TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

Auf dem Weg zu einer zentralen Infrastruktur zur Bereitstellung strukturierter Texte für Text- und Data-Mining

Jens Freund
Universitäts- und Landesbibliothek Darmstadt

06.10.2026

Die Präsentation ist, mit Ausnahme der darin enthaltenen Logos und Symbole, lizenziert unter der Creative Commons Attribution 4.0 Lizenz.

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

Projekt Workflow Digitale Medien

(01.01.2023 – 30.06.2026)

Arten der Literaturnutzung

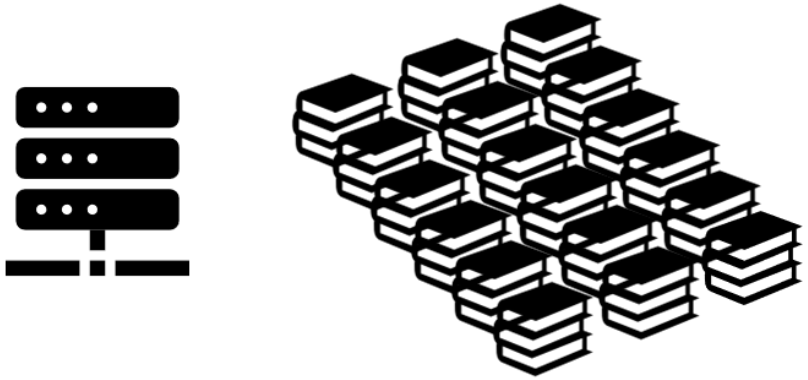
TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

