Communication Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 145–152 DOI: 10.15439/2025F0064
ISSN 2300-5963 ACSIS, Vol. 45

Towards a German VET Archive and its Integration into a Data Warehouse

Thomas Reiser* , Petra Steiner†, Kristine Hein†
* University of Koblenz, Germany,
Email: treiser@uni-koblenz.de
† Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany,
Email: steiner@bibb.de

***Abstract*—This paper presents a systematic evaluation and prototypical implementation of an information system for historical vocational education and training (VET) regulations in Germany. The focus of this study is on integrating structured outputs with the German Labor Market Ontology (GLMO) and a broader labor market data warehouse. A corpus of VET and CVET regulations, as published in the Federal Gazette from 1969 to 2022, was used to assess the functional and semantic requirements of the archival process. This analysis was complemented by a review of existing software frameworks, culminating in the proposal of a combined architecture utilizing Omeka S and TEI Publisher. In addition, the necessary transformations, metadata enrichment, and ETL processes required to integrate the resulting TEI XML documents into a semantically linked data environment are detailed. This work provides a concrete roadmap for the sustainable digitization and semantic integration of regulatory texts into modern labor market intelligence infrastructures.**

I. Introduction

VOCATIONAL education and training (VET) systems are of critical importance in maintaining a skilled workforce and supporting economic resilience. In Germany, a historically extensive corpus of VET and continuing VET (CVET) regulations has been published in the Federal Gazette over the course of several decades. These documents serve as the foundational elements of occupational standards and training frameworks, garnering substantial interest from researchers, policymakers, and labor market analysts. However, the archival form of these regulations as described in [1]—primarily as unstructured or semi-structured scanned documents—poses challenges for digital accessibility, analysis, and integration with contemporary data systems.

The digitization of archival material presents an opportunity to preserve, structure, and analyze regulatory knowledge in a form amenable to semantic linking, machine learning, and long-term data curation as discussed in our previous work [2], [3]. In this context, two fundamental questions emerge: first, which software tools and platforms are most suitable for the digitization, structuring, and management of historical training regulations, and second, how can the resulting digital records be semantically integrated into the German Labor Market Ontology (GLMO) and a broader data warehouse environment that supports longitudinal labor market research?

In order to address the aforementioned inquiries, the present document offers a technical design and evaluation of a digitization pipeline founded upon image preprocessing, optical character recognition (OCR), and TEI XML structuring. A comparative review of available archival systems is conducted, followed by the implementation of a dual-platform prototype using Omeka S for metadata management and TEI Publisher for structured transcript administration. Beyond the archival perspective, a methodology is proposed for mapping the structured training documents to historical occupation taxonomies. This methodology enables their integration into the GLMO and subsequent ingestion into a data warehouse through standard ETL procedures. The present study contributes to the development of interoperable digital infrastructures for vocational education and labor market data by addressing both archival and semantic integration concerns.

This study is organized as follows: The first section introduced our data set, consisting of records and scans of the occupations archive. Next, the research background is introduced to give an overview over research in the area of document digitization and information extraction. Then, we discuss related literature that addresses the existing solutions to similar problems, especially regarding the text structure recognition and the integration of such workflows into web applications. The fourth section introduces our methodology that aligns with the early phases of the software lifecycle, from analyzing the problem statement, requirement elicitation, and system design in order to implement a first prototype, that is presented afterwards. Finally, findings of strengths and shortcomings with this prototype are discussed, before summing up the paper and giving an outlook over future work.

II. Background

The digitization of historical documents has garnered significant interest in recent years, with a proliferation of methodologies to address this undertaking. Optical character recognition (OCR) is a foundational technology for digitization, with a significant research focus and a range of established tools.

A plethora of methodologies exists for the purpose of document digitization. One fundamental approach entails the detection of the complete text within the document images, as illustrated by the methodology employed in the Finnish newspaper digitization project[4]. In this project, the objective is to generate an ALTO XML document that encompasses all recognized text, leveraging the Tesseract OCR engine. An alternative approach is demonstrated in [5], wherein the

authors model the text structure of legal texts in Austria and align the recognized text to this predefined structure, thereby enhancing the structured recognition of text. More advanced methods employ the OCR results to construct structured data from the text images. The authors of [6] employ OCR to digitize invoices and to structure the recognized information, such as product description, quantity, and price. In a similar vein, the study by [7] involved the extraction of names of judges at German federal courts from 1950 to 2019. This was achieved by applying OCR to the Federal Gazette, a publication that contains the official gazette of the Federal Republic of Germany.

In order to optimize accessibility, a number of OCR workflows have been integrated into web-based applications [8]. A significant endeavor in this domain is OCR4all, which facilitates the implementation of diverse preprocessing steps, segmentation methodologies, and OCR models. Additionally, it facilitates interaction with each of the process steps, enabling users to make corrections at intermediate results and thereby improve the overall outcome. However, there are also less extensive tools that facilitate the management of digitized document collections. For instance, these tools can be found in [9], [10].

