updated daily: [Latest additions](#) to the Digital Collections of the Bavarian State Library. Items online: **1,049,045** [23.10.2014]



Close enough to touch: The Golden Evangeliary of Mainz at the Library of the Aschaffenburg Castle >>

From October 24 to November 21, 2014, the Library of the Aschaffenburg Castle is presenting the Golden Evangeliary of Mainz. The "BSB Explorer" provided to the exhibition by the MDZ, allows full access in high resolution to the digitized version of this precious manuscript of the 13th century. [15.10.2014]



Other recent news items >>



# History 1

---

- Keyboarding and/or correction - special projects since 1997
  - ❖ In cooperation with research institutes
  - ❖ high level of accuracy needed or postulated
  - ❖ Special Linking for persons, places
- OCR mass production by Google (PPP since 2007/ full text delivered since 2009)
  - ❖ taken as delivered
  - ❖ reiterated processes → living data
- OCR as part of the ZEND-workflow (since 2008)
  - ❖ only for **Antiqua** or **Roman typefaces**

# History 2

---

- Cooperation with research institutions
  - ❖ more than 40 projects with full-text generation
  - ❖ 2/3 outsourced to external service providers
  - ❖ highest level of accuracy: 99,95 % (high financial investment)
- Cooperation with and initiation of developments
  - ❖ EU-project IMPACT: dictionary for 16<sup>th</sup> century

# History 3 Outsourcing and own use of software

- ABBY Recognition Server docked on to the ZEND workflow tool

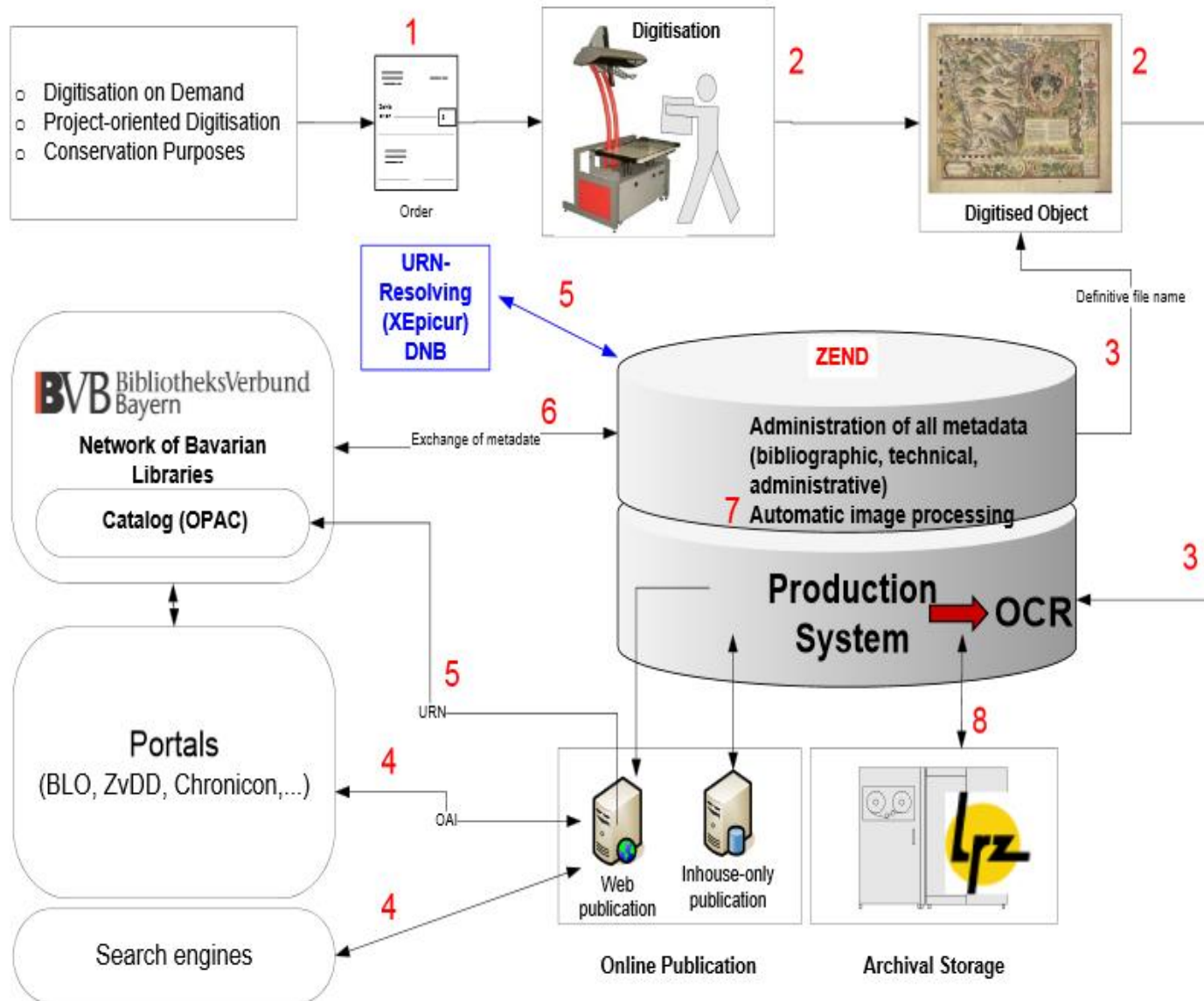
- ❖ unlimited license for Antiqua
- ❖ Gothic typefaces (Fracture) only for testing purposes

```
<!--016-->
<hi rendition="#segm_fontTimesNewRoman #segm_size10." xml:lang="de-DE">
  <seg type="alphaNum" sameAs="#bsb00044385_00018_b001_016_001">eine</seg>
  <seg type="char"> </seg>
  <seg type="alphaNum" sameAs="#bsb00044385_00018_b001_016_002">Neubewertung</seg>
  <seg type="char"> </seg>
  <seg type="alphaNum" sameAs="#bsb00044385_00018_b001_016_003">des</seg>
  <seg type="char"> </seg>
  <seg type="alphaNum" sameAs="#bsb00044385_00018_b001_016_004">Abwehrkampfes</seg>
  <seg type="char"> </seg>
  <seg type="alphaNum" sameAs="#bsb00044385_00018_b001_016_005">gegen</seg>
  <seg type="char"> </seg>
  <seg type="alphaNum" sameAs="#bsb00044385_00018_b001_016_006">Napoleon</seg>
</hi>
</-->
```