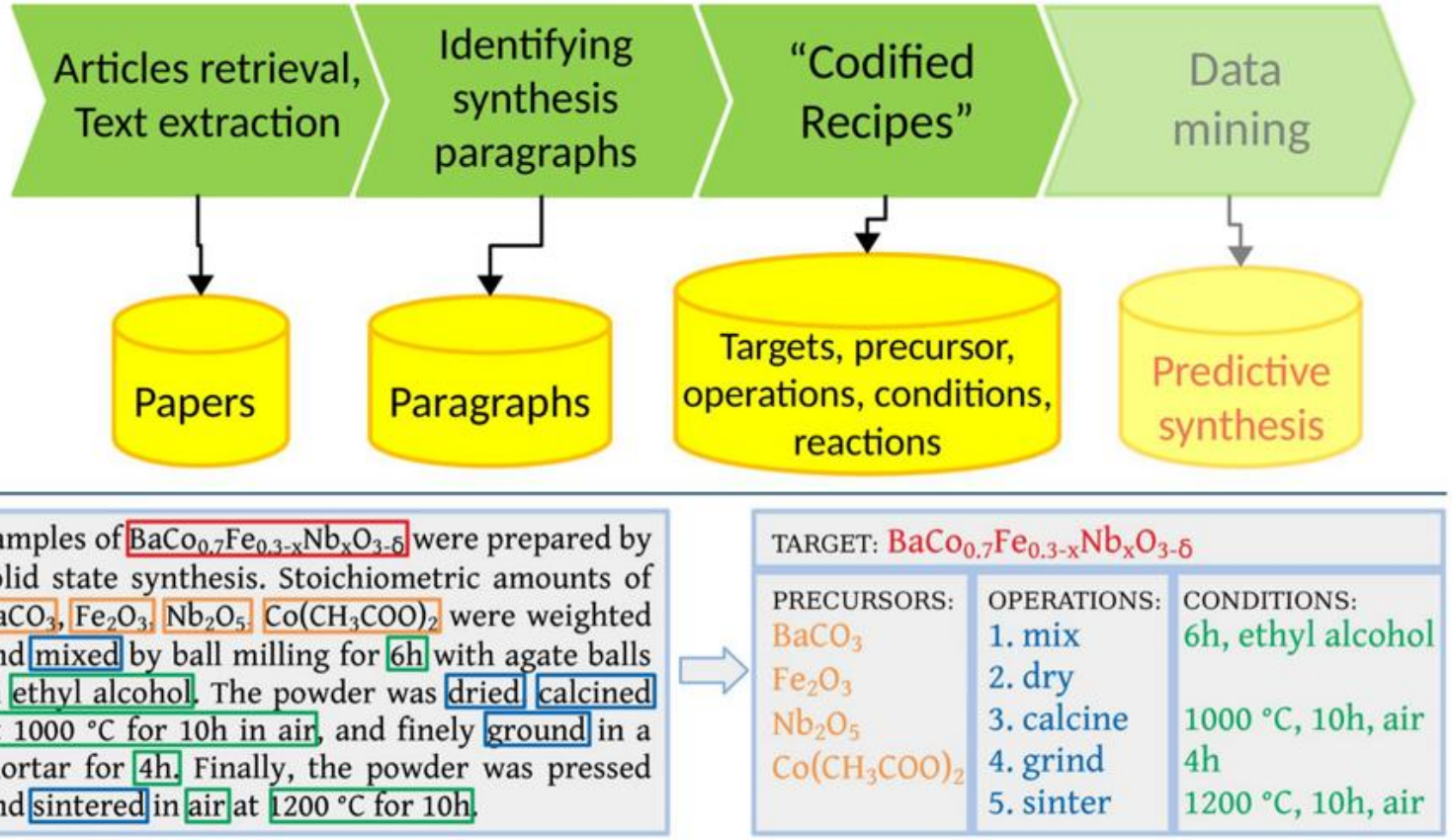
Individuelle Lektüre



Text- und Data-Mining (TDM)



TDM-Anwendungsbeispiel



Kononova, O., Huo, H., He, T. *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Sci Data* **6**, 203 (2019). https://doi.org/10.1038/s41597-019-0224-1

Verschiedene Bereitstellungswege der Verlage

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

Beispiele:



Einzelne Artikel über Webseiten, z. B. https://journals.openedition.org/jtei/



Zugriffsgeschützte FTP-Server, siehe z. B. https://www.mdpi.com/librarians



APIs, siehe z. B. https://dev.springernature.com/



Direkte Bulk-Downloads, siehe z. B. https://plos.org/text-and-data-mining/

Rechtliche Aspekte

Grundlegende Fragen

Besteht eine Erlaubnis für

- die **lokale Speicherung** und
- den **systematischen Download** der Dokumente?

Verschiedene Dokumentformate

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

Beispiele













Wissenschaftliche Qualität

- Sind die Informationen in den heruntergeladenen Dokumenten qualitativ hochwertig und vertrauenswürdig?
- Wurden die Dokumente z. B. einem wissenschaftlichen Begutachtungsverfahren unterzogen, um das sicherzustellen?

Projekt „Workflow Digitale Medien“ – WDM

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

- Ziel des Projekts war die Entwicklung eines prototypischen und leicht adaptierbaren Workflows, mit dem Bibliotheken digitale Forschungspublikationen aggregieren, aufbereiten und schließlich Forschenden in einem einheitlichen TEI/XML-Format für Text- und Data-Mining-Analysen zur Verfügung stellen können. (TEI = Text Encoding Initiative)
- DFG-gefördert (Projektlaufzeit: 01.01.23 – 30.06.26)
- Grundprinzip: Harvesting von Daten (Metadaten und Volltexte inkl. Bilddateien und andere Supplemente) für das Eigenhosting
- Fokus auf Open-Access-Literatur
- Die im Projekt entwickelten Skripte / Konzepte stehen unter freien Lizenzen zur Verfügung.

Erwerbungsentscheidung

Erwerbungs-entscheidung

On Demand / Just in Time

Anschaffungswünsche von Forschenden.

Just in Case

Sammlungsprofil der ULB Darmstadt.

Retrodigitalisierung

Scannen bereits vorhandener, gedruckter Literatur und Konvertierung in das TEI-Zielformat.

Erwerbungsentscheidung der Fachreferentin / des Fachreferenten

Rechteprüfung, Verlagskontakt

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

Erwerbungs-entscheidung → Rechteprüfung, Verlagskontakt

Rechteprüfung

Steht die gewünschte Literatur unter OA-Lizenzen?

Verlagskontakt

Ist der Verlag bereit, die OA-Literatur in strukturierten Dateiformaten bereitzustellen?

