

# Korpus 4.0 - Ein innovativer Workflow zur Erstellung eines Korpus wissenschaftlicher Texte

## Kalmer, Silke

silke.kalmer@tu-darmstadt.de  
Universitäts- und Landesbibliothek Darmstadt,  
Deutschland  
ORCID: 0009-0006-2292-058X

## Freund, Jens

jens.freund@tu-darmstadt.de  
Universitäts- und Landesbibliothek Darmstadt,  
Deutschland  
ORCID: 0000-0001-6232-7568

## Hammer, Angela

angela.hammer@tu-darmstadt.de  
Universitäts- und Landesbibliothek Darmstadt,  
Deutschland  
ORCID: 0000-0002-8711-3221

## Geißner, Andreas

andreas.geissner@tu-darmstadt.de  
Universitäts- und Landesbibliothek Darmstadt,  
Deutschland  
ORCID: 0000-0002-6996-4671

## Kampkaspar, Dario

dario.kampkaspar@tu-darmstadt.de  
Universitäts- und Landesbibliothek Darmstadt,  
Deutschland  
ORCID: 0000-0002-0118-0811

## Einleitung

Digitale Textkorpora bilden eine wichtige Grundlage für viele Gebiete der Digital Humanities, sei es als Forschungsgrundlage oder für das Testen und Anwenden von Analysewerkzeugen (Schöch et al., 2020). Der Aufbau solcher Korpora ist aber mitunter langwierig (Westergaard et al., 2018) und das Bereitstellen aufwändig. Die Bibliothek als traditioneller Ort zur Versorgung mit Literatur bietet sich für einen Workflow zur Erstellung sowie zur Bereitstellung von solchen Korpora an.

Im Rahmen des DFG-geförderten Projekts Workflow Digitale Medien<sup>1</sup> arbeitet die Universitäts- und Landesbibliothek Darmstadt (ULB) daran, ein Korpus aus wissenschaftlicher Open Access (OA) Literatur in einem einheitlich strukturierten XML-Format über frei zugängliche Schnittstellen bereitzustellen. Forschende sollen die Möglichkeit erhalten, alle Arten wissenschaftlicher Dokumente, zum Beispiel Zeitschriftenartikel, E-Books oder Konferenzbände, in großen Mengen über diese Schnittstellen abzurufen, um sie anschließend beispielsweise für Text- und Data-Mining-Analysen nutzen zu können. Zusätzlich zu den Texten werden alle für den Workflow entwickelten Konzepte und Skripte Open Source zur Verfügung gestellt, um deren Nachnutzbarkeit zu gewährleisten. Im vorliegenden Vortrag<sup>2</sup> wird ein Überblick über den Workflow, die Datengrundlage und Zugriffsmöglichkeiten sowie den aktuellen Entwicklungsstand des Projekts gegeben.

## Überblick und Ziele

In der ersten Projektphase konzentriert sich das Projekt unter anderem auf die Bereitstellung von OA-Zeitschriftenartikeln. Die Artikel werden von verschiedenen Verlagen aggregiert, sofern möglich unmittelbar in ein einheitliches XML-Format auf der Grundlage des Formats der Text Encoding Initiative (TEI Consortium, 2024) konvertiert und über Schnittstellen öffentlich abrufbar gemacht. Wann immer möglich, werden neben den eigentlichen Texten auch Begleitmaterialien, insbesondere die Abbildungen, mit aggregiert, um sie später ebenfalls bereitzustellen. Sämtliche im Rahmen des Projektes entwickelten Skripte und Tools werden ebenfalls unter freien Lizenzen veröffentlicht.

