Ein Blick hinter die Kulissen der Aggregation

Gregor Pirgie & Lukas Graf

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Inhalt

1. Was ist Aggregation?
2. Wie aggregieren wir?
   - A) Pipelines und Komponenten
   - B) Vorstellung edmlib
3. Learnings
4. Ausblick

Inhalt

1. Was ist Aggregation?
2. Wie aggregieren wir?
   - A) Pipelines und Komponenten
   - B) Vorstellung edmlib
3. Learnings
4. Ausblick

kulturpool

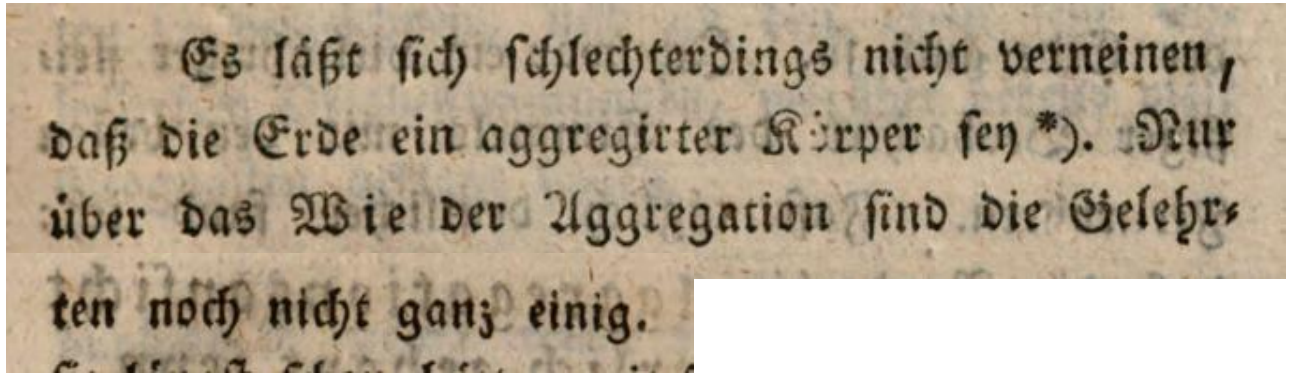
Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Was ist Aggregation?

- Abrufen und Zusammenführen von externen Daten, zur gemeinsamen Verarbeitung
- Einstiegspunkt der externen Daten in unser System
- Aufgaben der Aggregation
  - Vereinheitlichung
  - Sicherung Datenqualität
  - Nachhaltigkeit


Es läßt sich schlechterdings nicht verneinen, daß die Erde ein aggregirter Körper sey *). Nur über das Wie der Aggregation sind die Gelehrten noch nicht ganz einig.

Es läßt sich schlechterdings nicht verneinen, daß die Erde ein aggregirter Körper sey *).
Nur über das Wie der Aggregation sind die Gelehrten noch nicht ganz einig.

„Gedanken und Ansichten über die Ursachen der Erdbeben nach der Aggregations-Theorie der Erde“, Franz von Paula Gruithuisen, Seiten 5-6. o. D. Kulturpool.at. Kulturpool. Zugegriffen 14. Oktober 2024. https://kulturpool.at/institutionen/oenb/%252BZ18607510X

kulturpool
Bundesministerium Kunst, Kultur, öffentlicher Dienst und Sport
Finanziert von der Europäischen Union NextGenerationEU
naturhistorisches museum wien nhm

Aggregation im Kulturpool

- Ausgangslage
  - Viele Institutionen
    - Verschiedene Schnittstellen
    - Unterschiedliche Datenformate
  - Jede Anbindung einzigartig

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Es war klar, dass die wahren Probleme sich erst im weiteren Prozess zeigen würden.

