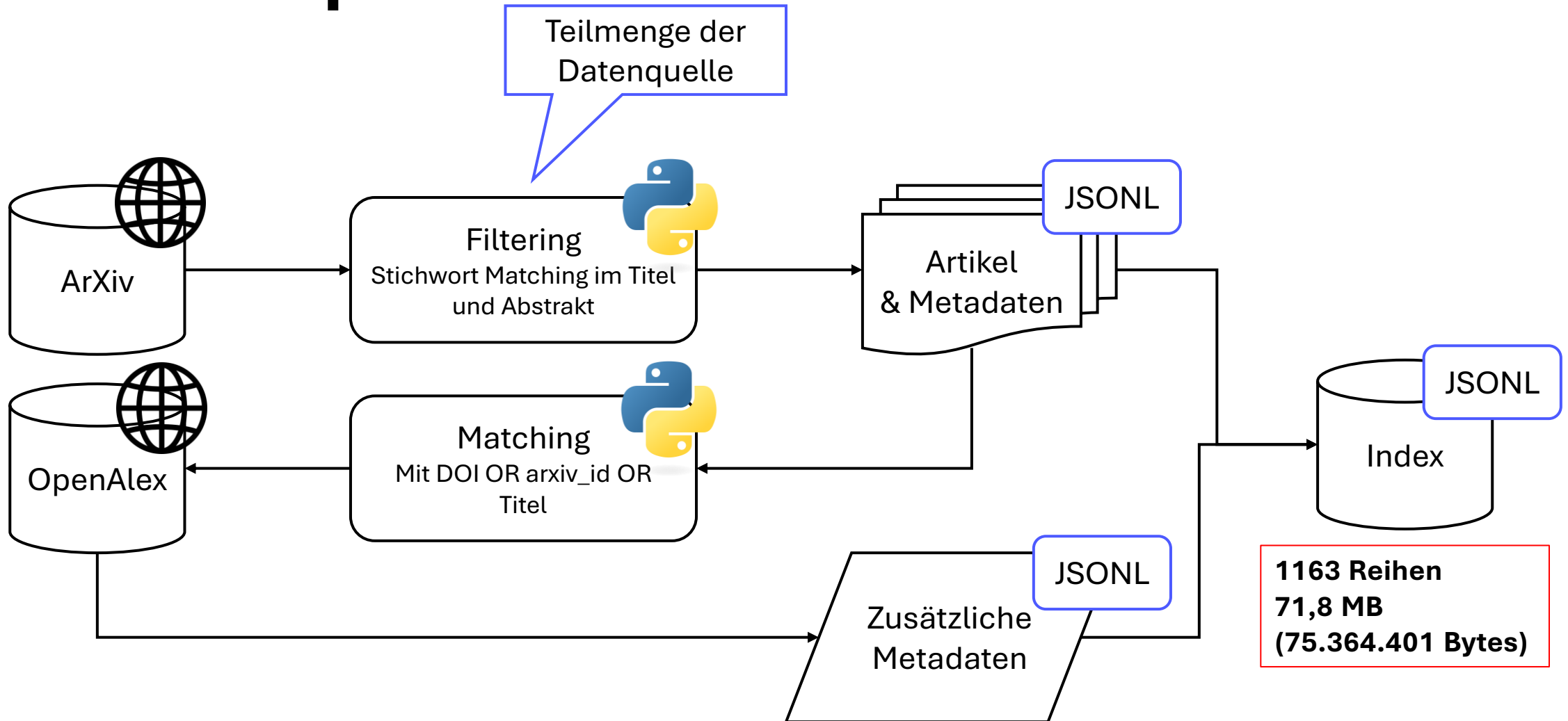# ORBIT

Aktueller Stand

Patryk Gadziomski

Hochschule der Medien

03.07.2025

Datenakquise

Datenbereinigung

Spalten entfernen:
- Link
- Authors
- Category
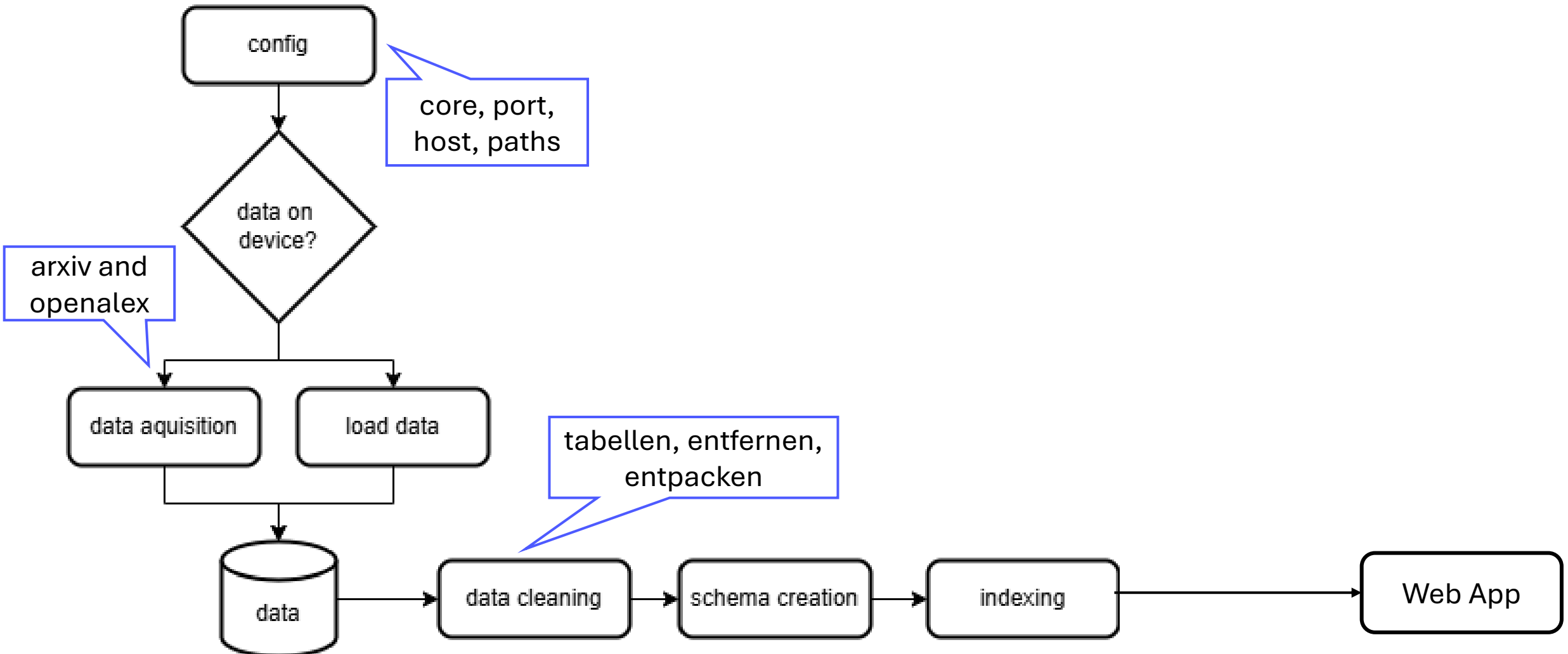- Host venue
- Host venue license
- Mesh
- Grants
- Is paratext

Volltext bereinigen:
- Tabellen
- Multiple Leerzeichen
- Ewige Zahlenreihen

Multiple Werte (Kategorie, etc.)
- Vernünftig speichern (V1)

Indexierung

```python
FIELDS = [
    {"name": "arxiv_id", "type": "string", "stored": True, "indexed": True, "required": True},
    {"name": "openalex_id", "type": "string", "stored": True, "indexed": True, "required": True},
    {"name": "doi", "type": "string", "stored": True, "indexed": True, "required": False},
    {"name": "title", "type": "text_general", "stored": True, "indexed": True},
    {"name": "summary", "type": "text_general", "stored": True, "indexed": True, "multiValued": True},
    {"name": "published", "type": "date", "stored": True, "indexed": True, "docValues": True},
    {"name": "updated", "type": "date", "stored": True, "indexed": True, "docValues": True},
    {"name": "pdf_link", "type": "string", "stored": True, "indexed": True},
    {"name": "fulltext", "type": "text_custom", "stored": False, "indexed": True},
    {"name": "cited_by_count", "type": "pint", "stored": True, "indexed": False, "docValues": True},
    {"name": "type", "type": "string", "stored": True, "indexed": True, "docValues": True},
    {"name": "referenced_works", "type": "text_general", "stored": True, "indexed": True, "multiValued": True},