III. Related Literature

In previous works, we analyzed the data set of the occupations archive [1] to obtain an overview over the different types of documents which vary a lot in language- and layout-style. For a selected data set of documents from the Federal Republic Germany that have been available on the internet, a structure analysis has been conducted on TEI XML transcripts that were created by a rule-based transcription pipeline [3], [2]. Both text structure and content were analyzed to get a first overview over commonalities in the selected documents. As all of the documents were from the same period, their layout followed mostly similar patterns, even across multiple decades. However, an evolution of wording and structure over time was be observed. While the development of a more advanced approach for the occupations archive is underway, this pipeline served as a preliminary feasibility analysis.

In the initial approaches to document structuring that emerged in the 2000s, human knowledge about the document was employed to delineate text- and layout-based rules for extracting the text structure [11], [12], [13]. The initial iteration of the digitization workflow employed predefined rules; however, future research endeavors will prioritize the automatic recognition of patterns in layout and text features to facilitate the structured organization of texts with diverse layouts. As with [5], a specific structure is delineated to replicate the text's hierarchical arrangement.

Despite the existence of standard conversion tools, such as Vertopal, which facilitate the transformation of text files into markup languages, these tools operate under the assumption that the text contained within documents to be converted is accompanied by accurate structural information [14] like in born-digital documents. However, this cannot be assured through the utilization of default OCR (Optical Character Recognition) models which are important to extract text from image information. While these tools can be utilized to generate files in HTML or TEI XML, the resulting output files frequently fail to accurately represent the text hierarchy or logical document structure. Instead, these files offer an alternative representation of recognized text areas, lines, and text.

The utilization of machine learning models, akin to those employed in GROBID [15], facilitates the extraction of metadata elements such as title and author information. Additionally, it facilitates the recognition of references and citations, as well as the detection of the abstract. Despite its extensive array of useful features, the model was trained exclusively on scientific articles, resulting in its exceptional performance on this specific domain. In order to employ GROBID in legal documents such as the training regulations examined in this article, it would be necessary to refine the model, a process that would require training data. The efficacy of this approach is contingent upon the quality of page segmentation and the reliability of text recognition.

The target data format has been determined to be TEI XML, as it is endorsed by the German Research Foundation (DFG) as a suitable standard for the long-term archiving of documents, see [16]. While PDF is a proprietary standard that is stored in binary files, markup languages such as XML can be read by almost any computer without the need for additional software designed for reading PDF files. Moreover, these files can be efficiently stored in XML databases such as eXist-db to manage the document collection [17] . eXist-db is a versatile system that facilitates the incorporation of plugins, including a versioning plugin. This plugin enables memory-efficient storage of multiple versions of the same documents, facilitating swift restoration of older versions when necessary. This is particularly salient in the context of automated systems, where errors are to be anticipated. Another beneficial plugin is TEI Publisher, which facilitates the management of XML databases and enables the viewing of documents in a human-readable manner. It also allows for the editing of uploaded TEI XML files. Additionally, it enables the visualization of the text alongside the original images, thereby facilitating the error correction process. Consequently, TEI XML has been identified as an optimal target data format for the digitization of extensive text corpora.

IV. Methodology

A. Software Requirements

The software requirements for the digitization of the vocational training archive encompass a broad set of functional and non-functional requirements aimed at ensuring the accurate and efficient transformation of historical training regulations into structured, searchable digital formats. The archival system must possess the capability to ingest scanned documents in a variety of formats, in particular PDF, PNG, JPG, and TIFF, while also employing an AI pipeline to automatically generate TEI XML transcripts. The system must support the

