EasyChair Preprint

№ 11925

From Legal Texts to Digitized Services for Public Administrations

Marianne Mauch, Sarah T. Bachinger, Philipp Bornheimer, Stephan Breidenbach, Daniela Erhardt, Leila Feddoul, Hannes Legner, Felicitas Löffler, Frank Löffler, Maximilian Raupach, Sirko Schindler, Jörg Schröder and Birgitta König-Ries

January 30, 2024

FROM LEGAL TEXTS TO DIGITIZED SERVICES FOR PUBLIC ADMINISTRATIONS

Marianne Mauch / Sarah T. Bachinger / Philipp Bornheimer / Stephan Breidenbach / Daniela Ehrhardt / Leila Feddoul / Hannes Legner / Felicitas Löffler / Frank Löffler / Maximilian Raupach / Sirko Schindler / Jörg Schröder / Birgitta König-Ries

Marianne Mauch(ORCiD 0000-0003-1478-1867), Research Associate, University of Jena, Competence Center Digital Research (Zedif), Leutragraben 1, 07743 Jena, DE, marianne.mauch@uni-jena.de; https://www.zedif.uni-jena.de, https://www.opendva.de

Sarah T. Bachinger(ORCID 0009-0005-5422-2164), Research Associate, University of Jena, Competence Center Digital Research (Zedif), Leutragraben 1, 07743 Jena, DE; sarah.bachinger@uni-jena.de; https://www.zedif.uni-jena.de

Philipp Bornheimer, Research Associate, University of Jena, Competence Center Digital Research (Zedif), Leutragraben 1, 07743 Jena, DE, philipp.bornheimer@uni-jena.de; https://shi-institut.de/ueber-uns/

Stephan Breidenbach, Professor, Entrepreneur, Europauniversität Viadrina Frankfurt/Oder, Center for Legislation and Digitization, Lückhoffstr. 24, 14129 Berlin, DE, stephan-breidenbach@gmx.de

Daniela Erhardt, Projektmanagerin Stadtverwaltung Jena, Zentrales Prozess- und Projektmanagement, Team Transformation, Am Anger 15, 07743 Jena, DE, daniela.erhardt@jena.de; www.jena.de

Leila Feddoul(ORCiD 0000-0001-8896-8208), Research Associate, University of Jena, Heinz Nixdorf Chair for Distributed Information Systems, Leutragraben 1, 07743 Jena, DE, leila.feddoul@uni-jena.de; https://www.fusion.uni-jena.de

Hannes Legner, Research Associate, University of Jena, Competence Center Digital Research (Zedif), Leutragraben 1, 07743 Jena, DE, hannes.legner@uni-jena.de; https://www.zedif.uni-jena.de

Felicitas Löffler(ORCID 0000-0001-6423-7427), Referentin, Thuringian Ministry of Finance, Ludwig-Erhard-Ring 7, 99099 Erfurt, DE, felicitas.loeffler@tfm.thueringen.de, https://www.thueringen.de

Frank Löffler(ORCID 0000-0001-6643-6323), Head, University of Jena, Competence Center Digital Research (Zedif), Fürstengraben 1, 07743 Jena, DE; frank.loeffler@uni-jena.de; https://www.zedif.uni-jena.de

Maximilian Raupach, Research Associate, University of Jena, Competence Center Digital Research (Zedif), Leutragraben 1, 07743 Jena, DE, maximilian.raupach@uni-jena.de; https://www.zedif.uni-jena.de

Sirko Schindler (ORCiD: 0000-0002-0964-4457), Acting Department Head, German Aerospace Center (DLR), Institute of Data Science, Department for Data Management and Enrichment; Mälzerstr. 3–5 07745 Jena, DE; Sirko.Schindler@dlr.de; https://www.dlr.de/dw/

Jörg Schröder, FIM-coach, managing director, Büro für praktische Informatik GmbH (BFPI), Fleckebyer Sraße 1. 18239 Satow, DE, schroeder@bfpi.de, https://www.bfpi.de

Birgitta König-Ries(ORCiD 0000-0002-2382-9722), Professor, University of Jena, Heinz Nixdorf Chair for Distributed Information Systems, Leutragraben 1, 07743 Jena, DE, birgitta.koenig-ries@uni-jena.de; https://www.fusion.uni-jena.de