Für Forschende sollen damit zukünftig die zeitaufwändigen Schritte des Dokument-Harvestings von unterschiedlichen Verlagsplattformen, die rechtliche Prüfung zur Nutzbarkeit der Dokumente und deren Homogenisierung in ein einheitliches Dateiformat entfallen, die häufig als Vorbereitungsschritte wissenschaftlicher Text- und Data-Mining-Projekte notwendig sind. In Absprache mit einem Advisory Board werden dabei die Anforderungen von Forschenden besonders mit berücksichtigt. Das Advisory Board besteht aus 11 Forschenden und Mitgliedern anderer Infrastruktureinrichtungen (Wissenschaftliche Bibliotheken, Universitäten, Hessisches Zentrum für Künstliche Intelligenz (hessian.AI), GESIS – Leibniz-Institut für Sozialwissenschaften) und berät das WDM-Projekt seit November 2023 zu allen Entwicklungen. Dies stellt nicht nur die Nutzung der Inhalte durch Forschende sicher, sondern auch die Nachnutzbarkeit der entwickelten Konzepte und Skripte durch andere Bibliotheken. Die Treffen mit dem Advisory Board lieferten bereits wertvolle Anregungen, beispielsweise zur Auswahl der Texte für das Korpus sowie zu deren Langzeitarchivierung und Bereitstellung.

Durch das Advisory Board und einen öffentlichen Workshop des Projekts haben sich Kontakte zu anderen Bibliotheken in Deutschland und der Schweiz ergeben, die an

Projekten arbeiten, aus deren einzelnen Aspekten sich Synergien zum WDM-Projekt ergeben könnten.

Perspektivisch ist geplant, gemeinsam mit anderen Bibliotheken eine arbeitsteilige Sammlungs- und Bereitstellungsinfrastruktur für OA-Literatur aufzubauen und Lösungen zu finden, wie darüber hinaus auch lizenzpflichtiges Material für berechnete Forschende der betreffenden Einrichtungen bereitgestellt werden könnte.

## Workflow

Ausgangspunkt des in Abbildung 1 schematisch dargestellten Workflows ist die Entscheidung darüber, welche Literatur bereitgestellt werden soll. Die ULB verfolgt dabei zwei Ansätze: On Demand (Anschaffungswünsche von Forschenden) und Just in Case (Erwerbungsentscheidungen der Bibliothek auf der Grundlage ihres Sammelprofils). Sobald auf diese Weise relevante Literatur identifiziert wurde, erfolgt im zweiten Schritt eine Prüfung, ob die Publikationen lizenzrechtlich im Rahmen des Workflows verarbeitet und bereitgestellt werden dürfen. Häufig geht dies mit einer Anfrage bei den betreffenden Verlagen einher, ob die Publikationen durch den Verlag bereits in einem XML-Format mit den zugehörigen Abbildungen kontinuierlich an die ULB geliefert oder, beispielsweise über eine API, zugänglich gemacht werden können. Hier konnte die ULB bereits mit mehreren Verlagen Vereinbarungen treffen.

Der Abruf der Dokumente auf die Speichersysteme der ULB erfolgt dann mittels selbst entwickelter Python-Harvesting-Skripte. Für die Konversion der Verlags-XML-Dokumente in das TEI/XML-Zielformat werden im Projekt entwickelte XSLT-Skripte genutzt. Vor der eigentlichen Konversion erfolgt eine Validierung, mit der geprüft wird, ob die Verlags-XML-Dokumente durch das Stylesheet verarbeitet werden können oder ob das Skript zuvor angepasst werden muss bzw. die Dokumente bei verlagsseitigen Fehlern an diesen zur Korrektur zurückgegeben werden müssen.

Nach erfolgreicher Validierung werden die Verlagsdokumente gemäß dem Open Archival Information System (OAIS) Modell (CCSDS, 2012) zunächst in Form sogenannter Archival Information Packages langzeitarchiviert. Dazu kommt die Open-Source-Software Archivematica<sup>3</sup> zum Einsatz. Anschließend werden Kopien der Daten aus Archivematica entnommen und in das DSpace<sup>4</sup>-Dokumentenrepositorium TUSTorage<sup>5</sup> eingespielt, das als Speicherort für die ursprünglichen Verlagsdokumente dient (Verlags-XML- und PDF-Dokumente sowie Bilddateien). Von dort werden die XML-Dokumente zur Konversion ins TEI/XML-Format in eine eXist-XML-Datenbank<sup>6</sup> kopiert und über Schnittstellen für den automatisierten Abruf durch Forschende zur Verfügung gestellt.