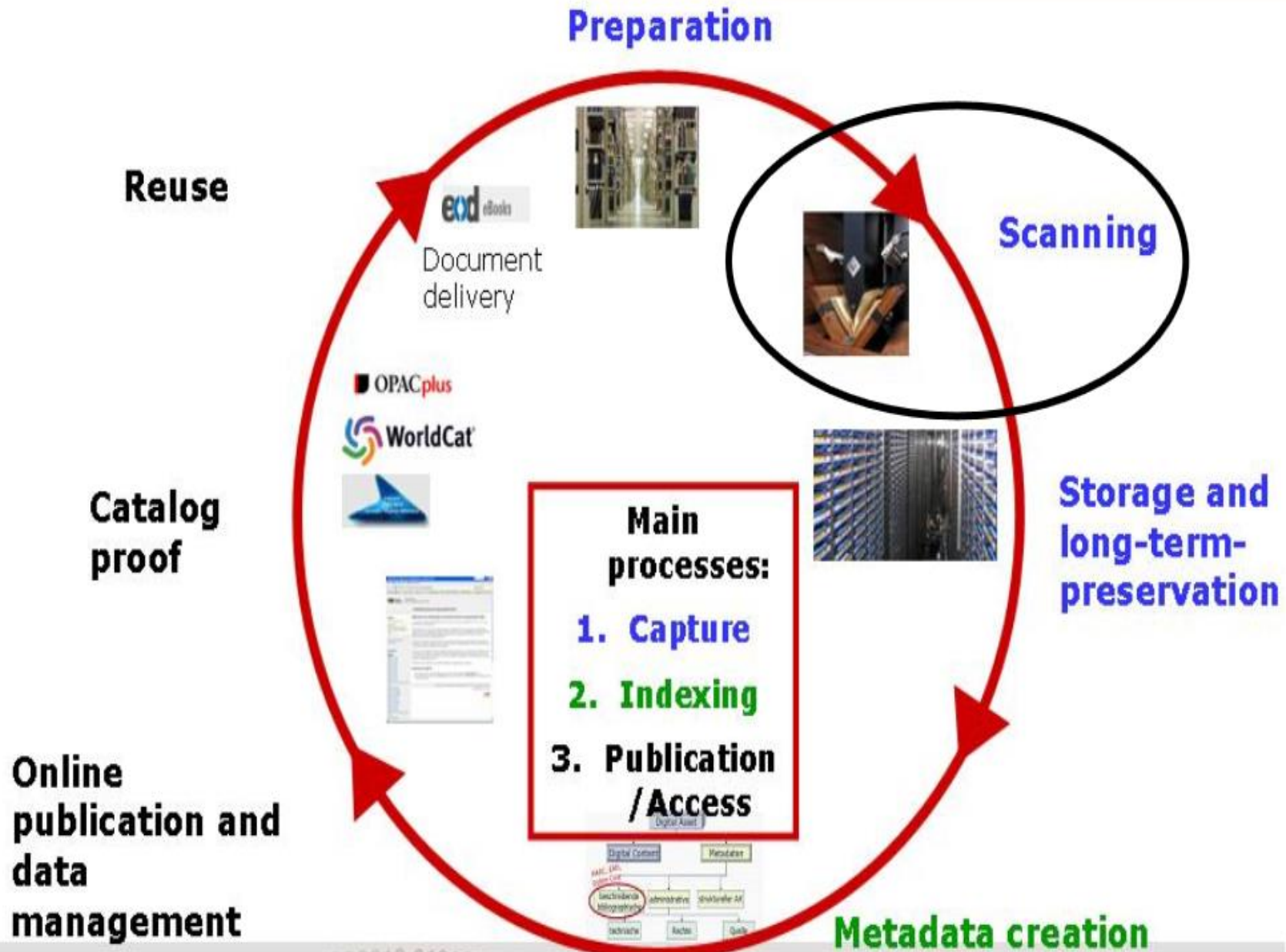
- Service providers

- ❖ mixed experience
- ❖ challenges: contracting rules, vendor guidance, quality management