Keywords: ***end-to-end digitization, digital public administration, Semantic Web, citizen developer***

Abstract: *For the end-to-end digitization of the German public administration, there is a lack of detailed, interoperable descriptions of legal regulations, existing standards, and specific requirements. Yet, they are needed by small companies, decision-makers, administrative staff, and future citizen developers to create fully digitized public services with No-Code/Low-Code platforms and to keep them semi-automatically up to date with changing legal texts. We describe how formal descriptions of processes, data fields, decisions are derived from analyzing legal texts for the service to be digitized and are supplemented by semantic annotations and links to standards, creating innovative services.*

1. Introduction

While implementing the German Online Access Law (OZG)[1], the focus is often put solely on giving citizens access to public services via websites and other digital means. A holistic approach providing true end-to-end digitization of administrative processes is still missing. This prevents taking full advantage of all arising opportunities. Consider a job center that is part of the municipal administration. There, first-time applications for unemployment benefits ("Bürgergeld") can be made either on-site, paper-based, or via a provided PDF-form. Once an application is received, the process may look like this: The case is manually assigned to a case handler. Their first task is to manually enter all relevant data into a specialized processing system. After verifying the application, the caseworker creates a draft for the decision using a predefined template in a common text editor. This case file is then forwarded to another unit for review. Following approval, the case is returned to the case handler, who may manually trigger further processes like rehabilitation measures for the labor market. This glance into current administrative reality already highlights two issues: (i) Discontinuities between the used systems result in repeated manual gathering of relevant data. (ii) The lack of (machine-readable) data impedes the integration of those systems.

Quite obviously, an end-to-end digitization would be beneficial here. This comes with a number of challenges, though, and is made even more challenging by the fact that digitization of processes is not a one-time effort. Rather, processes need to be constantly adapted to new or changing regulations and requirements – a challenging task even in a traditional, non-digital setting.

The basis for any such effort is a detailed understanding of all related aspects: involved processes, applicable regulations, relevant standards, and other requirements arising, e.g., from the existing IT landscape. Currently, only a few experts possess this knowledge. Since there is no consolidated publicly available source, gaining it requires considerable effort. This is not only a severe bottleneck but also poses a high entry barrier for stakeholders like software providers who wish to enter this domain, does not foster transparency of public administration, and severely limits the availability of personnel or partners with the technical skills and knowledge to perform those changes. This includes several challenges: (i) Individual organizations (communes, even smaller federal states) hardly possess the necessary resources for independent technical development. (ii) The often rather slow nationwide standardization cycles in the public sector often hinder faster local innovation, which means even if local expertise exists, possibilities to use them are limited. (iii) Local knowledge, especially of existing challenges, rarely influences political decisions, sometimes resulting in rather hard-to-implement regulations.

This paper outlines our vision for mastering these challenges at the „Offenes Design digitaler Verwaltungsarchitekturen (openDVA)"[2] working group. In three research projects, we investigate how the path from legal text to its digital implementation can be followed in an easier, semi-automated way, both for new legal texts as well as when they are changed, all the while complying with and – where absolutely necessary – extending existing standards. To this end, we are striving for a common knowledge base providing easy access to all relevant information in connection to administrative processes and intelligent services for citizen developers using existing No-Code/Low-Code platforms. The remainder of this paper is organized as follows. Section 2 briefly summarizes the used approaches and technologies. Section 3 describes the architecture and components of our concept for end-to-end digitization. Initial results from our ongoing efforts are presented in Section 4. We conclude in Section 5 with a summary and an outlook for future work.