    {"name": "related_works", "type": "text_general", "stored": True, "indexed": True, "multiValued": True},
    {"name": "grants", "type": "string", "stored": True, "indexed": True, "docValues": True},
    {"name": "landing_page_url", "type": "text_general", "stored": True, "indexed": True},
    {"name": "author_orcid", "type": "text_general", "stored": True, "indexed": True, "multiValued": True},
    {"name": "volume", "type": "string", "stored": True, "indexed": True},
    {"name": "issue", "type": "string", "stored": True, "indexed": True},
    {"name": "first_page", "type": "string", "stored": True, "indexed": True},
    {"name": "last_page", "type": "string", "stored": True, "indexed": True},
    {"name": "journal", "type": "string", "stored": True, "indexed": True, "multiValued": "true", "docValues": True},
    {"name": "concepts", "type": "string", "stored": True, "indexed": True, "multiValued": True, "docValues": True},
    {"name": "author_names", "type": "string", "stored": True, "indexed": True, "multiValued": True, "docValues": True},
    {"name": "open_access", "type": "boolean", "stored": True, "indexed": False, "multiValued": True, "docValues": True},
    {"name": "language", "type": "string", "stored": True, "indexed": True, "multiValued": True, "docValues": True},
    {"name": "institutions", "type": "string", "stored": True, "indexed": True, "multiValued": True, "docValues": True},
    {"name": "keywords", "type": "string", "stored": True, "indexed": True, "multiValued": True, "docValues": True},
    {"name": "spellcheck_base", "type": "text_general", "stored": False, "indexed": True, "multiValued": True}
]
```