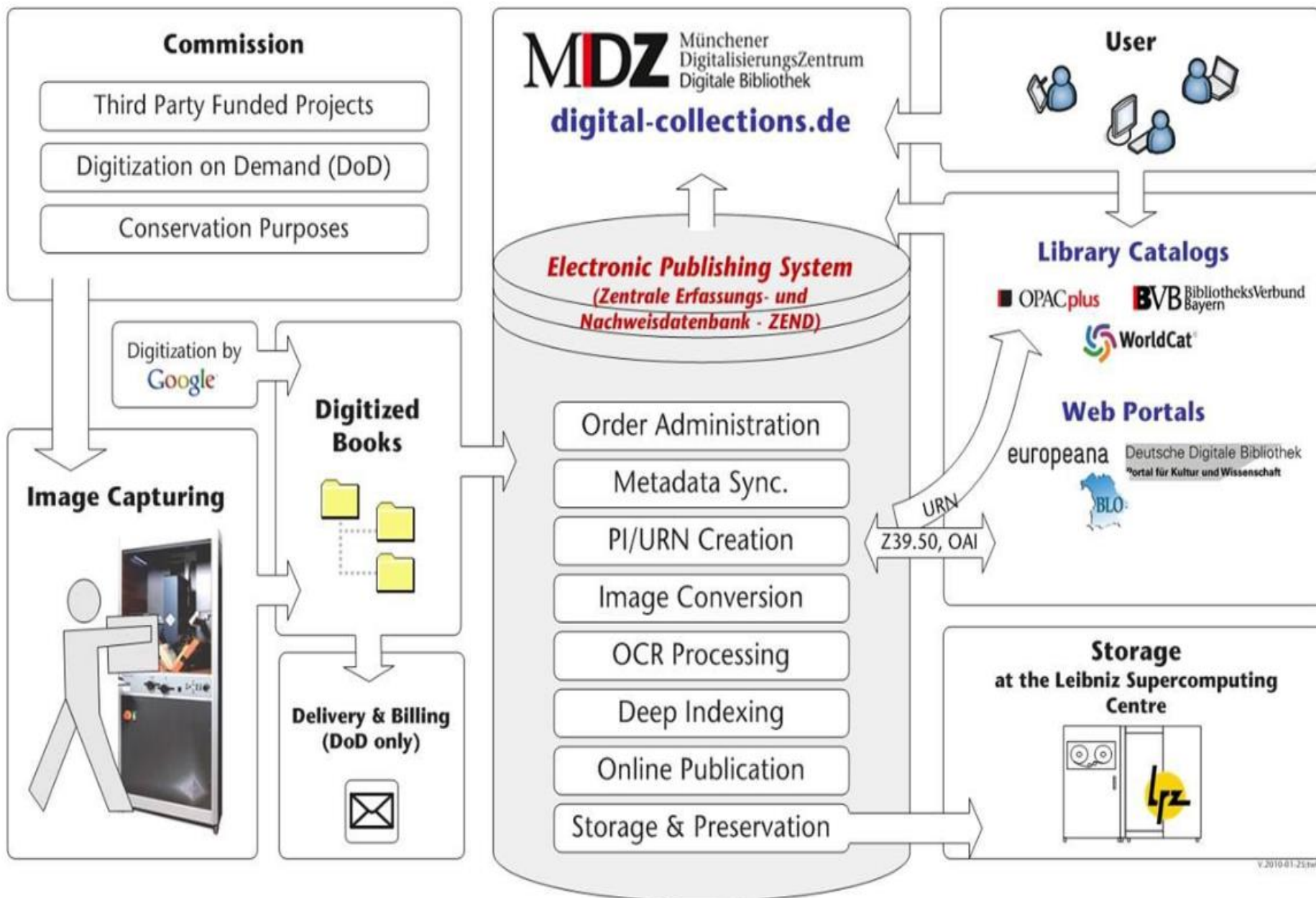
# ZEND at a glance



# Digitization process at a glance



# Stages of production process



# ZEND-OCR-Support

Organisational parameters:

Digitization inhouse or outsourced

Project affiliation

OCR yes/no

Technical parameters:

- Colour depth
- Resolution
- Language information for OCR recognition



# Examples

---

- all examples from 19<sup>th</sup>-20<sup>th</sup> century material (2011: 35 projects, among them 24 with service providers)
- all of explorative character
- not exhaustive
- 19<sup>th</sup> century material – not without problems
- 20<sup>th</sup> century: copyright issues
- Improvement through enrichment and marking (highlighting)
- Main question: **what serves whom and how?**

# Examples

---

- Partial OCR application: Session Reports of the German Reichstag
- Fully automatic approach only: Digi20
- Original with additions: biographical reference work by Lopowsky and Baader

# Little OCR, great results:

## Session Reports of the German Reichstag

- more than 500 volumes, mostly in 4°
- OCR applied only to the registers, but with linking to the indicated text

### Pros:

- + only a small part of the text
- + direct linking to the page
- + no „random“ links (intellectually controlled)

### Cons:

- only a small part of the text
- heterogeneous registers
- difficult text recognition (abbreviations, no sentences)

# Clean enrichment biographical dictionary by Lopowsky and Baader

- early project
- Original version remains intact
  - ❖ only entity markings in the original
  - ❖ additional block with additional data  
(GND, gender, editing)
  - ❖ automatic editing with intellectual control
  - ❖ intellectual enrichment
  - ❖ Combination of lemma and later additions

## Pros:

- + authenticity
- + added value through authority data
- + link to GND

# Automatic approach – Digi20

---

- Copyright-protected material from 3 publishers
- uncorrected OCR with machine coordinates
- Antiqua of 2nd half of 20th century
- Uneven levels of accuracy
- “systematic” misreadings depending on the object:
  - ❖ “m” for “rn”
  - ❖ “h” for “ir”
  - ❖ “k” for “li”
- automatic recognition of entities
- combined search: catalogue data and full text

## OCR – an overview

Prerequisites: a good scan (from a clean original)

- High resolution, deep resolution, if possible no distortions, stright pages
- Image enhancement (delete spots, binarisation, adjustment of pages)
  1. Layout analysis (structure of the page, images etc.)
  2. Segmentation: block, line, word, character
  3. Recognition of characters and classification of words
  4. Lexical analysis.