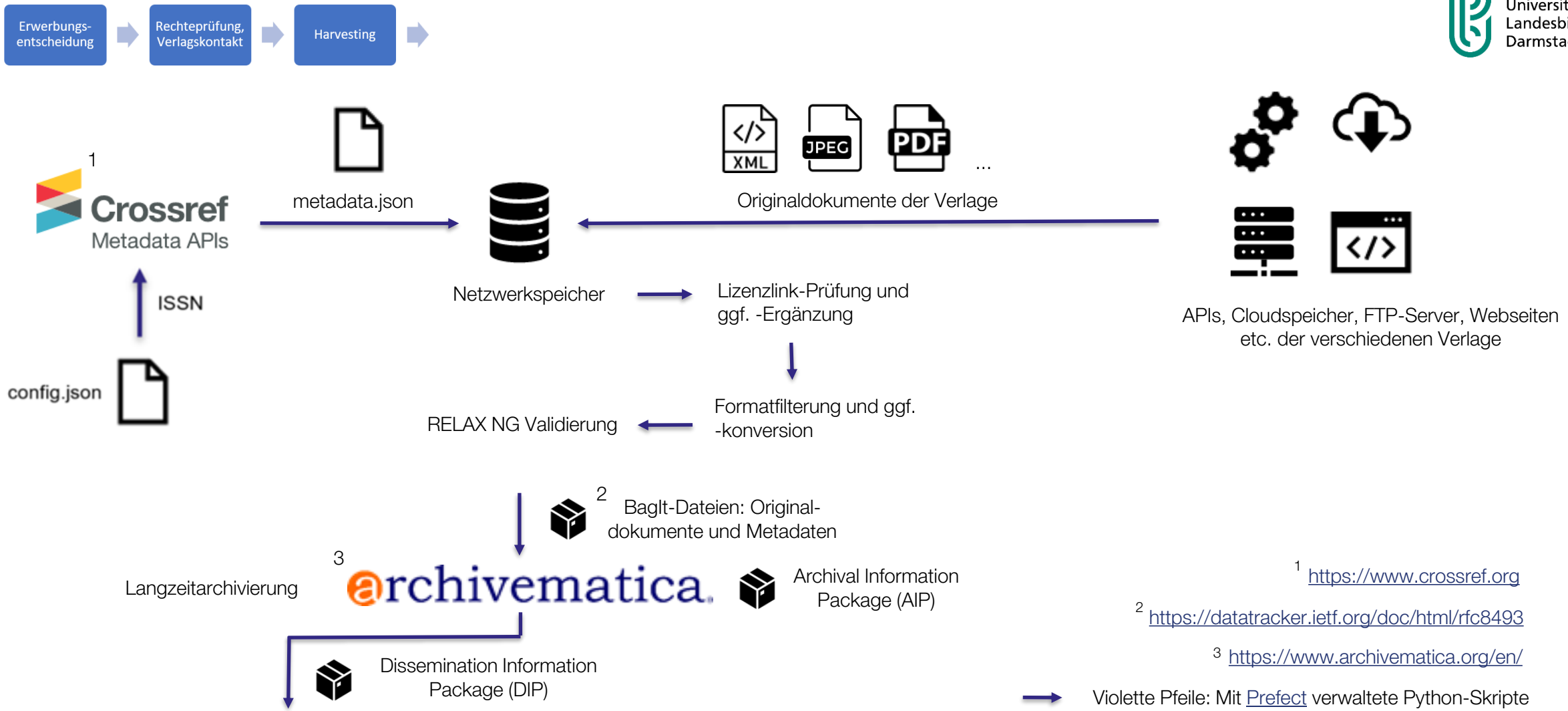
Bereits bestehende Vereinbarungen der ULB mit Verlagen zu OA-Literatur