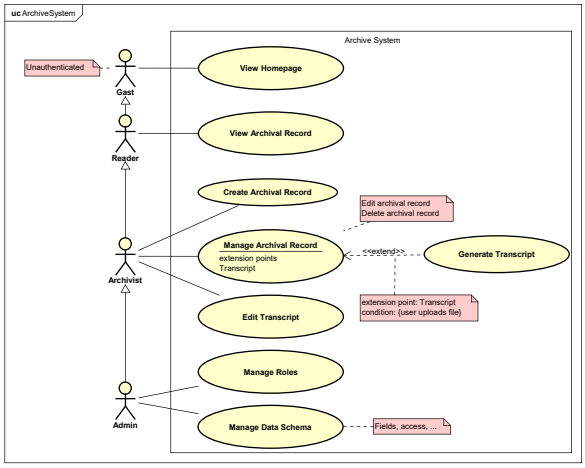

Fig. 1: Use Case Diagram for the Archive System

Name:	**Archiveintrag erstellen**
Beschreibung:	Ein Archivar kann Dokumente hochladen und mit den zugehörigen Metadaten annotieren. Das System speichert diese Informationen und erstellt automatisch ein Transkript im TEI XML Format.
Vorbedingung:	Auf dem Fileserver ist genug Speicherplatz für den hochzuladenden Scan verfügbar.
Nachbedingung:	Ein neuer Eintrag wurde im Archivsystem erstellt.
Fehlerfälle:	Der Transkriptionsworkflow kann den Upload nicht verarbeiten. Es ist nicht genügend Speicherplatz für den neuen Eintrag verfügbar.
Systemzustand im Fehlerfall:	Der Archivar erhält eine Meldung, dass der Upload nicht erfolgreich war.
Akteur:	Archivar
Auslöser:	Der Archivar hat ein Dokument gescannt und möchte es zum Archivsystem hinzufügen.
Standardprozess:	Neues Dokument speichern und Scan hinzufügen. 1. Der Archivar ruft die Seite zum Upload auf. 2. Der Archivar lädt den Scan hoch. 3. Das System erstellt ein Transkript für das hochgeladene Dokument. 4. Das System füllt die Felder für im Dokument enthaltene Metadaten mit automatisch generierten Vorschlägen aus. 5. Der Archivar korrigiert mögliche Fehler im Transkript oder den gefundenen Metadaten. 6. Der Archivar bestätigt seine Eingaben und der neue Eintrag wird im Archivsystem gespeichert. 7. Das System prüft, ob ein ähnlicher Archiveintrag bereits existiert und gibt dem Archivar die Möglichkeit, den Scan als neues Exemplar des Dokuments hinzuzufügen. 8. Das System speichert den neuen Eintrag und dass dieser überprüft wurde und korrekt sein sollte.
Alternativer Prozess:	Neues Dokument speichern und nicht sofort kontrollieren, ob alle Daten korrekt vom Workflow erkannt wurden. 1. Der Archivar ruft die Seite zum Upload auf. 2. Der Archivar lädt den Scan hoch. 3. Das System erstellt ein Transkript für das hochgeladene Dokument. 4. Das System füllt die Felder für im Dokument enthaltene Metadaten mit automatisch generierten Vorschlägen aus. 5. Der Archivar speichert den neuen Eintrag im Archivsystem. 6. Das System speichert den neuen Eintrag, sowie dass dieser noch nicht überprüft wurde und eventuell noch Fehler vorliegen.

Fig. 2: Scenario: Create Archival Record

management of metadata and version control, enable role-based access (i.e., guest, reader, archivist, administrator), and offer functions such as document upload, editing, and deletion. Furthermore, advanced features such as duplicate detection, exemplar linking, and prediction of metadata through machine learning are integral. The archival system under consideration should facilitate PDF export, full-text search, and integration with external tools such as TEI Publisher for transcript handling. Based on use case diagrams like shown in Figure 1, scenarios like in Figure 2 have been created to derive requirements for the information system.

B. Possible Software

A variety of open source software solutions were evaluated with the objective of meeting the requirements of the Archive project. A comprehensive overview of these solutions can be found in Table I. Because there are many already existing tools that should be able to solve the problem of designing an archival information system, from a maintenance perspective, it makes a lot sense to reuse these technologies instead of creating another one. The foundational platform for the archival

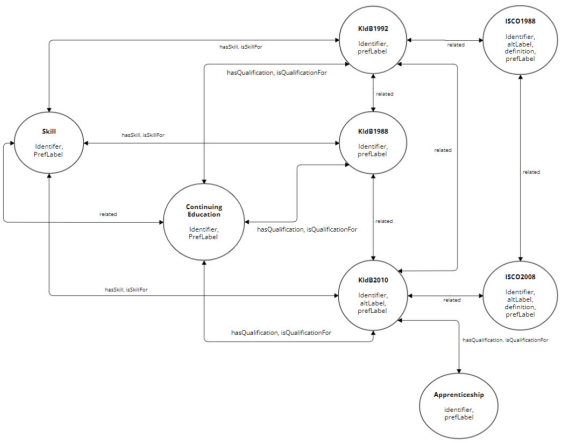

Fig. 3: Data Schema for historical KldB Data

system prototype is Omeka S, a digital archive framework that is modular and extensible. It offers role-based access control, ontology-based metadata management (e.g., Dublin Core), and integration capabilities via REST APIs. The architecture of this system is designed to support the organization of digital items and metadata-rich content, rendering it well-suited for general archival needs. However, a critical deficiency in Omeka S is its lack of native support for TEI XML transcriptions, a fundamental requirement for accurately representing the structure and hierarchy of historical training regulations. To address this, TEI Publisher was selected as a complementary tool, offering specialized support for managing, versioning, and displaying TEI-encoded documents. Constructed on the eXist-db platform, TEI Publisher facilitates seamless integration with Omeka S and offers a customized user interface for transcription workflows.