HOCHSCHULE
DER MEDIEN

Funktionen

- **Keyword search**

- **Wildcard search**: example*, *example, exam*le

- **Phrase search**: "AI" or "Explainable AI„

- **Faceted search**: Thehem, Open Access, Sprache, Journal, Typ
  - Facetten kombinierbar

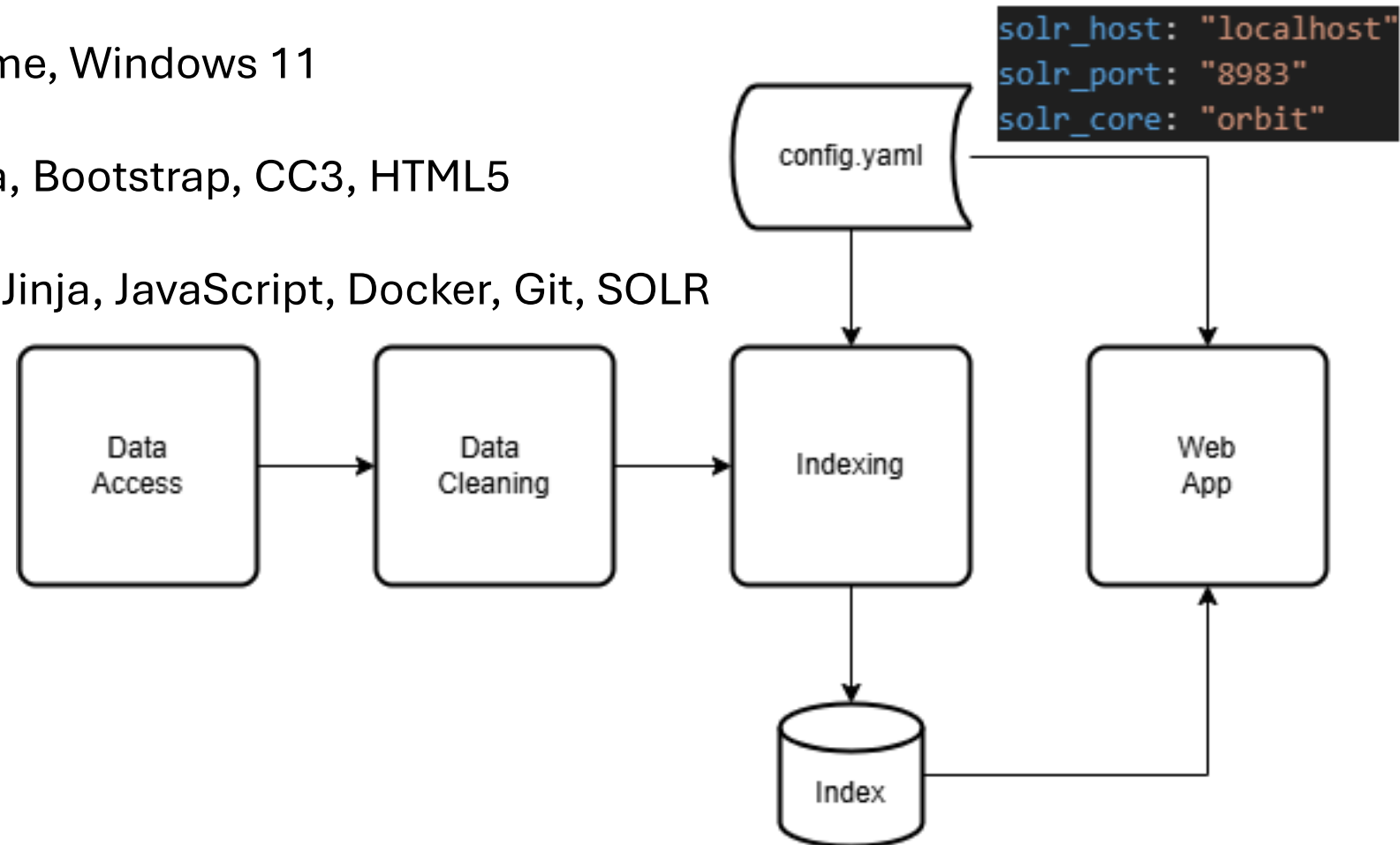- **Spellchecking**: Vorschläge bei keinen Treffern (wird weiterentwickelt)

Übersicht

Tested on: Google Chrome, Windows 11

Frontend/Design: Canva, Bootstrap, CC3, HTML5

Backend: Python, Flask, Jinja, JavaScript, Docker, Git, SOLR

Docs: Markdown


```
solr_host: "localhost"
solr_port: "8983"
solr_core: "orbit"
```

Live Demo

Ausblick

- Ähnliche Dokumente Identifizieren

- Autovervollständigung

- Suchterme hervorheben

- Spellchecking weiter ausbauen

- Frontend

- Dokumentation

ORBIT

Danke!

Patryk Gadziomski

Hochschule der Medien

03.07.2025