# Problems:

Heavily used books – cleaning  
Languages, mixed languages  
Special characters

## dMGH

- Project from 2004 to 2010
- The image remains the only reliable reference
- The full text is hidden
- Marking in three parts: Text, notes, text critical information
- High effort in quality control
- Since the relaunch in 2010: full text displayed



Research usually requires a 99,95% accuracy to deduce reliable information from the text – increase of costs

Lower qualities are sufficient – often in conjunction with the images (highlighting of hits) – for searching  
OCR quality

Very good – 99,6-99,95%

Good – 97-99,5%

Average – 90 – 96 %

Bad – below 90 %

OCR software (e.g., state of the art: 2011)

Open-Source-Software

Tesseract (Google, last release 2011)

OCROPUS (last release 2009)

Licensed Software

- Abbyy – different modules
  - Recognition server
  - SDK
  - XIX – for the time being only Omnifont-Gothic available
- Arpa (Paperin Book)
- B.I.T. Tomasi (BIT Alpha)
- Nuance (Omnipage)

# Preparation

- Project management, project aims, organisation
- Juridical aspects to be cleared: especially for 20th century material
- Analysis of the corpus: format, page numbers, physical condition, uncut pages etc.
- OCR via outsourcing:
  - Task description
  - Formal tender and decision: count 3 months

## Einfache Suche

[\[Hilfe\]](#) [\[Neue Suche\]](#) [\[Erweiterte Suche\]](#)


766 Bänd(e) in 3.005 Sek.

1 - 10 > >>

### The semantic organization of the Serbo-Croatian verb

Gorup, Radmila J. - München - Sagner - 1987 - 459 Seiten - [Relevanz: 100%]

Lokalschlüssel: *pv* ; *oe*

Kontext: Suche nach *event* in diesem Band - Mehr als 100 Treffer 

the emphasis on the *event* of 'meeting Mark'. The lower  
Focus meaning in b. tells the reader that this *event* is to  
be backgrounded to some other *event*. Based on this

[\[Suche im Band\]](#) [\[DFG-Viewer\]](#) [\[PDF-Download\]](#)


### Studien zur Semantik des Verbalaspekts im Russischen

Schwenk, Hans-Jörg - München - Sagner - 1991 - 271 Seiten - [Relevanz: 78.47%]

Lokalschlüssel: *pv* ; *oe*

Fachgebiet: Slawistik

Schlagwort: Russisch ; Aspekt <Linguistik> ; Semantik

Kontext: Suche nach *event* in diesem Band - 72 Treffer 

"An independent predicate contains in its meaning a causing  
*event* and a caused *event*. The causing *event* causes the caused  
*event* or, in other words, the caused *event* is the result of  
the causing *event*, i.e. causing *event* and caused event are  
connected by the causal relationship."

[\[Suche im Band\]](#) [\[DFG-Viewer\]](#) [\[PDF-Download\]](#)


### Ereignisse und andere Partikularien : Vorbemerkungen zu einer mehrkategorialen Ontologie

Kanzian, Christian - Paderborn ; München [u.a.] - Schöningh - 2001 - 267 Seiten - [Relevanz: 66.08%]

Lokalschlüssel: *pm*

Fachgebiet: Philosophie

Schlagwort: Ereignis ; Ontologie

Kontext: Suche nach *event* in diesem Band - 86 Treffer 

Suchergebnis einschränken

#### Verlag


- NICHT ANGEGEBEN (6)
- Bärenreiter (9)
- Bärenreiter-Verl. (1)
- Éd. Univ. [u.a.] (1)
- Fink (269)
- Fink [u.a.] (2)
- Finnish Exegetical Soc. [u.a.] (2)
- Fischer [u.a.] (2)
- [Mehr...](#)