In addition to the primary tools, alternative systems such as Paperless-ngx and Access to Memory (AtoM) were also considered. Paperless-ngx is an open-source document management system designed for individual or small-scale organizational use, offering basic OCR and tagging features. While it demonstrates notable strengths in terms of simplicity and usability, it exhibits deficiencies in terms of flexibility and extensibility, which are crucial for effective performance in complex archival tasks and structured text processing. Conversely, AtoM is designed to align with international archival standards and offers a web-based interface for institutional repositories. Despite its strengths in terms of compliance, the software does not offer sufficient support for AI-driven transcription pipelines or integration with TEI XML workflows. Consequently, the integration of Omeka S and TEI Publisher was determined to be the optimal solution for meeting the technical and semantic criteria of the archival digitization initiative.

C. Integration into GLMO

To facilitate comprehensive longitudinal analyses of vocational development in Germany, the German Labor Market

TABLE I: Overview of possible Software Products for the VET Archive

Software	Type	Key Features and Notes
Omeka S	Web Publishing Platform	• Modular architecture, extensible with custom modules • Secure authentication (e.g., via LDAP) • Supports ontologies (e.g., Dublin Core) • REST API and CSV import • Lacks native TEI XML support
TEI Publisher	XML Management	• Manages TEI XML transcriptions • eXist-db plugin with built-in versioning • UI for transcript administration • Suitable for integration with Omeka S
Paperless-ngx	Document Management System	• Open-source DMS with OCR and tagging • Focused on personal or small-business document workflows • May be limited for complex archival needs
Access to Memory (AtoM)	Archival Description Tool	• Focus on archival standards (e.g., ISAD(G)) • Web-based interface for archival institutions • Less suited for tight integration with AI pipelines

Ontology (GLMO) is being extended with historical occupational taxonomies from both the Federal Republic of Germany (FRG) and the former German Democratic Republic (GDR). This ontological enrichment involves the alignment of legacy classification systems, such as KldB 1988 and KldB 1992, with more recent taxonomies, including KldB 2010 and ISCO-08, through a series of conversion mappings, see Figure 3. These mappings facilitate temporal integration and semantic linking of occupational entities across decades. The establishment of bidirectional links between historical training regulations—often preserved as scanned and TEI-encoded documents—and the corresponding occupation nodes within the GLMO knowledge graph is of particular relevance to the occupations archive project. The utilization of persistent identifiers and relation types, such as "hasSource" or "referencesClassification," facilitates the embedding of archival artifacts directly into the ontology-driven representation of the labor market. This integration facilitates enhanced contextualization of archival data and supports link prediction and graph reasoning tasks, enabling researchers to infer structural trends, skill transitions, and educational pathways over time.

Currently, labels from the genealogy of vocational education that describes the history of vocational training in Germany are used for the mapping to the GLMO. Because the regulations in the occupations archive build the legal foundation for the records in the genealogy as they describe the time periods where training occupations had state recognition in Germany, we expect that records in the occupations archive should have a one-to-one mapping with the genealogy records. To match these two data sets, we use record linkage methods as described in [18].

D. Integration into Data Warehouse

The integration of the occupations archive into a more extensive data warehouse for vocational education and labor market research introduces a series of technical and semantic challenges. The initial task entails the harmonization of TEI XML document structures with the classification systems that are already present in the data warehouse. These include the German Classification of Occupations (KldB) and the German Labor Market Ontology (GLMO). This mapping is designed to ensure semantic interoperability and facilitate meaningful linkage to other datasets based on occupation, time, or region.

A secondary requirement is the design of reliable ETL (Extract–Transform–Load) processes to convert document scans and their corresponding XML outputs into structured records suitable for ingestion. These transformations are required to normalize formats, extract metadata, and validate consistency across records. Concurrently, data protection considerations must be addressed. Given the potential for scanned documents to contain sensitive or personally identifiable information, compliance with the General Data Protection Regulation (GDPR) is imperative. This involves the development of anonymization procedures, incorporating pseudonymization or masking techniques, with the objective of preserving analytical value while ensuring the protection of personal data.

Moreover, the historical nature of the documents necessitates the implementation of temporal modeling. Regulations must be indexed not only by document metadata but also by their effective periods, including enactment and expiration dates. This temporal axis facilitates longitudinal analysis and compatibility with existing time-based labor statistics. Another essential integration step is the enrichment of documents with metadata that might not be present in the original scans. The employment of natural language processing methodologies is imperative for the inference of document types, occupational domains, and geographic scope from the content.