Lukas Graf, Ende 2023

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Inhalt

1. Was ist Aggregation?
2. Wie aggregieren wir?
   - A) Pipelines und Komponenten
   - B) Vorstellung edmlib
3. Learnings
4. Ausblick

„Bau der Pipeline am Bodenseeufer“. o. D. Kulturpool.at. Kulturpool. Zugegriffen 14. Oktober 2024. https://kulturpool.at/institutionen/vorarlberger-landesbibliothek/o%253A258708.

kulturpool

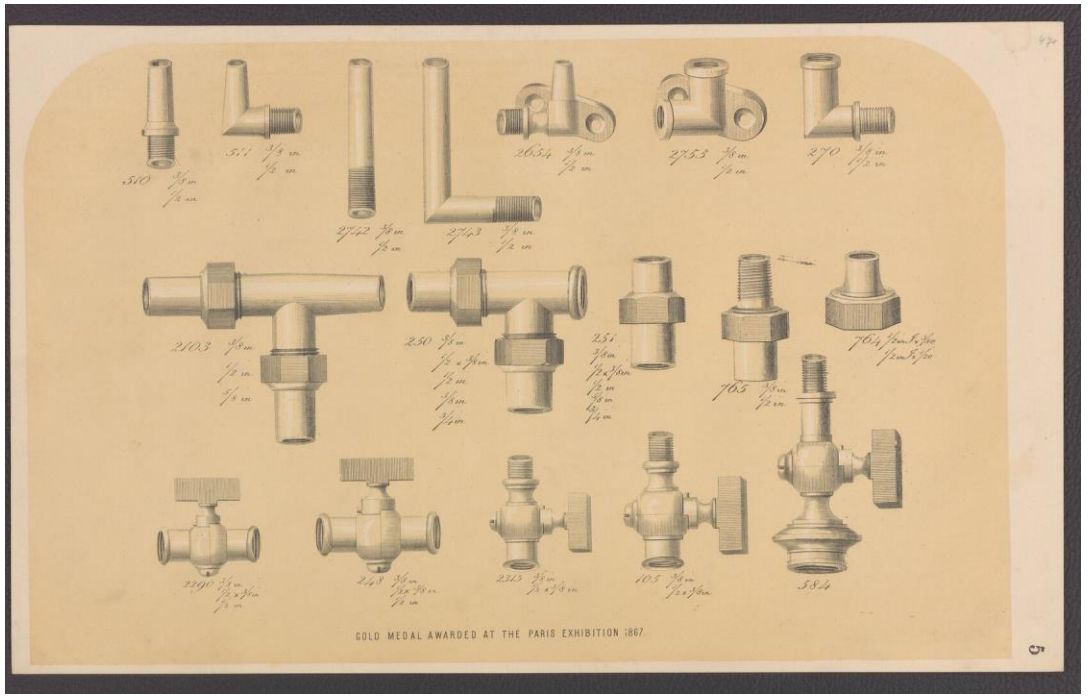
Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Technische Anbindung

- Pipeline
  - Besteht aus Phasen
  - Speziell für jede Anbindung
- Komponenten
  - Anpassbar
  - Wiederverwendbar



„Rohrverbindungen und Hähne für Gasleitungen“. o. D. Kulturpool.at. Kulturpool. Zugegriffen 12. Oktober 2024. https://kulturpool.at/institutionen/mak-wien/KI%252016888-5.

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Harvesting und Splitting

Harvesting:

Laden der Daten von externen Schnittstellen

Splitting:

Extrahieren einzelner Records aus den Daten

kulturpool

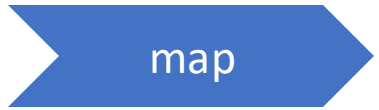
Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Mapping

- Umwandlung der Daten aus dem Ausgangsformat in EDM
- Grundmappingskripte für manche Standards (MARC 21, METS/MODS)
- Sonst basierend auf Mappingtabelle



kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Enrichment

- Erweiterungen und Verbesserungen der EDM-Daten
- De-Referenzierung
  - Abfragen von referenzierten Normdaten
  - Ergänzung der Daten um die dortigen Informationen
- In Zukunft weitere Enrichments möglich

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Validierung

- Die EDM Records werden validiert
  - Sind URLs (syntaktisch) valide
  - Sind alle Pflichtfelder des EDM enthalten
  - …
- Invalide Records werden zur weiteren Prüfung gesammelt

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Ingest

- Records werden in unser weiteres System eingespielt
  - Unabhängiger Service zur Analyse von Mediendaten
  - Typesense für Suchportal
  - Zukünftig APIs
- Um weitere Verwendung zu erleichtern Umwandlung in Zwischendarstellung "framed JSON-LD"

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Man schreibt nie zweimal die gleiche Anbindung. Jede Anbindung ist einzigartig.