2. Background and Related Work

When providing a public service, authorities must follow processes based on legal requirements (laws and regulations). Thus, recording and analyzing legal bases is the first step in the creation of digitized administrative services, whereby all process elements involved (e.g., steps, actors, etc.) are identified. The extracted information

[1] https://www.onlinezugangsgesetz.de.
[2] https://www.opendva.de.

is then converted into a formal description consisting of the actual process and its required data in a restricted notation. A major notation standard is the Business Process Modeling Notation (BPMN)[3] format. The Federal Information Management (FIM)[4] provides standardized information for administrative services using a restricted BPMN 2.0[5] notation to describe the process information and widely harmonized data structures to explain the data needed to fulfill the modeled administrative work. Both process and data information are strictly derived from legal requirements[6]. The process model uses only seven patterns to describe the kind of administrative work, such as "Formal check", "Tied decision", or "Decision with leeway". These patterns are called reference activity groups (RAGs)[7]. The retrieval of information to be modeled with FIM out of law, regulations, or affected standards is known as FIM norm analysis. It is currently an entirely manual task often done by highly trained public administration staff. It involves detecting relevant terms or phrases, creating a list of discovered processes and process steps along with the associated data fields, and finally combining all collected elements into a list as a basis for the later modeling. Modeling using the FIM-method creates master process models, master schemes of data, and textual information describing the public service. These three aspects are covered by three public standards: XProzess (process information), XDatenfelder (data structures), and XZufi (textual service information). They are part of the „XML in Public Administration" (XöV) family of standards coordinated by the Coordination Office for IT Standards (KoSIT)[8]. These standards define the structure and semantics for administrative services through cross-disciplinary and cross-project reuse of individual text modules, data fields, and process elements. The framework only maps legal requirements for a service and does not include any further actions by the citizen. Master information provided by FIM is a reliable basis for the development of digital public services.

2.1. From legal norms to formal descriptions

Named entity recognition in the legal domain. Currently, FIM norm analysis is a manual process. We aim to generate automatic suggestions that assign categories to the relevant terms/phrases by using techniques from natural language processing (NLP) for solving the task of Named Entity Recognition (NER)[9], to recognize categories in the text and thus enable machines to better understand legal texts. This is a first step towards more complex tasks, like, in our case, the automatic generation of the final administrative process. Existing research in this area is mostly limited to the English language. Existing research in this area is mostly for the English language; only a few researchers investigate NER in the German legal domain.

These first approaches (Glaser et al.[10], Leitner et al.[11], Zöllner et al.[12], Darji et al.[13]) apply a variety of techniques ranging from more traditional ones like DBpedia Spotlight[14] to Bidirectional Long Short-Term Mem-

[3] Chinosi/Trombetta, BPMN: An introduction to the standard. Computer Standards & Interfaces 34 (1), pp. 124–134, 2012, issn: 0920-5489, doi: 10.1016/j.csi.2011. 06.002.

[4] FITKO: Über FIM, https://fimportal.de/ueber-fim.

[5] BPMN Specification – Business Process Model and Notation: https://www.bpmn.org/.

[6] https://ozg.sachsen-anhalt.de/fileadmin/Bibliothek/Politik_und_Verwaltung/MF/OZG/Bilder/Veranstaltungen/OZG-Sprechstunden/6.FIM-Stammprozess_Datenschemata_Onlinedienst_MI.pdf.

[7] https://www.xrepository.de/details/urn:xoev-de:fim:codeliste:referenzaktivitaetengruppe.

[8] Koordinierungsstelle für IT-Standards (KoSIT, https://www.xoev.de/xoev-4987).

[9] Sun/Han/Li, A Survey on Deep Learning for Named Entity Recognition, IEEE Transactions on Knowledge and Data Engineering, 2022, vol. 34, no. 1, pp. 50–70, doi: 10.1109/TKDE.2020.2981314.

[10] Glaser/Waltl/Matthes, Named entity recognition, extraction, and linking in German legal contracts, IRIS: Internationales Rechtsinformatik Symposium, 2018, pp. 325–334.

[11] Leitner/Rehm/Moreno-Schneider, Fine-Grained Named Entity Recognition in Legal Documents, SE-MANTICS 2019. Springer, 2019, LNCS 11702, pp. 272–287, doi:10.1007/978-3-030-33220-4_20.

[12] Zöllner/Sperfeld/Wick/Labahn, Optimizing Small BERTs Trained for German NER, Inf. 12, 2021, 443. doi:10.3390/info12110443.

[13] Darji/Mitrović/Granitzer, German BERT Model for Legal Named Entity Recognition, ICAART. INSTICC, 2023, pp. 723–728. doi: 10.5220/0011749400003393.