Subsequent integration steps involve validation against reference datasets. Cross-referencing transcribed documents with existing administrative data, such as DAZUBI or QuBe, ensures data quality and enables automatic categorization. Prior to implementation in business intelligence (BI) environments

or dashboards, the content must undergo plausibility checks, formatting consistency tests, and, if necessary, additional aggregation or filtering. These processes must be meticulously documented, with version control and user access management implemented to ensure both reproducibility and secure data access for researchers and stakeholders.

E. DFG requirements

A plethora of methodologies pertain to the process of document digitisation. As outlined in the DFG Practical Guidelines on Digitisation [19], a comprehensive compendium on the execution of such a digitisation initiative is available. Digital preservation is an undertaking that requires meticulous planning from a multitude of perspectives. The initial step in this process typically involves the acquisition of an overview of the existing resources, encompassing personnel, financial, and material aspects. Furthermore, it encompasses the consideration of potential objections pertaining to damages that may arise during the digitisation process. Subsequent to this, a suitable scanning technology and method must be selected. The choice of format and quality of paper can have a significant impact on the final result. In the case of document images, the target format should be TIFF uncompressed, as this is a lossless format that has been in existence for a considerable time and is widely accepted by the majority of archival endeavours. These original images are also referred to as 'digital masters', and it is imperative that they undergo minimal post-processing to ensure that the integrity of the original information is preserved.

The generation of various derivatives is possible from this digital master. A derivative could for example be a copy of the image in JPG or PNG format, in black and white, or with reduced noise. With regard to more efficient web delivery, a PDF file can be utilised, in which the different scans are combined into a single document, potentially with embedded text.

To guarantee interoperability with other archives, one or more standards for the metadata needs to be selected. Commonly used standards include the Dublin Core Metadata Initiative (DCMI) terms, and the Europeana Data Model (EDM) which are represented as linked data. Other established formats include the METS/MODS XML format, ISAD(G), and others that are not mentioned here. It is common practice to store metadata in multiple formats so that the data becomes better interoperale with other archives that might use different formats and also for better interoperability with the Open Archive Initiative Protocoal for Metadata Harvesting (OAI-PMH). The DFG guidelines recommend METS/MODS in particular, but states that formats like DCMI are also a suitable alternative.

The key challenges associated with the scanning process include the potential for damage to documents during scanning, water damage, brittle paper, faded ink, and various other external factors. It is imperative to ensure that the original resources are not destroyed during the process of digitisation.

V. Presentation of Prototype

As outlined in Section IV-B, the majority of the tools mentioned therein are considered to be valid choices for the archive system. Indeed, combinations of different tools are easily possible. Archivematica, for instance, is equipped with a functionality that integrates it into AtoM with minimal effort. The microservice architectures of Archivematica and Islandora facilitate the extension of these tools and the distribution of their components across multiple machines. However, this architectural design invariably entails a compromise in terms of maintainability, see [20]. Whilst less OAIS-compliant tools, such as Omeka S and Islandora, can offer greater flexibility, particularly with regard to data schemas, more static solutions, including AtoM and Archivematica, are better at ensuring that stored metadata adheres to the most recent standards, such as METS or MODS. While it is possible to define such a schema in the other tools as well, the process is more prone to human error during system setup. In the context of storing disparate data, such as data scraped from social media, which can also be pertinent to labour market research, it would be advantageous to store their content within the record's metadata for enhanced findability in the database, rather than having to scrape media attached to a record in the system.

The initial prototype focuses on the representation of the already given metadata and integration of the transcription workflow. First, a suitable data format to store the records in the system has to be selected. As Omeka S is based on linked data concepts, the metadata terms from the Dublin Core Metadata Initiative (DCMI) have been selected as it supports all of the required terms. Although it is common practice to store archival information in multiple formats, the DCMI terms are a widely accepted standard for the description of digital archival records and considered to be sufficient enough for the first prototypical implementation. Using a linked data based format also increases the usability with other non-archival data.

Different RDF schemas are easy to import with Omeka S through the user interface shown in Figure 4 and the table that stored the different existing records can be imported with a mapping from columns in the CSV file to available fields in the Omeka S instance through the CSV Import module. The resulting overview over all records in the admin interface looks as depicted in Figure 5.

As the planned information system is to be initiated with the documents of the occupations archive, it is recommended that the data be migrated into Archivematica in order to structure it in METS, PREMIS or DCMI which are all considered to be best practice for archival metadata, see [19]. Subsequently, the data can be imported into one of the other tools to facilitate browsing the existing data in collections.

