Ähnlichkeiten

Berlin, 17. Januar 2024
Stiftung Deutsche Kinemathek

Ziele

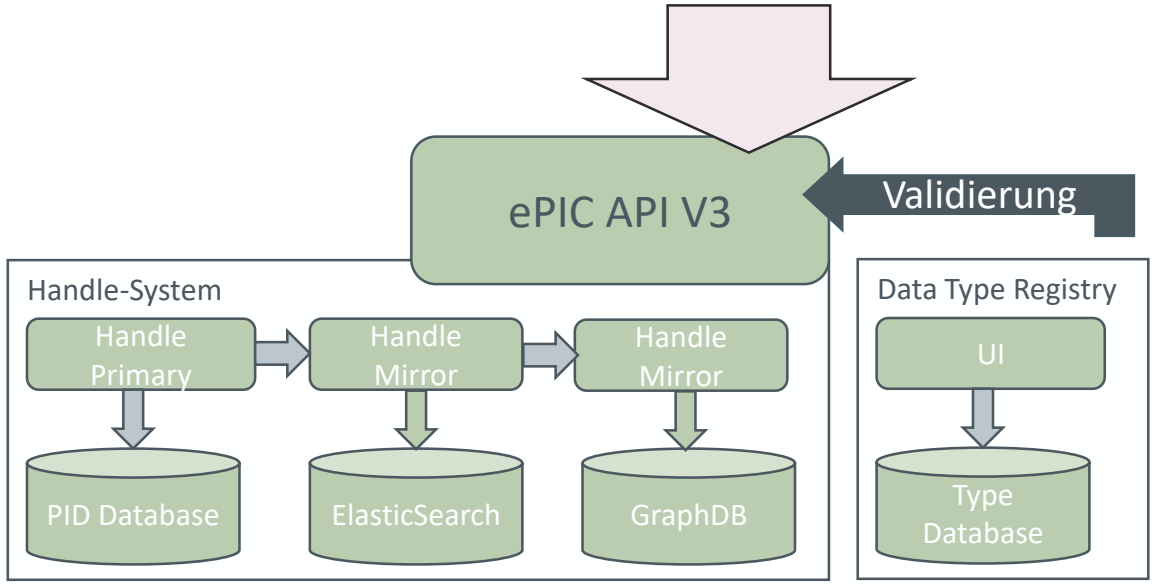
- Disambiguierung / Duplikaten
- Abgleich von Metadaten
- Verbesserung der Suche

Literatur (Ausschnitt)

- PFEIFER, Barbara; POLAK-BENNEMANN, Renate. Zusammenführen was zusammengehört – Intellektuelle und automatische Erfassung von Werken nach RDA. o-bib. Das offene Bibliotheksjournal / herausgegeben vom VDB, [S.l.], v. 3, n. 4, p. 144-155, dec. 2016. ISSN 2363-9814. Verfügbar unter: <https://www.o-bib.de/article/view/2016H4S144-155>. Date accessed: 25 oct. 2017. doi:http://dx.doi.org/10.5282/o-bib/2016H4S144-155.
- Pfeffer, Magnus. „Using Clustering Across Union Catalogues to Enrich Entries with Indexing Information.” In Data Analysis, Machine Learning and Knowledge Discovery, herausgegeben von Myra Spiliopoulou, Lars Schmidt-Thieme und Ruth Janning.437-445. Cham: Springer International Publishing, 2014. https://dx.doi.org/10.1007/978-3-319-01595-8
- García Mendoza, Juan Luis: Procedimiento para la mejora de la completitud en registros bibliográficos con formato MARC 21. https://dspace.uclv.edu.cu/handle/123456789/10674
- Jeffrey Beall: Measuring duplicate metadata records in library databases. Library Hi Tech News 27(9/10):10-12. DOI: 10.1108/07419051011110595 https://www.researchgate.net/publication/238599321_Measuring_duplicate_metadata_records_in_library_databases
- Guido Sautter - Klemens Böhm - David King: RefConcile – Automated Online Reconciliation of Bibliographic References. ICADL 2013: Digital Libraries: Social Media and Community Networks pp 161-170. https://link.springer.com/chapter/10.1007/978-3-319-03599-4_20
- Asma Abboura - Soror Sahri - Latifa Baba-Hamed - Mourad Ouziri - Salima Benbernou: Quality-Based Online Data Reconciliation. ACM Transactions on Internet Technology Vol. 16, 2016. No. 1. https://dl.acm.org/doi/10.1145/2806888
- Alain BOREL, Jan KRAUSE: MarcXimiL. the bibliographic similarity analysis framework http://marcximil.sourceforge.net/
- Years of experience in the preparation and processing of metadata meets methods of Machine Learning — https://swissbib.blogspot.com/2020/07/years-of-experience-in-preparation-and.html
- A Machine Learning Approach with Ensemble Methods for Deduplication of Swissbib Data — https://swissbib.blogspot.com/2020/07/a-machine-learning-approach-with.html
- source code: https://github.com/swissbib/clustering_metadata/tree/capstone_project_andreas_jud_epfl_final_defense

Vorteil des Verbundsystems

- Standardisierung durch Vorgabe des Schemas
- Nur validierte Daten werden in das System geschrieben
- Keine Mehrdeutigkeit bei Attributen (Keys)
- Werte zumindest rudimentär geprüft (Integer, String, usw.)

