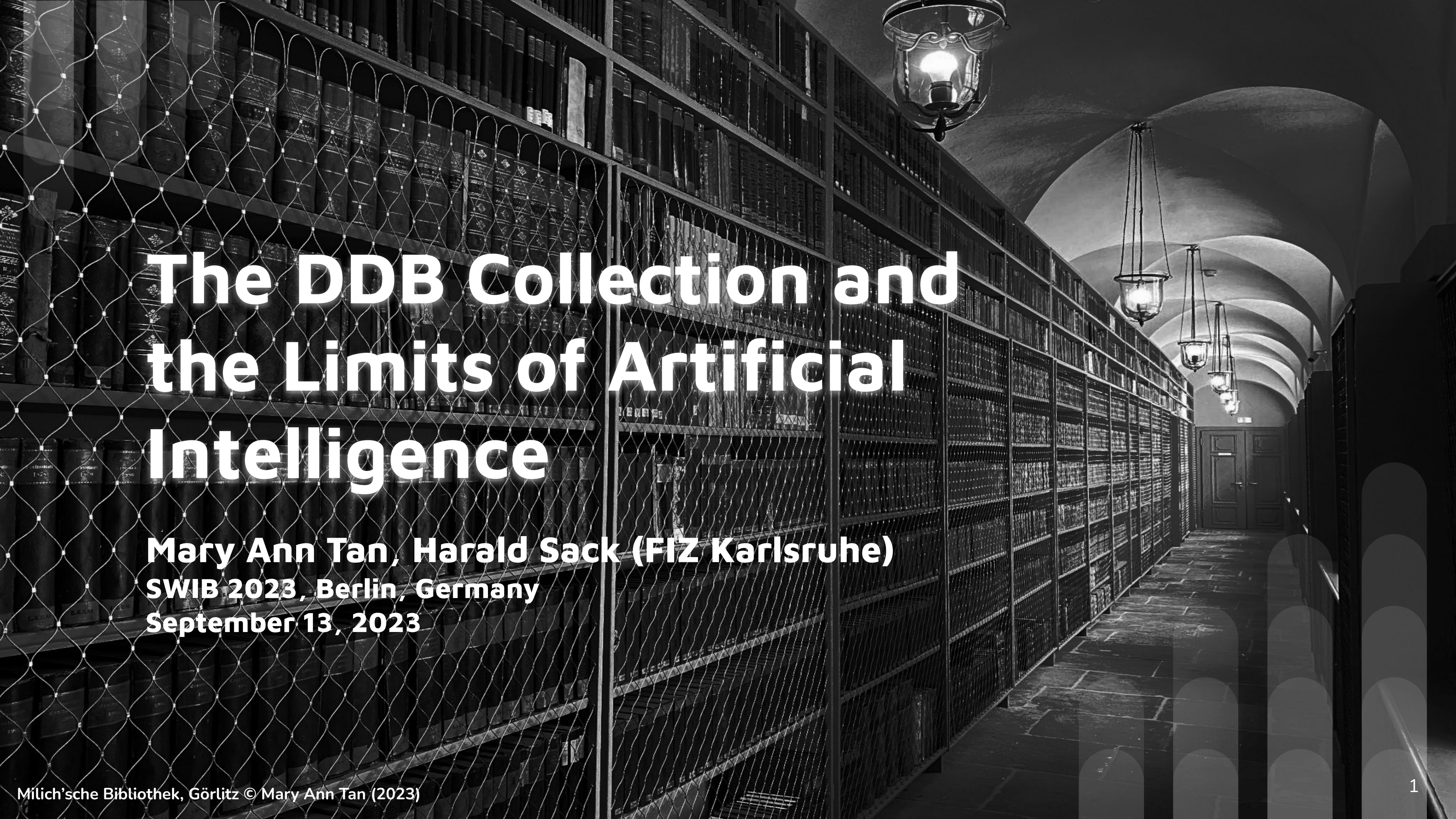


The DDB Collection and the Limits of Artificial Intelligence

Mary Ann Tan, Harald Sack (FIZ Karlsruhe)

SWIB 2023, Berlin, Germany
September 13, 2023

Milich'sche Bibliothek, Görlitz © Mary Ann Tan (2023)

The DDB Collection and the Challenges of Legacy Metadata

Mary Ann Tan, Harald Sack (FIZ Karlsruhe)