Given that the target format is TEI XML, the use of an XML database such as eXist-db can be highly advantageous for advanced querying within the documents, for example by employing the xquery query language. Additionally, eXist-db can be extended by the TEI Publisher module, which

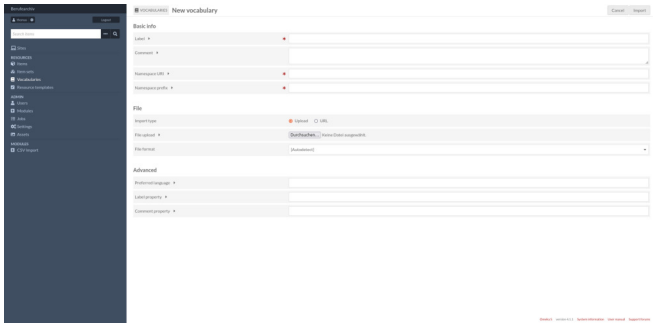

Fig. 4: Ontologies import interface

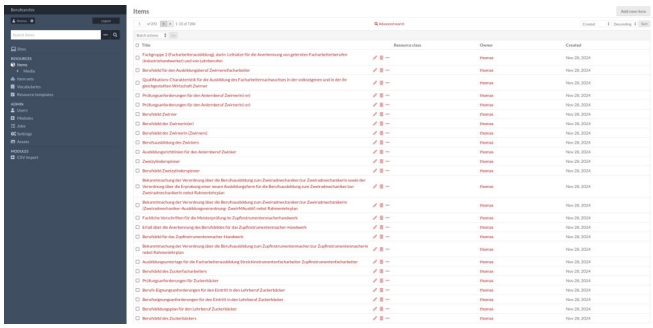

Fig. 5: Items overview in the admin interface

facilitates the straightforward viewing and editing of transcripts. The software also incorporates a functionality for the tagging of named entities within documents, thus rendering it conducive to the annotation of documents. In the context of the occupations archive, the term 'occupation' is understood to encompass not only names, but also skills, tools and other labour market-related terminology. A screenshot of how the TEI XML document that was generated by the prototypical rule-based transcription pipeline is shown in Figure 6

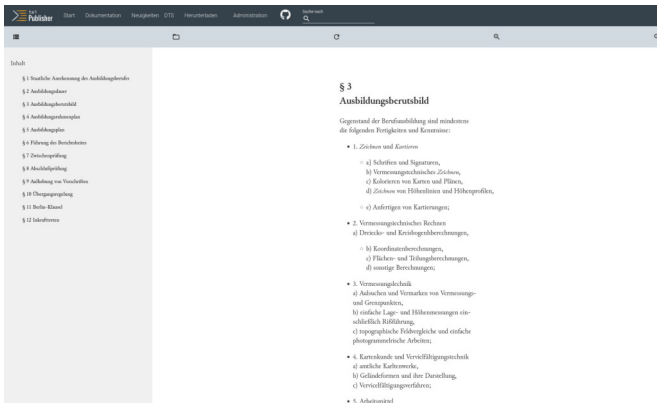

Fig. 6: TEI XML file rendered by TEI Publisher. The interface also allows the annotation of the document and automatically generates an outline.

VI. Results

The implementation of the prototype system, based on Omeka S in combination with a custom AI-powered transcription pipeline, has demonstrated the feasibility and effectiveness of the proposed digital archiving approach. It was demonstrated that both central use cases, namely 'Create Archival Record' and 'OCR-Based Transcription Pipeline', were successfully supported, albeit with complementary tools for full functionality.

Omeka S facilitated the structured creation and management of archival records. Archivists were able to upload scans in multiple formats and annotate them using ontology-based metadata fields, following the Dublin Core Metadata Initiative (DCMI) standard. The system's modular architecture permitted the integration of additional features, such as duplicate entry detection and role-based access control. It is possible for users to differentiate between verified and unverified entries by employing customised resource templates. It was demonstrated that the configuration under consideration was compatible with both standard and alternative workflows.

The AI-based pipeline was implemented externally and linked to the archive system. The system was able to successfully process scanned documents through a series of preprocessing, layout analysis, OCR, and transcript generation steps. The resulting TEI XML files were stored in an XML-native database and made accessible via references within Omeka S. Despite its inability to render TEI XML natively, Omeka S functioned as a stable metadata and document management layer. Integration with TEI Publisher facilitated the visualisation, versioning, and semantic navigation of the structured transcripts.

The efficacy of the modular integration strategy was demonstrated by the outcomes of the study. In this context, Omeka S functioned reliably as the front-end and metadata management layer. Concurrently, specialised tools were utilised for the purpose of handling complex document processing and TEI rendering. This architecture facilitated scalable, standards-compliant digitisation of vocational training regulations spanning over a century of historical data.

While this prototype provided key features of the required functionalities, there are still some open issues and missing features. Omeka S for example only uses a set of predefined roles which does not exactly match the requirements. Especially regarding user privacy, there are some issues. Every authenticated user is able to see all other users in the system. This is not desirable in our context as some users might not want to share their email address and in the case where an intruder manages to gain access to the system, he will also be able to figure out which users might give him the highest privilege. At the moment, there is no built-in way to manage these access rights to restrict this access.