- Copernicus
- Frontiers
- MDPI
- PLOS
- Springer Nature

Technischer Workflow

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt



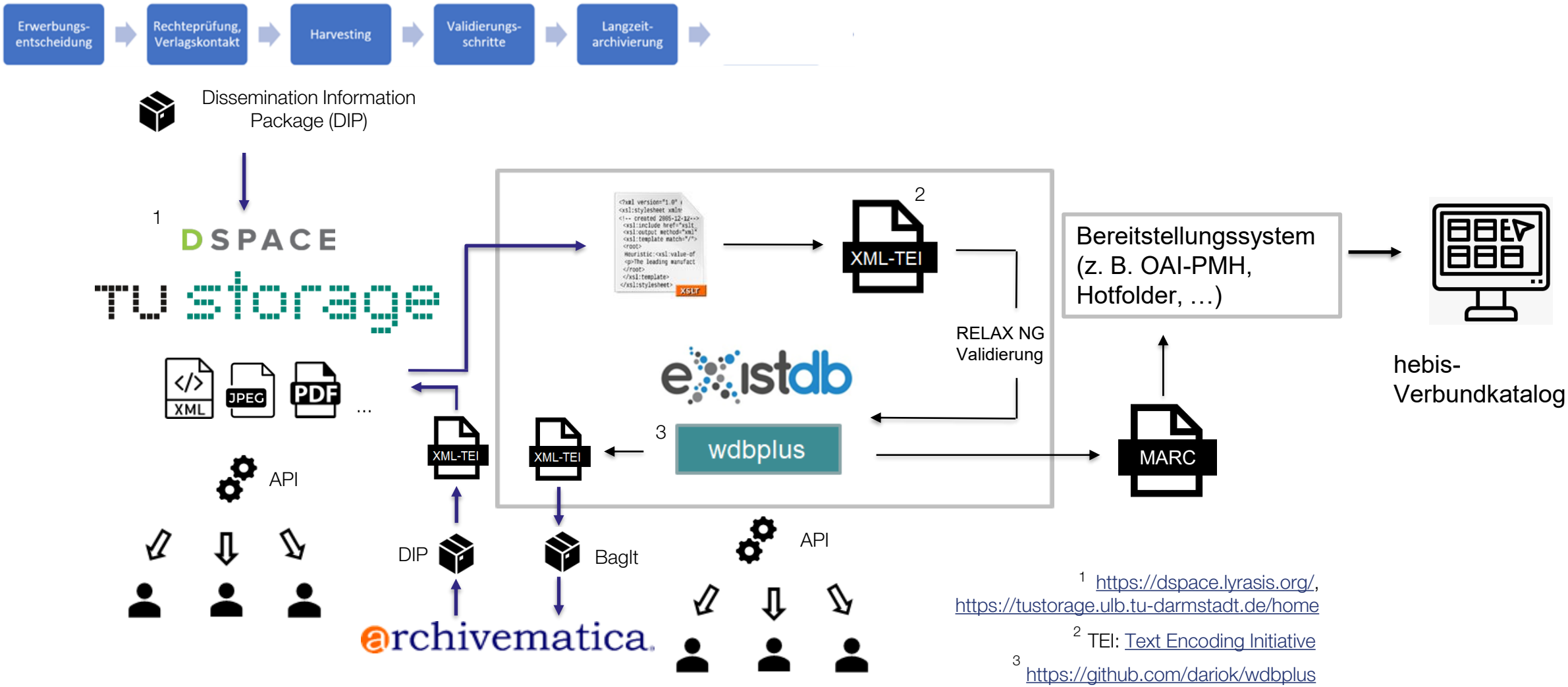
[1] https://www.crossref.org

[2] https://datatracker.ietf.org/doc/html/rfc8493

[3] https://www.archivematica.org/en/

→ Violette Pfeile: Mit Prefect verwaltete Python-Skripte

Technischer Workflow



[1] https://dspace.lyrasis.org/, https://tustorage.ulb.tu-darmstadt.de/home

[2] TEI: Text Encoding Initiative

[3] https://github.com/dariok/wdbplus

Bereitstellung von TEI-Dokumenten über eXist-db

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

- TEI-Beispielartikel in der eXist-Datenbank

```
−<tei:TEI xml:id="jz000091-0001">
  −<tei:teiHeader>
    −<tei:fileDesc>
      −<tei:titleStmt>
        −<tei:title>
           Ischemic-Hemorrhagic Stroke in New Onset Type 1 Diabetes Mellitus with Diabetic Ketoacidosis in a Two-Year-Old Toddler: The First Reported Case in Indonesia
         </tei:title>
        −<tei:author ref="#d2e53">
           <tei:persName>Halim, Anita</tei:persName>
         </tei:author>
        −<tei:author ref="#d2e63" role="correspondingAuthor">
           <tei:persName>Harsono, Felicia</tei:persName>
         </tei:author>
        −<tei:author ref="#d2e79">
           <tei:persName>Orlin, Sisilia</tei:persName>
         </tei:author>
        −<tei:author ref="#d2e92">
           <tei:persName>Octavius, Gilbert Sterling</tei:persName>
         </tei:author>
        −<tei:editor ref="#d2e109">
           <tei:persName>Lunevicius, Raimundas</tei:persName>
         </tei:editor>
       </tei:titleStmt>
```