Während die bisher beschriebenen Workflow-Schritte bereits prototypisch implementiert wurden, befinden sich die folgenden, in Abbildung 1 hellblau dargestellten Schritte noch in der Entwicklung. Alle über den Workflow bereitgestellten Dokumente sollen durch automatisierte Metada-

teneinspielungen im hebis-Verbundkatalog<sup>7</sup> nachgewiesen und damit einfach auffindbar gemacht werden. Neben der Bereitstellung über die eXist-Datenbank sollen die konvertierten TEI/XML-Dokumente ebenfalls langzeitarchiviert und, zusätzlich zur Bereitstellung über die eXist-Datenbank, auch über TUSTorage bereitgestellt werden.



Schematischer Workflow von einer Erwerbungsentscheidung der ULB bis zur Zugänglichmachung der Literatur im TEI/XML-Format. Workflowschritte in Dunkelblau funktionieren bereits prototypisch, solche in Hellblau befinden sich noch in der Entwicklung.

## Datengrundlage und Basisformat

Die wissenschaftlichen Texte, die die Grundlage des Korpus bilden, sind aus einem breit gefächerten Spektrum. Inhaltlich werden unterschiedliche Wissenschaften und Wissenschaftsbereiche abgedeckt, es gibt keine thematische Fokussierung. Dadurch werden Forschungsgebiete sowohl aus den MINT-Fächern als auch aus den Geistes- und Sozialwissenschaften gesammelt. Bei den Textsorten handelt es sich um Journalartikel, Monographien, Konferenzbände und Dissertationen.

Die Eingangstexte liegen sowohl in unterschiedlichen Dateiformaten, wie XML, Word, PDF etc. vor, als auch in verschiedenen semantischen Auszeichnungen wie JATS<sup>8</sup>, BITS<sup>9</sup> oder TEI. Aufgrund dieser heterogenen Datenformate ist das Ziel, sie in ein einheitliches Zielformat zu überführen. Zu diesem Zweck wird ein Basisformat entwickelt, welches auf TEI P5 basiert. TEI/XML bietet sich als Zielformat an, da es sowohl menschen- als auch maschinenlesbar und -auswertbar ist. Zudem ist es, anders als zum Beispiel JATS, das eine XML-Auszeichnungssprache für Journalartikel ist, nicht auf eine bestimmte Textsorte beschränkt. Ein weiterer Vorteil ist, dass mathematische Formeln, die mittels MathML<sup>10</sup> kodiert worden sind, exakt übernommen werden können, so dass es zu keinem Datenverlust kommt. Dieses Basisformat soll darüber hinaus nicht nur für den Workflow, sondern übergreifend für andere Texte in der ULB, wie etwa digitale Editionen, genutzt und entsprechend ergänzt werden, je nach den Bedürfnissen der jeweils anderen Textsorten. Momentan werden diverse Konvertierungsskripte für JATS-Formate, die je nach Verlag andere Ausformungen haben können, zum TEI-Basisformat erstellt.

Hauptaugenmerk beim Basisformat liegt dabei auf dem TEI-Header, der die Metadaten des Textes enthält und sie einheitlich und bibliothekskonform abbilden soll. Die Textinhalte werden, wie sie sind, übernommen. Bei Bedarf können im späteren Verlauf eigene Annotationen hinzugefügt werden. Zur Abbildung der Entitäten, zunächst nur aus den Metadaten des Textes entnommen, wie etwa die Au-

tor:innen und Editor:innen des Textes in Form einer list-Person, wird ein TEI-Standoff verwendet, das in weiteren Schritten, wie etwa bei Named Entity Recognition auf den Textinhalt beliebig erweiterbar wäre.