Additionally, Omeka S is particularly well suited for digital exhibitions. However, for now, our focus lies more on data management than presentation. While Omeka S is a very flexible system, stricter rules regarding data types for some

fields can increase data quality as certain fields like the release date should only contain dates, and not e.g. textual information, to have better harmonized data. Such restrictions are not straightforward to implement in Omeka S, although some lighter rules are possible, e.g., that a certain fields must contain media or a URI.

Islandora, for example, allows the creation of custom data schemas and an additional mapping to RDF elements, which also allows to create a linked data representation of the records, thus also an additional representation in form of DCMI terms. It also allows the creation of custom roles with a predefined selection of rights which are sufficient to match the actors shown in Figure 1.

It remains to test the combination of Islandora instead of Omeka S with the other tools to see if all requirements can be fulfilled. Although Islandora comes with more customization options, its microservice architecture also reduces the system maintainability.

The DFG recommends storing data according to the Open Archive Information System (OAIS) model which is currently not implemented in the proposed prototype since Omeka S structurally does not exactly follow this reference model.

Furthermore, there are even more extension that would improve the system's archival capabilities. Archivematica is a tool that is designed to align exactly to the OAIS model and allows an easy integration with AtoM and some other tools. It is in particular designed for the management of the different information packages across the OAIS model and designed to be integrated with other software. However, as we are planning to add more information to the records than is supported by AtoM, Omeka S was selected for the prototype to increase flexibility. A successful integration of Archivematica with Omeka S has been demonstrated in [21]. The authors of [22] on the other hand have been able to create a OAIS compliant system with Archivematica and Islandora.

Given the existence of certain domain-specific metadata fields, and in view of the fact that Islandora provides a greater degree of flexibility while ensuring higher data quality, it has been selected for the purpose of browsing the occupations archive in a future implementation of the system to replace Omeka S.

VII. Conclusions and Outlook

This work presents the occupations archive at the Federal Institute for Vocational Education and Training, in addition to the ongoing endeavour to establish an information system for the digital management of the records in the aforementioned archive. The system will provide an interface for the management of metadata and will also generate structured transcripts in the TEI XML format of the document images. The following tools are presented: The initial design of the system and its components is outlined herein, and the final implementation of a preliminary prototype is currently underway.

Furthermore, there remains a paucity of ground truth data for the training of AI models. It is imperative that the annotations appended to the transcripts are of the utmost precision, incorporating such elements as layout, tables, the sequence of reading, and text hierarchy, in order to facilitate comprehensive comprehension of the document contents, as delineated in the transcripts. Utilising the established models, a combination of the diverse predictions will be employed to generate TEI XML files.

The subsequent stage is the implementation of the proposed system and its testing in a variety of scenarios in order to evaluate its use in creating an interoperable archive system that can be readily extended. There are several challenges that still require resolution, including the question of how consistency can be ensured between Omeka S or Islandora, Archivematica, and the XML database. Following the implementation of a stable prototype, the system can be advanced in several ways. For instance, the transcription pipeline can be expanded to encompass additional information extraction tasks, such as Named Entity Recognition, or alternatively, the implementation of additional pipelines can be contemplated.

Furthermore, the capacity to incorporate additional data into the system is a potential benefit. In the study [23], researchers at the BIBB analysed social media data to ascertain information regarding vocational education and training. This data can assist labour market researchers in a number of ways. While these data are not classical archival data, providing them in the same web service helps create not just an archive, but a diverse information system, while still adhering to archival best practices with the occupations archive.

In addition to the system itself, the TEI XML transcripts can be useful in a variety of ways. The use of Large Language Models in the training of such transcripts is facilitated by their inherent structural nature. The comparison of the regulations themselves is rendered more straightforward, as illustrated in [24]. Furthermore, these models enable more efficient reasoning and referencing to specific paragraphs in the text.

References

[1] T. Reiser, J. Dörpinghaus, P. Steiner, and M. Tiemann, "Towards a datatset of digitalized historical german vet and cvet regulations," *Data*, vol. 9, no. 11, 2024.

[2] T. Reiser, J. Dörpinghaus, and P. Steiner, "Analyzing historical legal textcorpora: German vet and cvet regulations," in *INFORMATIK 2024*. Gesellschaft für Informatik eV, 2024, pp. 2007–2018.

[3] ——, "Learning from historical vet and cvet regulations in germany: What should vet look like and whom should it serve?" in *NORDYRK 2024 BOOK OF ABSTRACTS*, 2024, p. 75.

[4] M. Koistinen, K. Kettunen, and J. Kervinen, "How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine," *Proc. of LTC*, pp. 279–283, 2017.