#### Erscheinungsjahr

- NICHT ANGEGEBEN (218)
- 1931 (1)
- 1959 (1)
- 1962 (1)
- 1963 (2)
- 1966 (2)
- 1967 (2)
- 1969 (3)
- [Mehr...](#)

#### Automatisch erkannte Personen (Mehr als 100)

- Wilhelm Fink (263)
- Friedrich Wilhelm (161)
- Max Weber (151)
- Karl Marx (134)
- Thomas Mann (132)
- Georg Wilhelm (126)
- Martin Luther (123)
- Jean Paul (116)
- [Mehr...](#)

Zurück zum Suchergebnis *event* 

### The semantic organization of the Serbo-Croatian verb

Autor / Hrsg.: Gorup, Radmila J.

Verlagsort: München | Erscheinungsjahr:

1987 | Verlag: Sagner

Signatur: Z 60.523-214

[\[Suche im Band\]](#) [\[DFG-Viewer\]](#)

[\[PDF-Download\]](#)

### Wörter

Automatisch erkannte Personen

Automatisch erkannte Orte

Ähnliche Dokumente

ABKÜRZUNGEN ABTEILUNG

ägypten AKZENT ALBRECHT

alois ANALYSE ANALYSEN

ANGABE ANGELES APRIL

AUFLAGE AUGUST AUSSPRACHE

BAND BARBARA BEGRÜNDET

BEITRÄGE BENJAMIN berger

BERÜCKSICHTIGUNG

BESCHREIBUNG BEZUG BILD

BILDER BRIEF cambridge



CHANCE CHARAKTERISIERUNG

CHURCH CITY college

DÄMONEN details

development EAST EDITION


### Suche im Band


 event 

### Optionen

Zeige nur Treffer von Seiten die alle Suchwörter enthalten

Trefferliste nach Seiten gruppieren

Sortierung nach Relevanz, absteigend 

10 Treffer pro Seite 

[\[Hilfe\]](#) [\[Neue Suche\]](#) [\[Neue Suche in diesem Band\]](#)

Suche in diesem Band - 172 Treffer in 0.051 Sek.

1 - 10 > >>

 Scan 97 | 100%

instructed to withhold any emphasis but instead to place the emphasis on the **event** of 'meeting Mark'. The lower Focus meaning in b. tells the reader that this **event** is to be backgrounded to some other **event**. Based on this instruction and the immediate context the hearer infers the

 Scan 181 | 99.4%

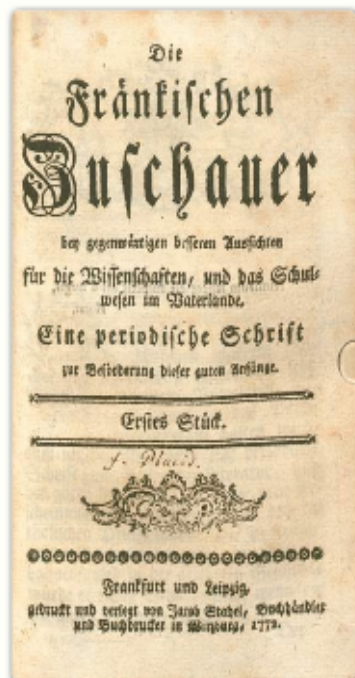
narrative. This move specifies the point in the story at which another **event** will take place. As it is concerned only with the specification of the **event** it is less important than the **event** itself and that is why the author downgrades the Focus of attention one step.

 Scan 191 | 98.8%

## For OCR

- copyright free material
- copyright protected material in the context of cooperations:
  - whole book
  - only text (no illustrations)
  - only parts of the book, articles
- selection of appropriate copy (undommaged, no pages missing)
- choice of production line according to script: antiqua inhouse, gothic outsourced.

## Daily and weekly newspapers, scholarly journals



### Gelehrte Journale und Zeitungen als Netzwerke des Wissens im Zeitalter der Aufklärung

No information in English language available yet.