Heraklit, o.D.

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Modulare Pipelines

Pipeline A: OAI-PMH Schnittstelle mit MARC21-Daten

Pipeline B: ZIP-File mit EDM-Daten

Pipeline C: REST-Schnittstelle mit Daten in eigenem Format

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Beispiel einer Pipelinedefinition

```python
class InstitutionPipeline(AggregationPipelineBase):
    def get_phases(self) -> list[BasePhase]:
        return [
            HarvestingPhase(
                harvesters=[
                    OAIHarvester(
                        api_url=settings.API_URL,
                        metadata_prefix=settings.OAI_METADATA_PREFIX,
                        sets=["ulbtirolhdsstams"],
                    )
                ],
            ),
            SplitOAIRecordsPhaseLxml(harvest_name=self.harvest_date),
            EDMPythonMappingPhase(
                processing_function=ULBMetsModsMapper.process_record,
                arguments={"data_provider": "Universitäts- und Landesbibliothek Tirol"},
            ),
            ValidationPhase(
                validators=[
                    edm_validator_factory(),
                ],
            ),
            EDMToJSONLDPhase(save_to=self.get_ending_persistence()),
        ]
```

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Inhalt

1. Was ist Aggregation?
2. Wie aggregieren wir?
   - A) Pipelines und Komponenten
   - B) Vorstellung edmlib
3. Learnings
4. Ausblick

B) edmlib

Bildet das EDM in Python ab

Erlaubt:

- Records direkt in Python zu schreiben
- Records von XML in Python Objekte einzulesen
- Records von Python Objekten als xml, json, turtle, json-ld, usw. auszuspielen
- Validiert dabei gegen alle Vorgaben des EDM

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

edmlib

- Kommt überall zum Einsatz, wo Records transformiert werden
- Kann leicht geupdatet werden
- Kann Teile von Pipelines ersetzen
- Dadurch, dass Records zu Python Objekten werden, hat man alle Möglichkeiten einer Programmiersprache

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Aggregation

Pipelines: Ermöglichen und Regeln die Befüllung des Kulturpool

edmlib: Funktion einer Schleuse – Kontrolliert das Wie des Zu- und Abflusses

„[Silvrettasee gegen Hohes Rad - Piz Buin - Silvrettahorn und Klostertaler Egghorn / Vorarlberg]“. o. D. Kulturpool.at. Kulturpool. Zugegriffen 16. Oktober 2024. https://kulturpool.at/institutionen/vorarlberger-landesbibliothek/o%253A42818.

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Inhalt

1. Was ist Aggregation?
2. Wie aggregieren wir?
   - A) Pipelines und Komponenten
   - B) Vorstellung edmlib
3. Learnings
4. Ausblick

Learnings

1. Planung wichtig, aber Herausforderungen werden oft erst im Verlauf klar
2. Trotz Standards: jede Anbindung ist einzigartig
3. Für Mapping Kommunikation und Datenexpertise wichtig
4. Schnittstellen/Daten verändern sich
5. Anbindungen im laufenden Betrieb stetig verbessern

„Hatsuhana 初花“. o. D. Kulturpool.at. Kulturpool.
Zugegriffen 12. Oktober 2024.
https://kulturpool.at/institutionen/mak-wien/KI%25207628-34.

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Ausblick

„Bauarbeiter auf Baustelle“. o. D. Kulturpool.at. Kulturpool. Zugegriffen 16. Oktober 2024. https://kulturpool.at/institutionen/vorarlberger-landesbibliothek/o%253A359528.

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm

Vielen Dank für die Aufmerksamkeit

kulturpool

Bundesministerium
Kunst, Kultur,
öffentlicher Dienst und Sport

Finanziert von der
Europäischen Union
NextGenerationEU

naturhistorisches
museum wien
nhm