[5] A. Nabizai and H.-G. Fill, "Eine Modellierungsmethode zur Visualisierung und Analyse von Gesetzestexten," *Jusletter IT*, February 2017. [Online]. Available: http://eprints.cs.univie.ac.at/5131/

[6] V. N. Sai Rakesh Kamisetty, B. Sohan Chidvilas, S. Revathy, P. Jeyanthi, V. M. Anu, and L. Mary Gladence, "Digitization of Data from Invoice using OCR," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 2022. doi: 10.1109/ICCMC53470.2022.9754117 pp. 1–10.

[7] H. Hamann, "The German Federal Courts Dataset 1950–2019: From Paper Archives to Linked Open Data," *Journal of empirical legal studies*, vol. 16, no. 3, pp. 671–688, 2019. doi: https://doi.org/10.1111/jels.12230

[8] C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe, "OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings," *Applied Sciences*, vol. 9, no. 22, p. 4853, 2019. doi: https://doi.org/10.3390/app9224853

[9] J. M. Jayoma, E. S. Moyon, and E. M. O. Morales, "OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines," in *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 2020. doi: 10.1109/HNICEM51456.2020.9400000 pp. 1–6.

[10] S. Van Nguyen, D. A. Nguyen, and L. S. Q. Pham, "Digitalization of Administrative Documents A Digital Transformation Step in Practice," in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 2021. doi: 10.1109/NICS54270.2021.9701547 pp. 519–524.

[11] S. Tsujimoto and H. Asada, "Major components of a complete text reading system," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1133–1149, 1992. doi: 10.1109/5.156475

[12] J. v. Beusekom, D. Keysers, F. Shafait, and T. Breuel, "Example-based logical labeling of document title page images," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007. doi: 10.1109/ICDAR.2007.4377049 pp. 919–923.

[13] S. Klink and T. Kieninger, "Rule-based document structure understanding with a fuzzy combination of layout and textual features," *International Journal on Document Analysis and Recognition*, vol. 4, no. 1, pp. 18–26, 2001. doi: https://doi.org/10.1007/PL00013570

[14] P. Pathirana, A. Silva, T. Lawrence, T. Weerasinghe, and R. Abeyweera, "A comparative evaluation of pdf-to-html conversion tools," in *2023 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, vol. 6, 2023. doi: 10.1109/SCSE59836.2023.10214989 pp. 1–7.

[15] P. Lopez, "Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications," in *Research and Advanced Technology for Digital Libraries*, M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. doi: https://doi.org/10.1007/978-3-642-04346-8_62. ISBN 978-3-642-04346-8 pp. 473–474.

[16] R. Altenhöner, A. Berger, C. Bracht, P. Klimpel, S. Meyer, A. Neuburger, T. Stäcker, and R. Stein, "DFG-Praxisregeln "Digitalisierung". Aktualisierte Fassung 2022." Feb. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7435724

[17] W. Meier, "exist: An open source native xml database," in *Web, Web-Services, and Database Systems*, A. B. Chaudhri, M. Jeckle, E. Rahm, and R. Unland, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. doi: https://doi.org/10.1007/3-540-36560-5_13. ISBN 978-3-540-36560-0 pp. 169–183.

[18] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated, 2012. ISBN 3642311636

[19] R. Altenhöner, A. Berger, C. Bracht, P. Klimpel, S. Meyer, A. Neuburger, T. Stäcker, and R. Stein, "DFG practical guidelines on digitisation. updated version 2022," 2023.

[20] M. Söylemez, B. Tekinerdogan, and A. Kolukısa Tarhan, "Challenges and solution directions of microservice architectures: A systematic literature review," *Applied Sciences*, vol. 12, no. 11, 2022. doi: 10.3390/app12115507. [Online]. Available: https://www.mdpi.com/2076-3417/12/11/5507

[21] B. Kim, S. Nakamura, and H. Watanave, "Using archivematica and omeka s for long-term preservation and access of digitized archive materials," in *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries*, Y.-H. Tseng, M. Katsurai, and H. N. Nguyen, Eds. Cham: Springer International Publishing, 2022, pp. 241–250.

[22] M. Klindt and K. Amrhein, "One core preservation system for all your data. no exceptions!" in *iPRES 2015 - Proceedings of the 12th International Conference on Preservation of Digital Objects*, 2015, pp. 101 – 108. [Online]. Available: http://phaidra.univie.ac.at/o:429551

[23] J. Dörpinghaus and M. Tiemann, "Vocational education and training data in twitter: Making german twitter data interoperable," *Proceedings of the Association for Information Science and Technology*, vol. 60, no. 1, pp. 946–948, 2023.

[24] M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Slezak, Eds., *Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2024.