SWIB 2023, Berlin, Germany
September 13, 2023

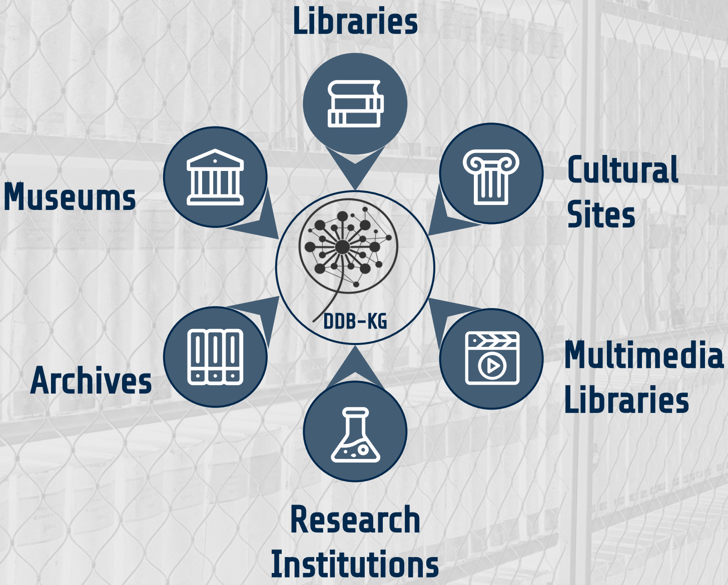
Milich’sche Bibliothek, Görlitz © Mary Ann Tan (2023)

OUTLINE

1. DDB in a Knowledge Graph
2. The DDB Pie and Art in Marzipan
3. Metadata Models
4. FRBR-ization
5. Work Reconciliation Results
6. Cataloging Standards
7. Conclusion

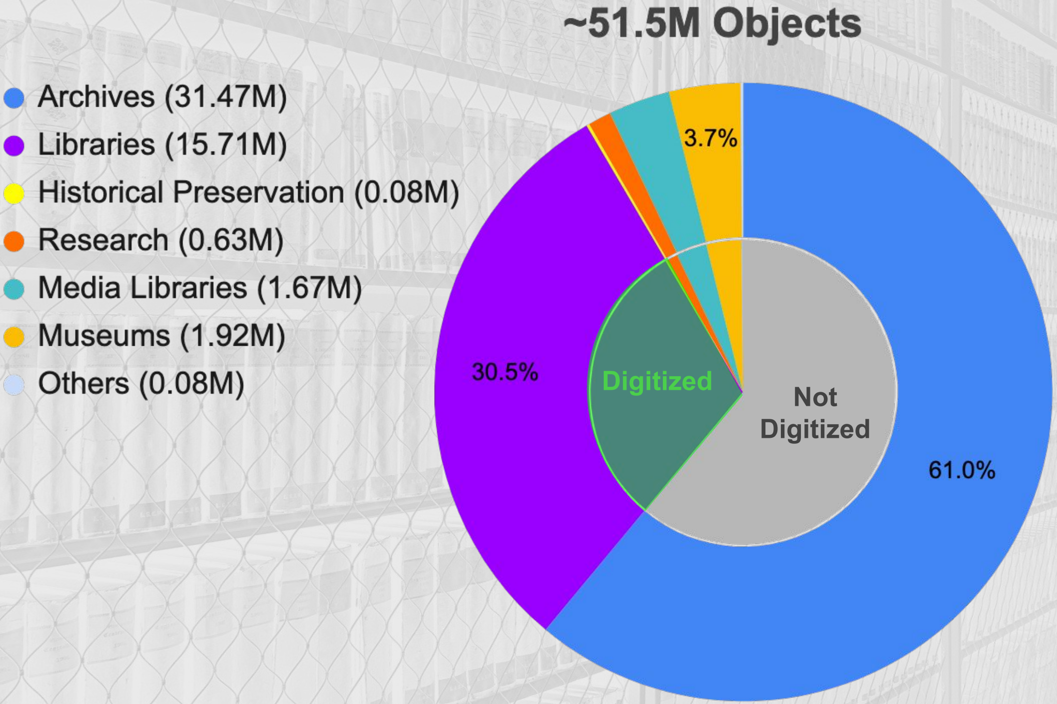
DDB in a Knowledge Graph

How to help users find what they are looking for?



- Semantic Search
- Exploratory Search

The DDB Pie



As of August 2023:

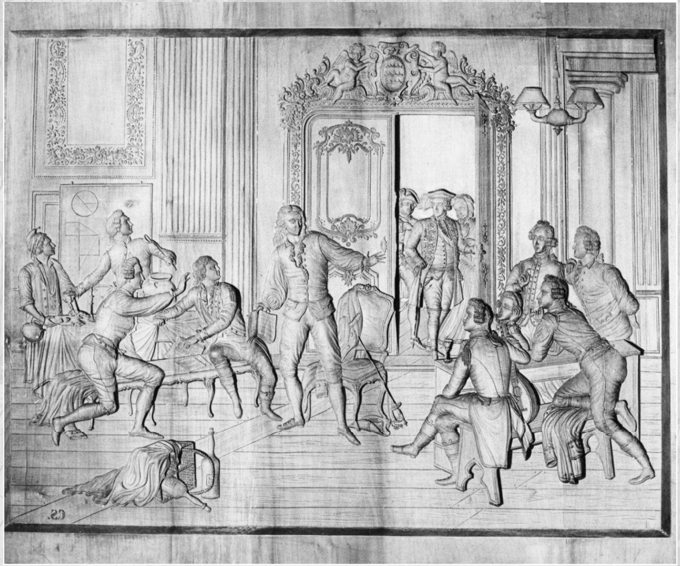
- Libraries have the most number of digitized objects
  - 12.67M **digitized texts**