[14] Mendes/Jakob/García-Silva/Bizer, DBpedia spotlight: shedding light on the web of documents, I-SEMANTICS 2011, ACM, 2011, pp. 1–8, doi:10.1145/2063518.2063519.

ory (Bi-LSTM) and BERT-based transformer models. For NER on legal texts in other languages, Naik et al.[15] worked on Indian and Georgoudi et al.[16] on Greek legal texts. More recently, works leveraging large language models (LLM) for solving the NER task have been published (e.g., Gonzáles-Gallardo et al.[17], Shao et al.[18]).

Although some work exists for NER in the German legal domain, the unique nature of the legal texts and categories that need to be identified to create processes for public services makes it difficult to reuse existing models. Moreover, categories to detect have different complexity degrees, which makes simple rule-based techniques not always suitable. For these reasons, we aim to combine different approaches and select the best-performing alternative for each category. This also includes creating our own training corpus and pre-training/fine-tuning existing and new models. We also evaluate different prompt variants for solving the same task using LLMs (cf. Subsection 3.1 for more details).

2.2. From formal descriptions to digitized services.

Norm analysis is used to create formal descriptions of legal requirements. These alone are not sufficient for automation using workflow management systems. We further need more detailed and, above all, executable processes. Nevertheless, even service artifacts available in fimportal.de are not error-free. As barely anyone has used them so far, this has largely gone by unnoticed. In the document quality assurance criteria for OZG reference processes[19], the FIM-Processes component presents how OZG reference information can be created with a higher level of detail using a reduced BPMN standard, called OZG-BPMN, with recurring patterns, so-called typified tasks from FIM master information.

This approach is taken up by the **MODULO process creation method.** This is an interactive approach to model processes[20]. It is based on predefined modules that can be used for collaborative development and visualization of process flows at a uniform level of abstraction. The method is based on BPMN, works with predefined process modules based on FIM, and addresses typical administrative tasks. OZG reference processes can thus be modeled from FIM master processes. Unfortunately, there is currently no official data format for OZG reference processes. Exporting them as BPMN results in FIM information being lost.

Decision trees, created with **Rulemapping**[21], are another option to model public service processes. Rulemapping captures all data and processes in a system and maps them in hierarchically structured, logical trees – the decision trees – extended by some metadata, documents, and data field definitions. These trees represent the rules resulting from a public service‘s legal basis. The aim is to translate lawyers‘ logic and analytical thought processes and describe the legal norm as practical actions (e.g., checking or deciding). The result is a decision structure and can, with the No-Code platform Logos[22], be directly automated, including the necessary data fields. Missing documents or data are automatically requested, and interim results are communicated. Each administrative decision leads to an automatically generated notification.

---

15 Naik/Patel/Kannan, Legal Entity Extraction: An Experim. Study of NER Approach for Legal Documents, Intern. Journal of Advanced Computer Science and Applications 143, 2023, doi: 10.14569/IJACSA.2023.0140389.

16 Georgoudi et.al., Towards Knowledge Graph Creation from Greek Governmental Documents, Advances and Trends in Artificial Intelligence, 2023, pp. 294–299. doi: 10. 1007/978-3-031-36819-6_26.

17 González-Gallardo/Boros/Girdhar/Hamdi/Moreno/Doucet, Yes but... Can Chat-GPT identify entities in historical documents?, In: arXiv preprint arXiv:2303.17322, 2023.

18 Shao/Hu/Ji/Yan/Fan/Zhang, Prompt-NER: Zeroshot Named Entity Recognition in Astronomy Literature via Large Language Models, arXiv preprint arXiv:2310.17892, 2023.

19 https://www.xrepository.de/api/xrepository/urn:xoev-de:xprozess:codeliste:referenzaufgabe_2022-03-23:dokument:Erl_uterung_zu_den_Referenzaufgaben.

20 Löbel/Schuppan, Prozessmanagement neu denken oder wie verengtes Prozessmanagement Innovationen in der Verwaltung verhindert, pp. 255–268, 2023, doi: 10.36198/9783838559292.