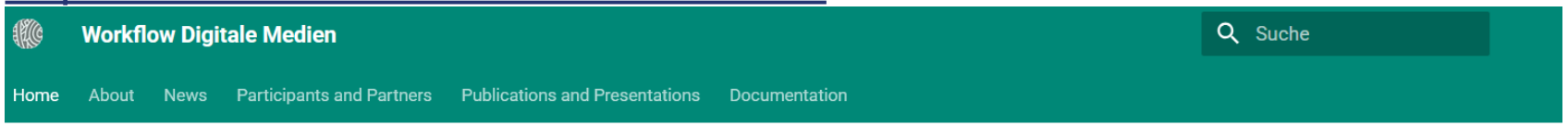
Beispielartikel: https://exist.ulb.tu-darmstadt.de/2/g/jz000091-0001

WDM-Webseite mit Workflow-Dokumentation

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

https://docs.ulb.tu-darmstadt.de/wdm/



Digital Media Workflow

An innovative concept for building digital collections to implement text and data mining methods.

The Digital Media Workflow (WDM) enables libraries to harvest open-access literature from academic publishers, convert it into a machine-readable, uniformly structured XML format based on the Text Encoding Initiative (TEI), and provide it to researchers for text and data mining analyses. The workflow combines technical components with organizational procedures. It is open-source and intended to be reusable by other infrastructure institutions.

For Infrastructure Institutions

If you would like to implement the WDM at your institution to harvest, process, archive and provide digital publications in XML/TEI format, start with the Documentation. It is currently under construction and will explain the overall workflow, the required systems and the relation between organizational and technical work steps.

Institutions interested in exchange around the follow-up project Digital Media Workflow II (WDM II) and a possible participation in a collaborative harvesting process can join the WDM community mailing list.

For Researchers and API Users



Hinweis: Die Seite befindet sich noch im Aufbau und wird kontinuierlich aktualisiert und erweitert.

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

Projekt Workflow Digitale Medien II: Kooperative Dateninfrastrukturen für KI und TDM

(seit 01.07.2026)

WDM-II-Projekt

- Gemeinsames DFG-Projekt der SUB Göttingen und ULB Darmstadt
- Projektstart: 1. Juli 2026
- Laufzeit: 36 Monate

Produktvision

Gemeinsam schaffen wir eine überregionale, offene Infrastruktur, die wissenschaftliche Literatur für TDM inklusive KI-Anwendungen einfach, verlässlich und nachhaltig nutzbar macht.



TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

Hauptziele von WDM II

01	WDM-Hub	• Zentral gehostete Infrastruktur zur Bereitstellung von TEI/XML-Dokumenten, der darin verlinkten Bilddateien und idealerweise weiterer Supplemente wie Videos. • Zunächst Implementierung nur für OA-Literatur.
02	Community-Building	• Aufbau einer Community aus Infrastruktureinrichtungen, die gemeinsam wissenschaftliche Literatur harvesten, ins TEI/XML- Format konvertieren und in den WDM-Hub einspielen möchten.
03	Lizenzpflichtige Literatur	• Verhandlungen mit Verlagen zur Bereitstellung lizenzpflichtiger Literatur für TDM-Analysen und Entwicklung einer Handreichung zur Rechteklärung. • Prüfung, ob WDM-Hub und WDM-I-Bereitstellungssysteme auf lizenzpflichtige Literatur erweitert werden können. • Entwicklung von Ad-hoc-Geschäftsgängen zur Literaturbereitstellung für individuelle Forschungsprojekte.

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

2029 vorgesehen (Auswahl)

In einer zentralen Infrastruktur werden in großem Umfang strukturierte OA-Texte angeboten.

Rasches Wachstum der Inhalte auf der Basis von koordiniertem und kooperativem Harvesting.

Zentrale und einrichtungsspezifische Schulungen können dank umfangreichen Materialien einfach angeboten werden.

Bibliotheksmitarbeitende sparen sich aufgrund der Handreichung viel Arbeit bei der Rechteklärung.