Art in Marzipan

Search results of “Schillers Räuber”

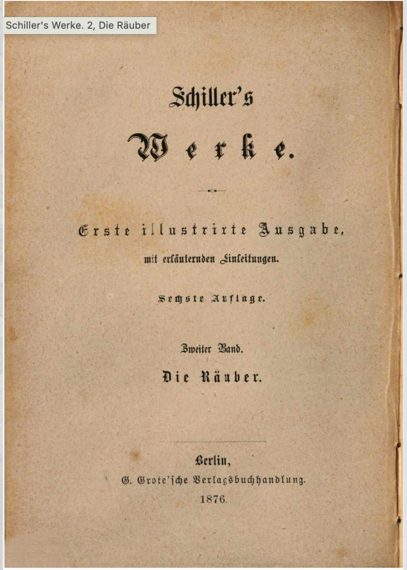
http://bit.ly/schillers-raeuber-results

Friedrich Schiller reciting “The Robbers”



http://bit.ly/fs-results-marzipan

Schiller’s Works “The Robbers”



Schiller's Werke. 2, Die Räuber

Schiller's

Werke.

Erste illustrirte Ausgabe,

mit erläuternden Einleitungen.

Sechste Auflage.

Zweiter Band.

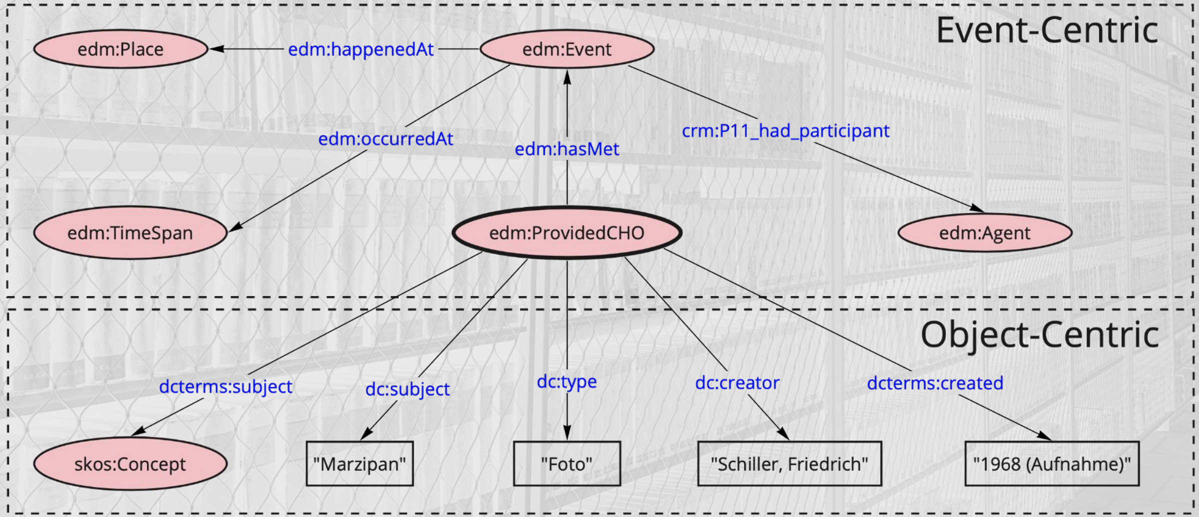
Die Räuber.

Berlin,
G. Grote'ſche Verlagsbuchhandlung.
1876.

http://bit.ly/die-raeuber-aufzug3

DDB-EDM Metadata Model

- Extension of the Europeana Data Model
- Event- and object-centric modeling paradigm