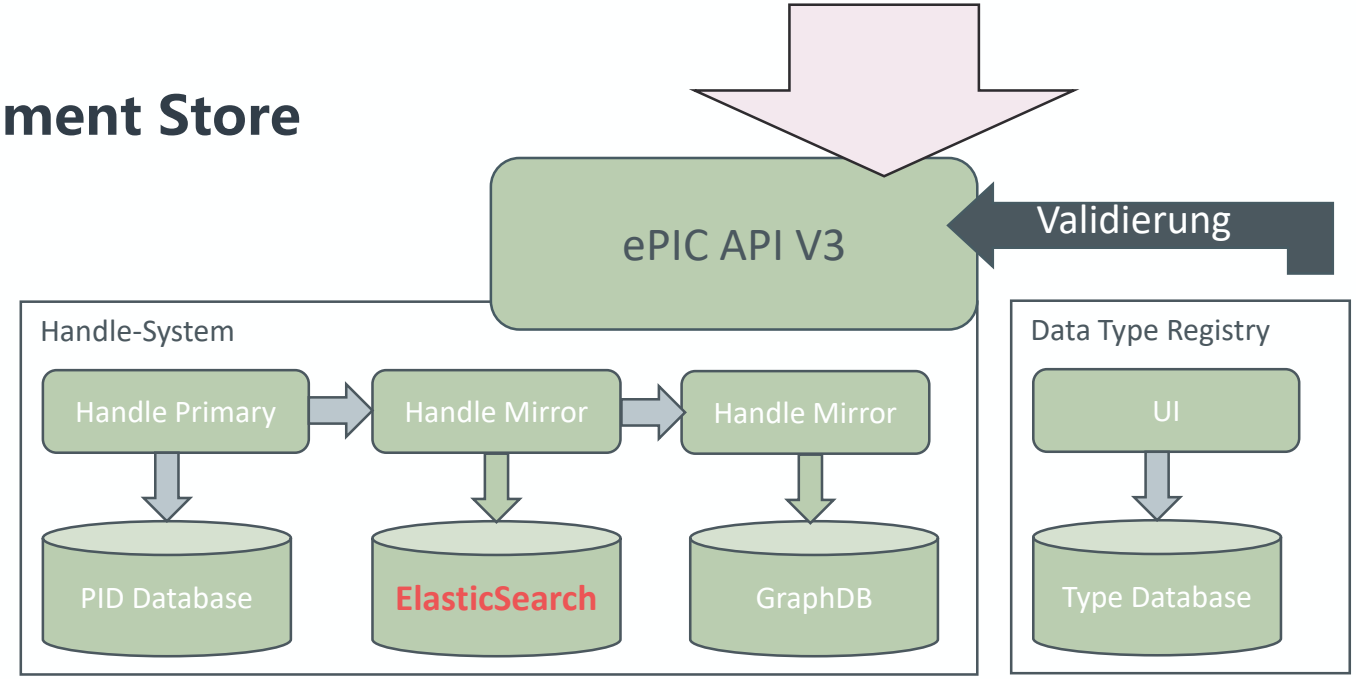


Vorverarbeitung von Werten

Alle Werte die als Text (String) gekennzeichnet sind.

- Stopwortlisten https://github.com/solariz/german_stopwords
- Clusterverfahren: Objekte mit ähnlichen Eigenschaften zusammen fassen
- und viele andere NLP Machine Learning Möglichkeiten

Document Store



ElasticSearch enthält schon viele Funktionalitäten:

- Text Embedding
- Text Similarity
- Text Classification
- Named entity recognition
- Fill mask
- More like this query
- …

https://www.elastic.co/guide/en/machine-learning/current/ml-nlp.html

Nächste Schritte

- Ersten Datenimport um die verschiedenen Verfahren zu testen

AVefi

MEHR INFORMATIONEN

https://wiki.tib.eu/confluence/pages/viewpage.action?pageId=257984794

Kontaktdaten
Sven Bingert: sven.bingert@gwdg.de
Elias Oltmanns: elias.oltmanns@gwdg.de