## Zugriffsmöglichkeiten

Bevor die Konvertierung stattfindet, werden die Eingangsdateien mit eigens entwickelten Schemadateien im RelaxNG-Format geprüft und validiert. Erst nach der erfolgreichen Validierung der Eingangsdateien kann die Konvertierung ins TEI/XML-Format erfolgen. Abgelegt werden die resultierenden XML-Dateien in eXist-db. Hier sind sie über REST-Schnittstellen abrufbar, sowohl als einzelne Datei, etwa ein Journalartikel, als auch als Collection, die den Überblick über zum Beispiel die Liste der Artikel in einem Journal ausgibt. Dadurch ist das Zusammenstellen einer eigenen Textsammlung bzw. eines Unterkorpus aus dem Korpus möglich. Ein Update der seit dem letzten Abruf geänderten Dateien ist über die Schnittstellen ebenfalls möglich, so dass bereits zusammengestellte Textsammlungen aktuell gehalten werden können.

Daneben sind die Texte durch eine Konvertierung ins HTML-Format als Lesetext auf der ULB-eigenen Plattform TUeditions im Open Access zu finden. Sie sind dabei in das Framework wdbplus (Kampkaspar, 2024) zur Anzeige eingebettet, welche unter anderem auch eine Volltextsuche anbietet. Die Ansicht, exemplarisch in Abbildung 2 gezeigt, ist so geteilt, dass sich auf der einen Seite der reine Lesetext befindet und auf der anderen die Übersicht über Informationen des Textes in Reitern. In einem Reiter sind die Metadaten des Textes, insbesondere die bibliographischen Angaben, in einem zweiten ist die Anzeige der Fußnoten des Textes und in einem weiteren Reiter ist das Inhaltsverzeichnis bzw. die Übersicht der Kapitel zu finden. Das XML-Format ermöglicht darüber hinaus auch eine schnelle Generierung ins EPUB-Format, wodurch dieses Format ebenfalls heruntergeladen werden können.

so dass jeder Text im Korpus auch durch die Katalogsuche der Bibliothek auffindbar sein wird.

## Zusammenfassung und Ausblick

Mit dem Projekt Workflow Digitale Medien verfolgt die ULB das Ziel, Forschenden wissenschaftliche OA-Literatur in einem einheitlich strukturierten TEI/XML-Format über Schnittstellen zum automatisierten Download bereitzustellen. Auf diese Weise soll der bislang zeitaufwändige Prozess der Korpuserstellung für Forschende vereinfacht werden. Während oben der aktuelle Stand der Arbeiten beschrieben wurde, existieren bereits Pläne für weitere Schritte.

Neben der Konvertierung von JATS ins Basisformat sind momentan die Konvertierung von anderen Eingangsformaten, wie etwa andere TEI-Formate, die nicht dem Basisformat entsprechen, und anderen Dateiformaten wie Word, in Planung. Außerdem sind Weiterentwicklungen des Basisformats, welches zurzeit hauptsächlich wissenschaftliche Artikel abdeckt, in Vorbereitung.

Ein weiteres, übergeordnetes Ziel des Projekts ist, dass sich auch andere Bibliotheken dem Vorhaben anschließen, um ebenfalls OA-Literatur zu harvesten und im Basisformat bereitzustellen. Durch die unterschiedlichen fachlichen Sammelschwerpunkte der einzelnen Bibliotheken kann sich idealerweise eine Arbeitsteilung ergeben, bei der jede Bibliothek die für ihre Nutzenden relevante Literatur aufbereitet, die dann aufgrund der offenen Lizenzen des Ursprungsmaterials weltweit frei zur Verfügung gestellt werden kann. Bereits in der laufenden ersten Förderperiode wird zu diesem Zweck Wert auf die Nachnutzbarkeit der an der ULB entwickelten Konzepte und Skripte gelegt, beispielsweise durch deren geplante Veröffentlichung unter freien Lizenzen und die Nutzung von Open-Source-Software wie Archivematica, DSpace und eXist-db. In einer möglichen zweiten Förderperiode könnte die kooperative Bereitstellung technisch und organisatorisch durch einen Hub vertieft werden, einen Service, über den die geharvesteten Datenbestände der einzelnen Einrichtungen für die Nutzenden zentral zugänglich gemacht werden.