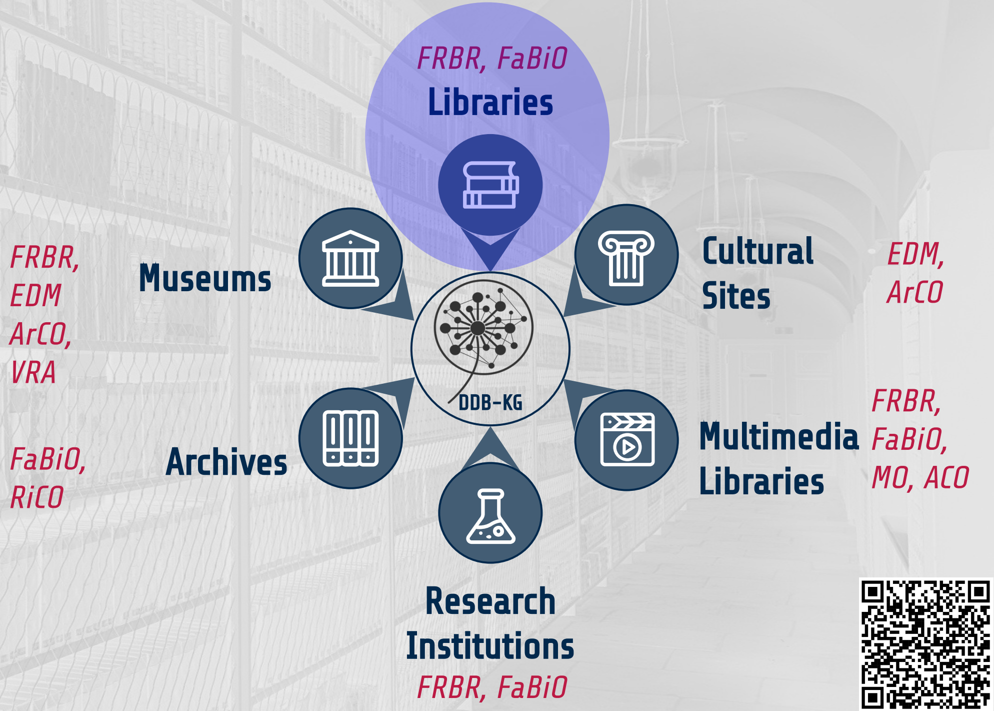


! ALL objects are instances of edm:ProvidedCHO

DDB-O: Domain Interoperability

Between art in marzipan and a book:

- How to **distinguish**?
- How are they **related**?



https://ise-fizkarlsruhe.github.io/ddbkg/ddbo

FRBR-ization

How DDB objects are aligned:

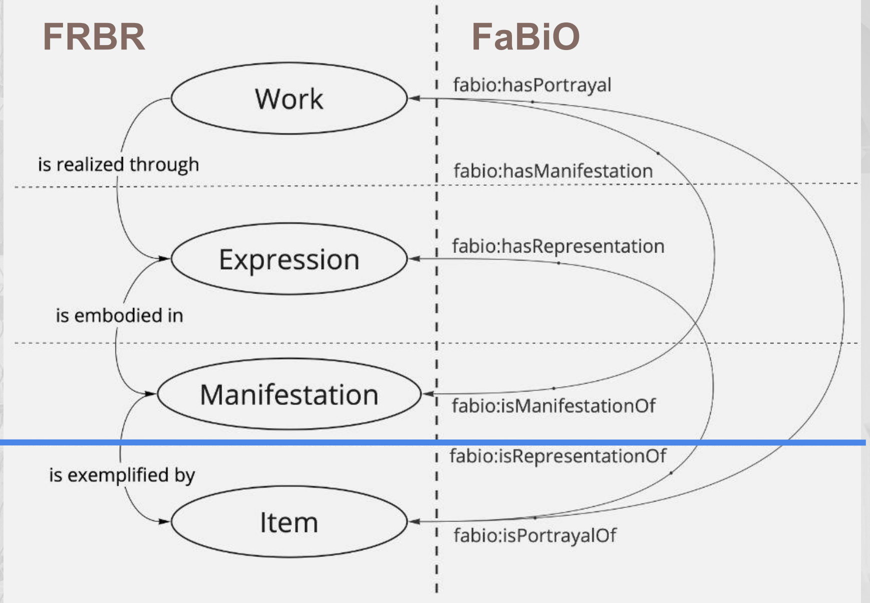
Most users seek Works & Expressions

Work

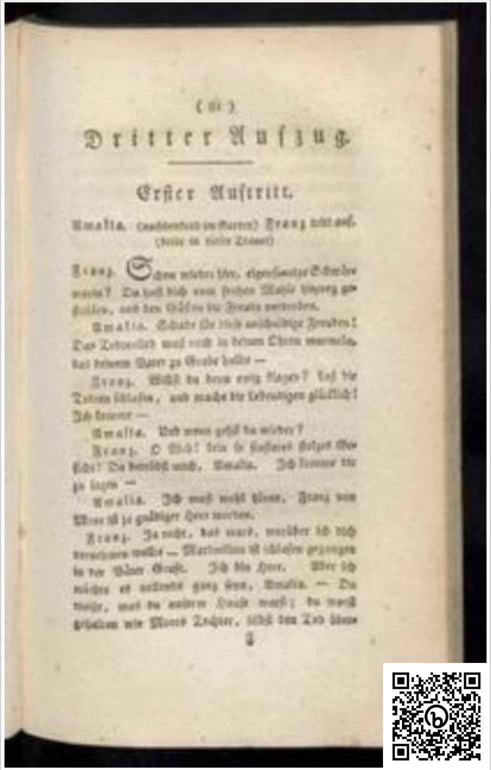
Expression

Manifestation

edm:ProvidedCHO →



Hierarchy and Granularity



(61)

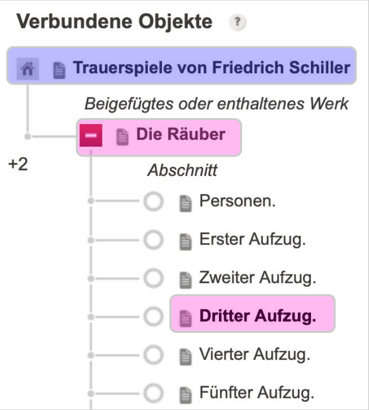
Dritter Aufzug.