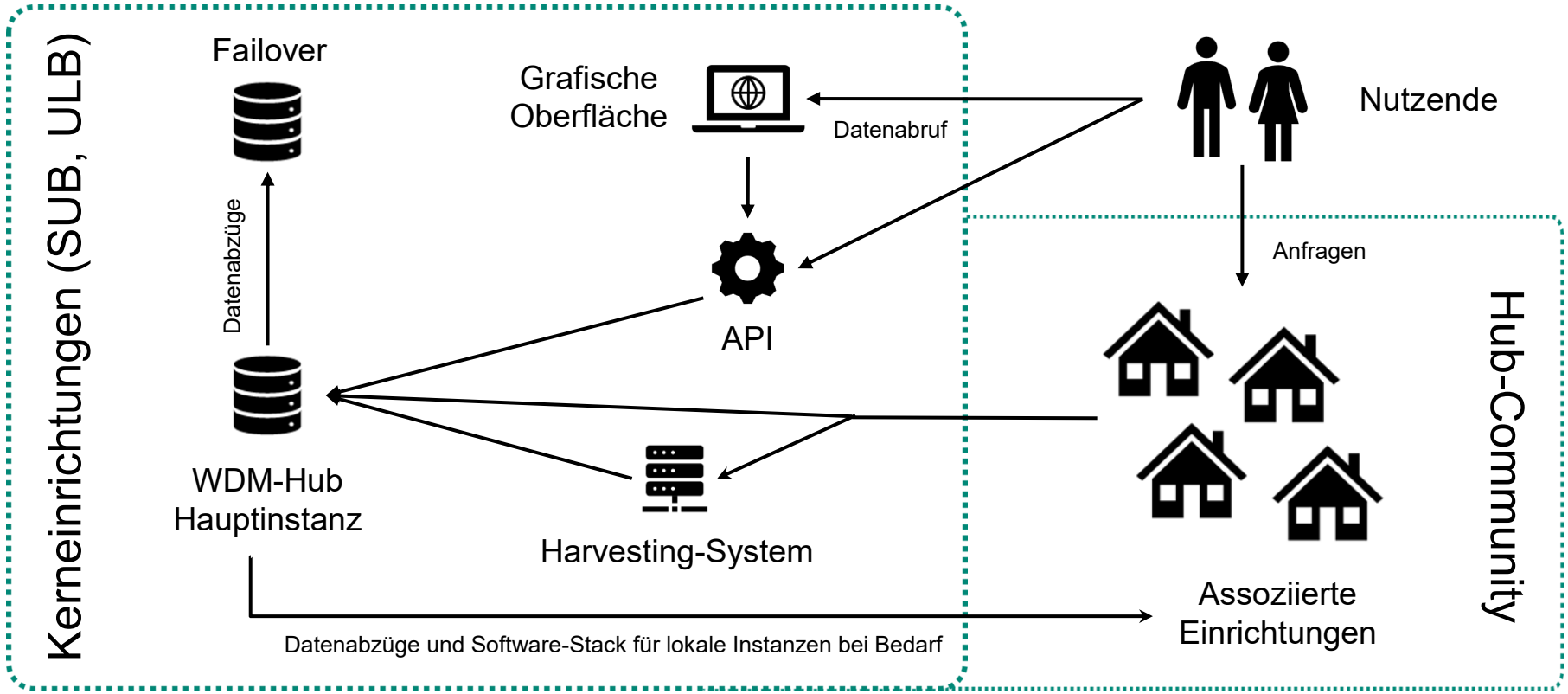
Forschende können sich über das Hub-Frontend und die API einfach OA-Textkorpora zusammenstellen.

Forschende können ihre erstellten Korpora einfach referenzieren und zur Nachnutzung anbieten.

Nutzende können direkt in der EZB sehen, ob Daten im Hub vorhanden sind.

Ein Modell für einen dauerhaft gemeinsam finanzierten Betrieb existiert und wird ausgerollt.

Geplante WDM-II-Infrastruktur



Community Building

Schritt 1: Mailingliste

- Sie können gerne beitreten unter https://www.listserv.dfn.de/sympa/info/wdm-community

… und viele weitere Schritte folgen noch …

TECHNISCHE UNIVERSITÄT DARMSTADT

Universitäts- und Landesbibliothek Darmstadt

Vielen Dank für Ihre Aufmerksamkeit!

Kontaktdaten Projektleitung

Angela Hammer
Universitäts- und Landesbibliothek Darmstadt
Magdalenenstr. 8
64289 Darmstadt
Gebäude S1|20, Raum 309
+49 6151 16-76204, -76431
angela.hammer@tu-darmstadt.de

E-Mail an das Projektteam (ULB Darmstadt)

wdm@ulb.tu-darmstadt.de

Projekt-Website:
https://docs.ulb.tu-darmstadt.de/wdm/

Gefördert durch

DFG Deutsche Forschungsgemeinschaft

Projektnummer WDM: 505256794
Projektnummer WDM II: 576318340