'Was ist Aufklärung?' – Nicht zufällig wurde die vielleicht berühmteste Frage des Jahrhunderts 1784 in einer Monatsschrift ausgeschrieben, denn Zeitungen und Zeitschriften waren längst ‚die Vorratskammern des menschlichen Verstandes‘ geworden. In Kooperation mit und auf Initiative der Göttinger Akademie der Wissenschaften, der SUB Göttingen und der UB Leipzig werden an der Bayerischen Staatsbibliothek Zeitschriften des 18. Jahrhunderts erschlossen und digitalisiert. Im Mittelpunkt des Münchener Digitalisierungsanteils, das durch Kontingente in Göttingen und Leipzig

ergänzt wird, stehen dabei katholische Zeitschriften aus dem letzten Drittel des 18. Jahrhunderts, die die besondere Ausprägung der bayerischen Spätaufklärung verdeutlichen.

To Start Page

## Search the collections



Title/Author/Label



Data basis: 1.049.045 Digitized Works

## Explore



Subjects | Authors | Time | Place of pu



Cimelia of the Augsburg State and City Library- chronicles, liturgical and literary texts as well as natural history objects

New



The Bellifortis by Konrad Kyser illustrates on more than 300 pages military technology from the perspective of the Middle Ages. It also contains the first known picture of a chastity belt. This manuscript was created between 1402 and 1405 in Bohemia.

New

Volltext

**Finden!**

[\[Erweiterte Suche\]](#) [\[Hilfe\]](#)

**Suche einschränken**  
(gefundene Seiten)

**Zeitung**

- Mittelbayerische Zeitung (217)
- Passauer Neue Presse (1741)
- Passauer neue Presse : Niederbayerische Zeitung (1315)

**Jahr**

- 1945 (3)
- 1946 (37)
- 1947 (31)
- 1948 (96)
- 1949 (164)
- 1950 (144)
- 1951 (71)
- 1952 (98)

• [Mehr...](#)

**Monat**

- Januar (250)
- Februar (237)
- März (325)
- April (334)
- Mai (282)
- Juni (281)
- Juli (276)
- August (250)

• [Mehr...](#)

**Tag**

**Suche nach (konditorei) \*.\* - 3273 Treffer**

Relevanz: 1

**Passauer Neue Presse Nr. 252 - Seite 16 - vom 31.10.1962**

**Überschrift**

Relevanz: 1

**Passauer Neue Presse Nr. 298 - Seite 8 - vom 24.12.1964**

**Überschrift**

Relevanz: 0.8333

**Passauer neue Presse : Niederbayerische Zeitung Nr. 228 - Seite 34 - vom 02.10.1965**

**Überschrift**

**Seitentext**



## Einfache Suche

historie

[Hilfe] [Neue Suche] [Erweiterte Suche]

3367 Bänd(e) in 0.994 Sek.

1 - 7 >

### Historie von einem armen Waißlein Victoria genannt

S.I. - 1645 - 4 Bavar. 2197,V,1/70#Cah.57 - 8 Seiten - [Relevanz: 100%]

Signaturenfach: Bavarica


[Suche im Band] [PDF-Download] [OPAC]

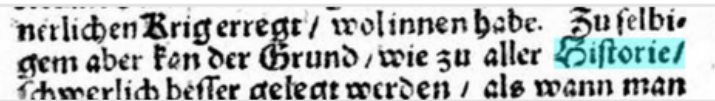
### Worinnen der Sachen wahrer Verlauff, vom Absterben Königs Carl des Zweyten in Spanien her, biß auff instehende Zeit, getreulich erzehlet, und mit den gehörigen Documenten, bewähret, erläutert und vorgestellt wird

Hartmann, Johann Jacob - Cölln [i.e. Nürnberg] - Marteau i.e. Hoffmann - 1703 - Bavar. 113-1 - 778 Seiten - [Relevanz: 85.93%]

Reihe: Außführliche Historie Des jetzigen Bayrischen Kriegs [...]; [1]