Erſter Auftritt.

[illegible]

1 “Object” = 1 Metadata ≤ 1 Item

- Primary
  - “Trauerspiele von Friedrich Schiller”, the item itself
- Secondary - **“Die Räuber”, “Dritter Aufzug”**
  - Parts of the item, that could also be aligned to a Work

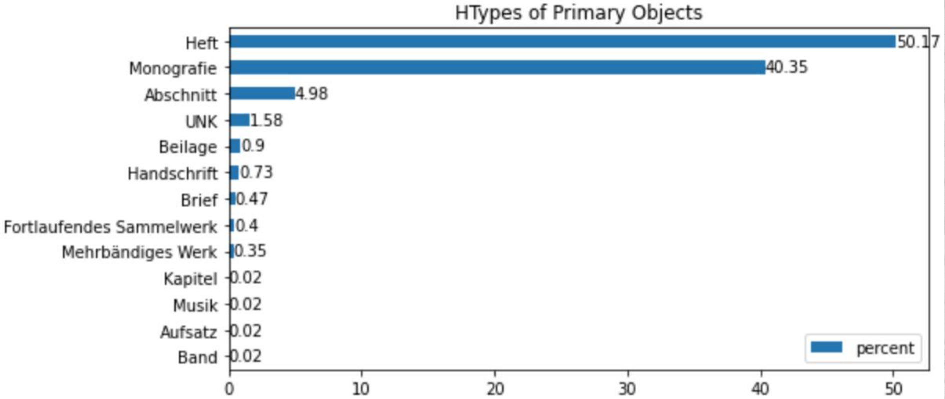


Verbundene Objekte ?

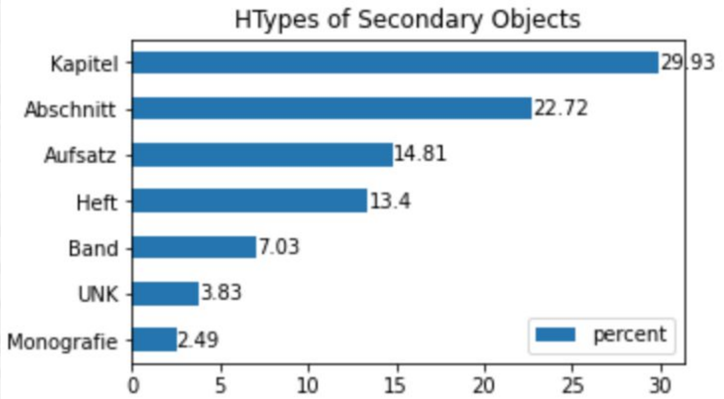
- Trauerspiele von Friedrich Schiller
  - *Beigefügtes oder enthaltenes Werk*
  - Die Räuber
    - *Abschnitt*
    - ○ Personen.
    - ○ Erster Aufzug.
    - ○ Zweiter Aufzug.
    - ○ **Dritter Aufzug.**
    - ○ Vierter Aufzug.
    - ○ Fünfter Aufzug.
  - +2

Hierarchy Types (HTypes)

- Primary Objects



- Secondary Objects



“Content” HTypes that could be Works:

- Monografie (Monography)
- Abschnitt (Section)
- Handschrift (Handwriting)
- Musik

- Kapitel (Chapter)
- Aufsatz (Essay)
- Mehrbändiges Werk (Multi-volume Work)
- Band (Volume)