Idealerweise soll über den Workflow zukünftig auch lizenzpflichtige Literatur bereitgestellt werden können. Daraus ergeben sich zwei Herausforderungen: Zum einen müssen Lizenzvereinbarungen mit Verlagen geschlossen werden, die die Bereitstellung der Literatur für den automatisierten Download durch Forschende erlauben. Teilweise ist dies an der ULB bereits erfolgt, auch wenn der Fokus in der laufenden Förderperiode auf der Bereitstellung von OA-Literatur liegt. Zum anderen muss ein Rechte- und Rollenmanagement in den Bereitstellungssystemen implementiert werden, mit dem sichergestellt werden kann, dass nur berechtigte Personen auf die Daten zugreifen können. Aufgrund der komplexen Rechtesituation soll, im Gegensatz zum oben beschriebenen Hub für OA-Literatur, hier kein zentrales Bereitstellungssystem entwickelt werden, sondern vielmehr eine prototypische Lösung, die dezentral von



Beispielanzeige eines Textes in TUEditions mit ausgewähltem Reiter „Metadaten“

Die Nachnutzbarkeit der XML-Dateien wird zudem dazu verwendet, um daraus MARC-21-XML-Dateien zu generieren. Diese sollen perspektivisch zur Katalogisierung der Texte an die hebis-Verbundzentrale weitergeleitet werden.

anderen Bibliotheken an ihren jeweiligen Standorten nachgenutzt werden kann, um die von ihnen lizenzierte Literatur an die Forschenden ihrer Hochschulen ausliefern zu können.

## Danksagung

Das Projekt Workflow Digitale Medien wird von der Deutschen Forschungsgemeinschaft (DFG) seit 2022 unter der Projektnummer 505256794 gefördert.

## Fußnoten

1. [https://www.ulb.tu-darmstadt.de/forschen\\_publicieren/forschen/wdm.de.jsp](https://www.ulb.tu-darmstadt.de/forschen_publicieren/forschen/wdm.de.jsp) (zugegriffen: 24. Juli 2024)
2. Contributor Roles: Silke Kalmer (Conceptualization, Software, Writing – original draft), Jens Freund (Conceptualization, Software, Writing – original draft), Angela Hammer (Conceptualization, Supervision, Writing – review & editing), Andreas Geißner (Conceptualization, Software, Writing – review & editing), Dario Kampkaspar (Conceptualization, Software, Visualization, Writing – review & editing)
3. <https://www.archivematica.org/en/> (zugegriffen: 24. Juli 2024)
4. <https://dspace.lyrasis.org/> (zugegriffen: 24. Juli 2024)
5. <https://tustorage.ulb.tu-darmstadt.de/home> (zugegriffen: 24. Juli 2024)
6. <http://exist-db.org/exist/apps/homepage/index.html> (zugegriffen: 24. Juli 2024)
7. <https://www.hebis.de/dienste/katalog/> (zugegriffen: 24. Juli 2024)
8. Journal Article Tag Suite, <https://jats.nlm.nih.gov/> (zugegriffen: 24. Juli 2024)
9. Book Interchange Tag Set, <https://jats.nlm.nih.gov/extensions/bits/> (zugegriffen: 24. Juli 2024)
10. <https://www.w3.org/Math/> (zugegriffen: 24. Juli 2024)

## Bibliographie

**CCSDS.** 2012. “Reference Model for an Open Archival Information System (OAIS). Recommended Practice. CCSDS 650.0-M-2. Magenta Book”. <https://public.ccsds.org/Pubs/650x0m2.pdf> (zugegriffen: 24. Juli 2024).

**"Kampkaspar, Dario.** 2024. “W. Digitale Bibliothek (wdbplus).” [Computer software]. <https://github.com/dariok/wdbplus> (zugegriffen: 24. Juli 2024).

**Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, Jörg Röpke.** 2020: “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen.” In *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel. DOI: 10.17175/2020\_006.

**TEI Consortium.** 2024. “TEI P5: Guidelines for Electronic Text Encoding and Interchange.” Version 4.8.0. <http://www.tei-c.org> (zugegriffen: 22. Juli 2024).

**Westergaard, David, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, Søren Brunakl.** 2018. “A comprehensive and quantitative comparison of text-mining in 15 million full-textarticles versus their corresponding abstracts.” In *PLoS Computational Biology* 14 (2), e1005962. DOI: 10.1371/journal.pcbi.1005962.