Signaturenfach: Bavarica

Kontext: Suche nach *historie* in diesem Band - 2 Treffer 




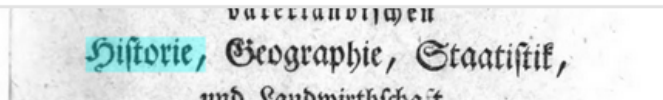
[Suche im Band] [PDF-Download] [OPAC]

### Beyträge zur vaterländischen Historie, Geographie, Staatistik, etc. ; 5. 1794

Westenrieder, Lorenz von - München - Lindauer - 1794 - Bavar. 4812 s-5 - 468 Seiten - [Relevanz: 83.33%]

Signaturenfach: Bavarica

Kontext: Suche nach *historie* in diesem Band - 2 Treffer 



[Suche im Band] [PDF-Download] [OPAC]

## Suchergebnis einschränken

### Erscheinungszeitraum

- 1800 bis \* (2237)
- 1700 bis 1799 (1021)
- 1600 bis 1699 (102)
- \* bis 1599 (3)

### Automatisch erkannte Personen (mehr als 100)

- Kaiser Karl (983)
- Johann Georg (929)
- Kaiser Ludwig (907)
- König Ludwig (900)
- Herzog Ludwig (864)
- Kaiser Heinrich (798)
- Kaiser Friedrich (767)
- Herzog Wilhelm (766)
- Mehr...

### Automatisch erkannte Orte (mehr als 100)

- München (2359)
- Rom (2027)
- Nürnberg (1993)
- Wien (1985)
- Bayern (1970)
- Augsburg (1879)
- Frankreich (1868)
- Deutschland (1851)
- Mehr...

### Signaturenfach

- Ascetica (Geistliche Erbauungsliteratur, Gebetbücher, Mystik) (1)
- Bavarica (3263)

# Perspectives

---

- use of catalogue data for OCR
- choice of dictionaries according to publication year
- choice of dictionaries for subject indexing
- use of index data
- Solr as a dictionary created “on the fly”
- image search

# Perspective – DFG project

---

- DFG coordinated initiative/project for further development of OCR

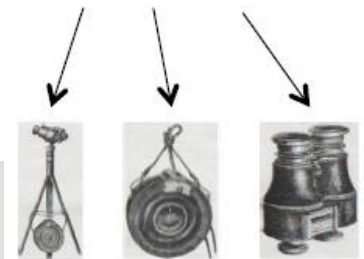
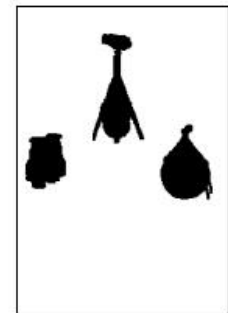
(BSB, Wolfenbüttel, Deutsches Textarchiv):


Tasks, concepts, corpora, open source tools, standardisation of workflows, interoperability of data, scalability, impact, long term archiving, crowdsourcing

Definition of modules (practical experiences), work packages for further DFG projects in this line of initiative

# ImageSearch

- <http://bildsuche.digitale-sammlungen.de/>
- separation of colors and contours
- automatic separation of image and text in very large corpora
- further development: application to texts?  
→ preparation for OCR?



Suche 



Places



Highlights



Topics



Literature Portal Bavaria



Similarity Based Image Search

**bavarikon**  
Kultur und Wissensschätze Bayerns



Maps



3D Objects



Historical Encyclopedia of Bavaria



People



Institutions



Objects

# Bavarikon: the Bavarian cultural portal

# Perspective - Data for research

---

- = in-house project of MDZ in cooperation with LRZ
- aims at a customer-induced self-service for images
  - JPEG from the original in 300 dpi with 95 % text accuracy
- Pilot project to evaluate technical challenges (with a limited number of downloads per user)
- Cost evaluation: Administration of data, storage space, maintenance + development costs for a highly performative service

Research ...

Images – Structural data – full text (hiding, showing)

Downloading of full text:

Juridical issues, technical issues

Searchable pdf: generation of full text

Uploading of full text - no programs

Transcribed texts – no highlighting

Thank you for your attention!