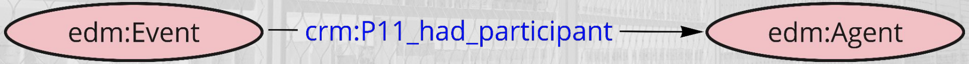
Properties for Reconciliation

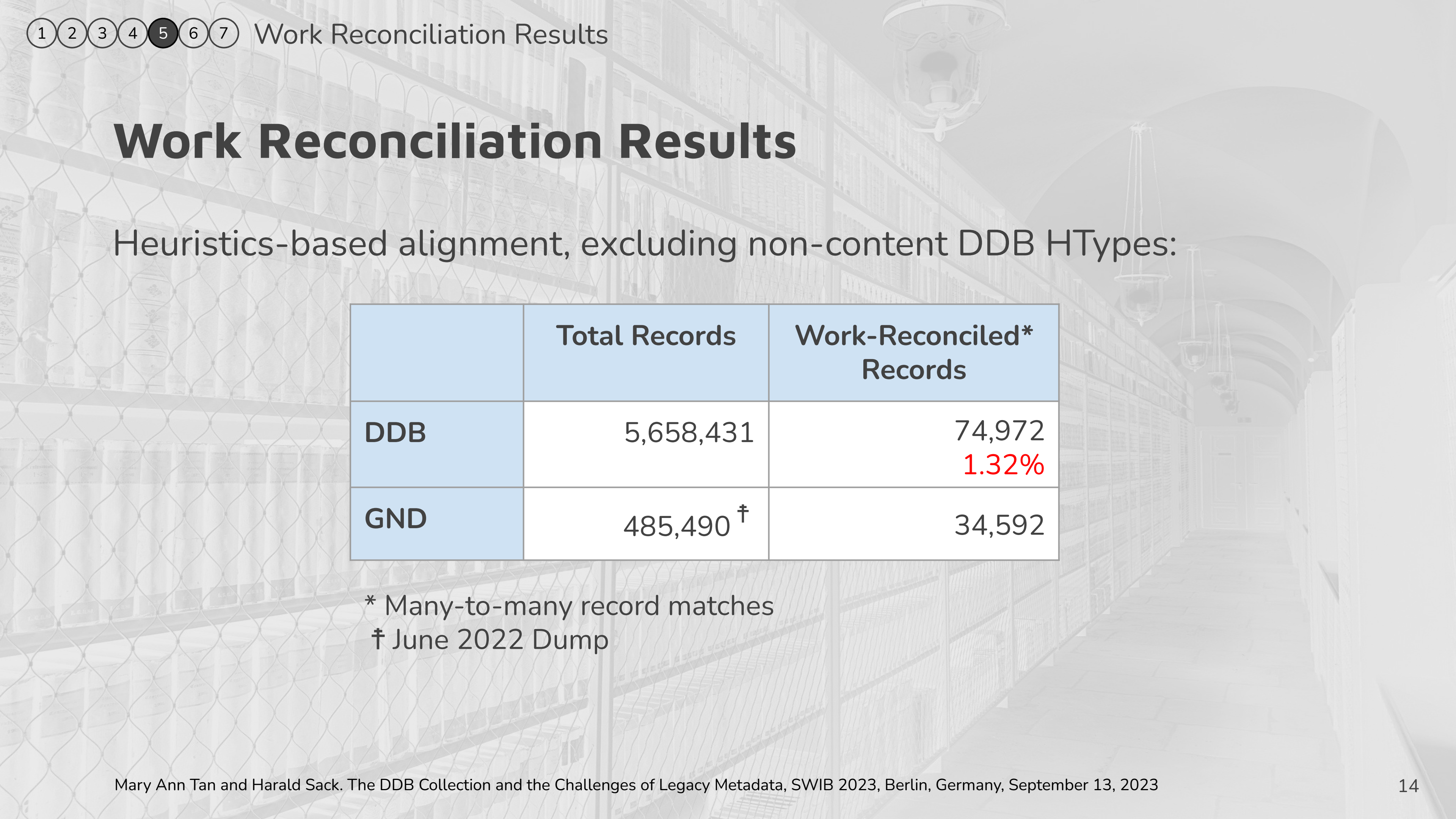
- With **GND Werke** as a reference

Fields	Properties	Matching Criteria
Title	dc:title dc:alternative	Full-text search (FTS)
Agent	dc:creator dc:contributor dc:publisher	GND ID FTS
Event	dcterms:created dcterms:issued	Item date >= GND date

DDB Property Values

Flexibility in values:

- Literals
- Controlled Vocabularies
- Linked Open Data:
    - Agents: 
      edm:Event
      crm:P11_had_participant
      edm:Agent
    - creators / publishers / producers
    - Only **36.64%** linked to (GND, VIAF, etc.)



Work Reconciliation Results

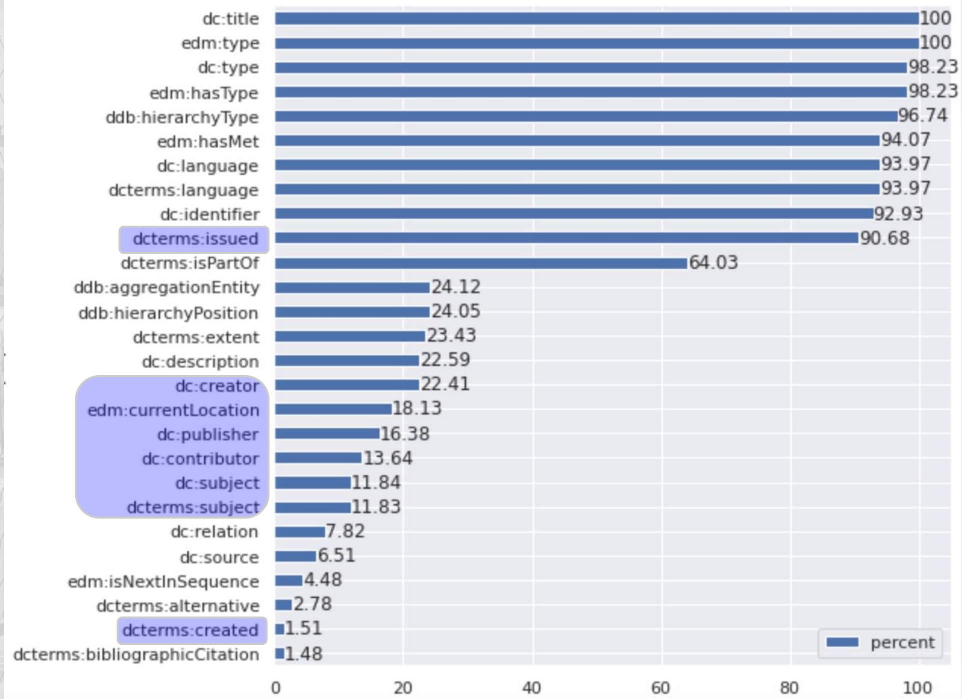
Heuristics-based alignment, excluding non-content DDB HTypes:

	Total Records	Work-Reconciled* Records
DDB	5,658,431	74,972 1.32%
GND	485,490 ‡	34,592

* Many-to-many record matches
‡ June 2022 Dump

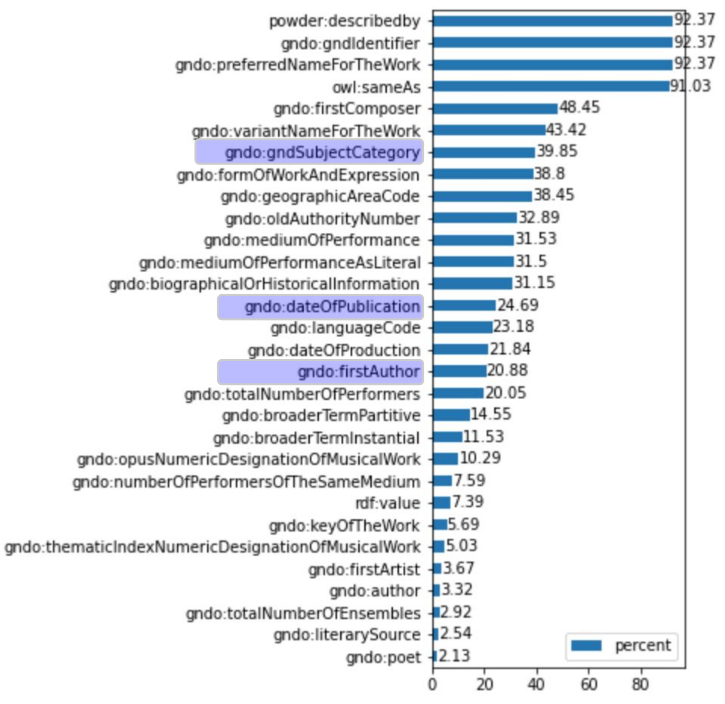
DDB Data and Object Properties

! Properties to identify an item as a possible creative work are **rarely present**.



Missing Context Descriptions

GND Properties:

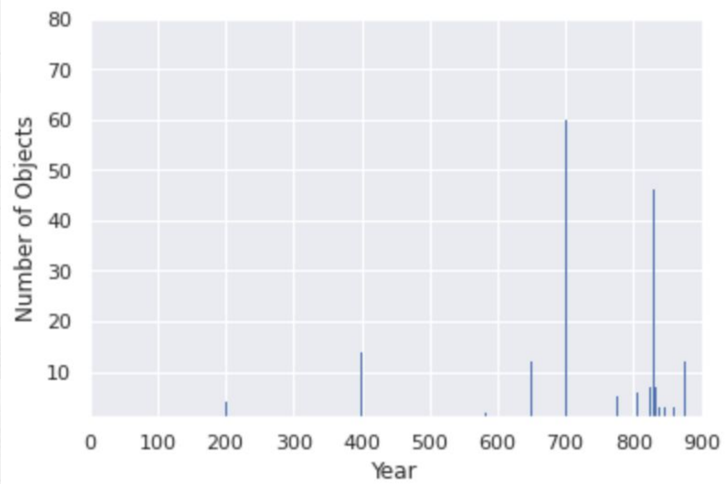


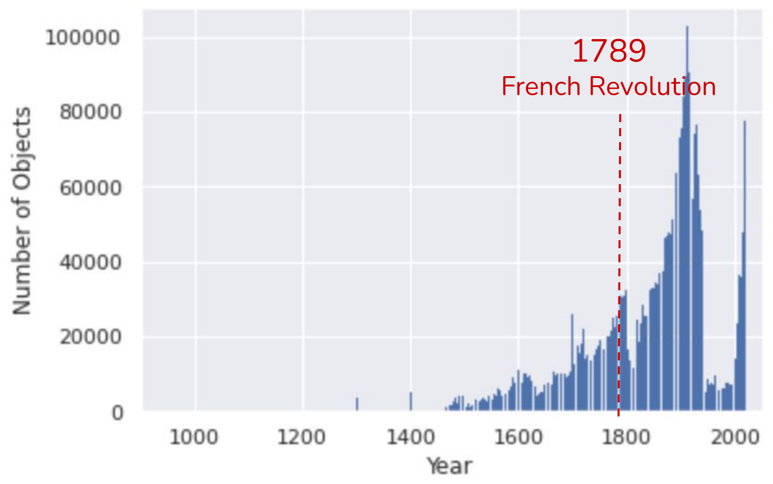
	Total Records	Records without Agents	Records without Dates
DDB	5,658,431	59.27%	11.84%
GND	485,490	33.68%	58.85%

Why are pertinent information missing?

Cataloging Standards

- Objects in the DDB span several millennia
- Heterogeneous cataloging standards [LOC 2017]
- Items from antiquities may have no author attribution

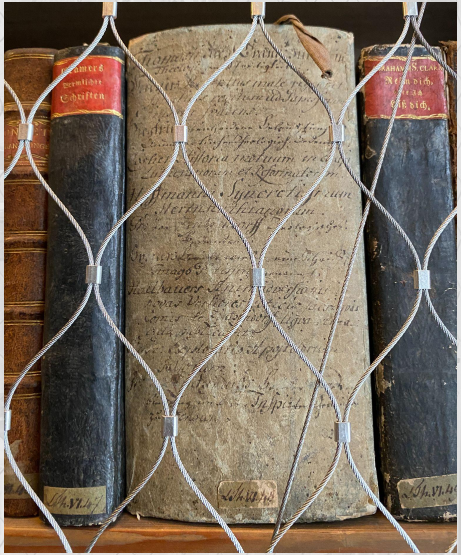




Can we compute similarity measures on titles alone?

Matching Non-Contemporary Titles

- Lengthy titles
- Included extraneous information



A tome from the Milich'sche Library, Görlitz.

Average token lengths of top 5 languages found in titles*:

- de - 13.27
- en - 8.63
- la - 25.64
- fr - 25.60
- it - 42.54 †

* Using both *langid (Lui et al., 2011)* and *bi-lstm (Toftrup et al., 2021)* with ~60% accuracy

† *identifiers confuse similar languages*

Semantic Similarity on Titles

Examples from using Pre-trained Language Models (PLMs + sentence transformers + FAISS †)

GND Title	DDB Title	Similarity Score †
Ilias	Bd. 1: Ilias ✓	0.6991
	Homeri Ilias : Accedunt Hymni Homeridarum Et Epigrammata Collection	0.4121
	Bd. 1: Homers Iliade : Travestirt nach Blumauer Parody	0.1904
	Bd. 4 = Abtheilung 2: Des Homerus Werke. Vierter und letzter Band. Iliade. Zweite Abtheilung ✓	0.1175

Conclusion

- Significant challenges in the FRBR-ization of legacy metadata.
  - Insufficient information.
  - Non-uniform metadata quality.
  - External reference also with incomplete information.
- Using PLMs to find matches based on titles alone is not reliable.
- Further initiatives require inputs from:
  - Data provider for metadata quality improvement.
  - Domain experts to confirm possible work matches.

Thank You for Your Attention!

Want to learn more about Knowledge Graphs?
Join our free online course at OpenHPI!

https://open.hpi.de/courses/knowledgegraphs2023

📧ann (dot) tan (at) fiz (dash) karlsruhe (dot) de
🌐https://ise-fizkarlsruhe.github.io/ddbkg/

FIZ Karlsruhe

KIT
Karlsruher Institut für Technologie

NEU START KULTUR

References

1. Mads Toftrup, Søren Asger Sørensen, Manuel R. Ciosici, and Ira Assent. 2021. **A reproduction of Apple's bi-directional LSTM models for language identification in short strings.** In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 36–42, Online. Association for Computational Linguistics.

2. Marco Lui, and Timothy Baldwin. 2011. **Cross-domain Feature Selection for Language Identification.** In Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, pp. 553—561. Available from http://www.aclweb.org/anthology/I11-1062

3. The Library of Congress. 2017. **The Card Catalog: Books, Cards, and Literary Treasures.** Chronicle Books, San Francisco.

Images:

- Title Page: Milich'sche Bibliothek, Barockhaus Görlitz. © Mary Ann Tan (2023)