21 Breidenbach/Glatz, Rechtshandbuch Legal Tech, C.H.Beck, 2021, isbn: 978-3-406-73830-2.

22 No-Code platform Logos, https://www.knowledgetools.de.

Using Low-Code/No-Code platforms[23] **in public administration**, citizen developers, i.e., domain experts with little to no coding experience, can create processes and associated forms for digitizing public services. Platforms such as formsflow.ai[24] promise faster and more cost-efficient development of digital services as they eliminate communication bottlenecks between software development and public administrations. These platforms are usually focused on specific categories of business applications, e.g., data entry, reporting, workflows, or analysis[25]. Due to the sensitivity of the data at hand, data privacy and security are a prime concern in our research. Formsflow.io is an open-source, modular platform with an identity management engine, a process execution engine, and a form creation tool, each of which can be used independently. Forms can be combined with processes via scripts. Unfortunately, FIM-BPMN and OZG-BPMN are not supported.

To analyze, model, and describe specific public services starting from the legal text and resulting in a complete set of formal processes and data fields, several questions need to be answered: Which standards, patterns, and requirements can be used to model processes and data fields in an executable way? What is needed to be able to use these standards digitally? Digitizing processes in Low-Code/No-Code platforms is simple but very time-consuming. They use few or no norms. We aim to encode the knowledge of processes, data fields, and norms semantically in a way that Low-Code/No-Code platforms can retrieve this knowledge.

2.3. Semantic description of knowledge in public administration

Only small parts of the e-government domain have been described in machine-readable form. The European Union (EU) provides general core vocabularies for public services – the e-Government Core Vocabularies[26]. The EuroVoc Thesaurus[27] is a multilingual and multidisciplinary terminology of the EU. The vocabulary provides terms and phrases with descriptions in semantic formats and has started connecting to existing knowledge graphs[28] (KG), such as Wikidata[29]. To our knowledge, German- or EU-specific knowledge about legal resources, public services, standards, and architectures is unavailable in semantic formats. However, there are some for other countries, e.g., Finland[30], via FINTO – Finnish thesaurus and ontology service[31]. Providing knowledge in machine-readable formats is essential to promote the development of IT-skills in German public administration and to provide easy access to this area for other stakeholders, such as decision-makers, IT companies, and developers. In Feddoul et al.[32], we show that existing efforts for semantic modeling of the knowledge in public administration are, in most cases, not process-oriented, do not follow a known standard, are either too granular or too general and do not link to existing terminologies. Therefore, we aim to model the missing context knowledge for the digitization of public services in Germany (cf. Subsection 4.1).

---

[23] Sahay et.al., Supporting the understanding and comparison of low-code development platforms, 2020, doi: 10.1109/SEAA51224.2020.00036.

[24] formsflow.ai, https://formsflow.ai/de/.

[25] Bock/Frank, Low-Code Platform. Business & Information Systems Engineering 63 (6), pp. 733–740, 2021, issn: 1867-0202, doi: 10.1007/s12599-021-00726-8.

[26] EU, https://joinup.ec.europa.eu/collection/semic-support-centre/solution/e-government-core-vocabularies.

[27] EU, EuroVoc, http://publications.europa.eu/resource/dataset/eurovoc , 2023.

[28] Hogan et.al., Knowledge Graphs. ACM Comput. Surv. 54 (4), 2021, doi: 10.1145/ 3447772.

[29] Vrandečić/Krötzsch, Wikidata: A Free Collaborative Knowledgebase. Communications of the ACM 57 (10), pp. 78–85, 2014, doi: 10.1145/2629489.

[30] JHS 183 WORKING GROUP, National Library of Finland, Semantic Computing Research Group (SeCo), The Finnish Terminology Centre TSK: JUPO – Finnish Ontology for Public Administration Services, 2020.

[31] FINTO: centralized service for interoperable thesauri, ontologies, classification schemes, https://finto.fi/en/.

[32] Feddoul/Raupach/Löffler et al, On which legal regulations is a public service based? Fostering transparency in public administration by using knowledge graphs, INFORMATIK 2023, pp. 1035–1040.

3. Approach

Figure 1 outlines our vision for an end-to-end digitization of public services. Legal texts (e.g., laws) are the foundation of any offered public service. So, we start by analyzing the sources relevant to a specific service. This can be achieved by using several techniques to identify relevant process elements like actors, activities, or links to other legal norms. The goal is to create a formal description of the process at hand.

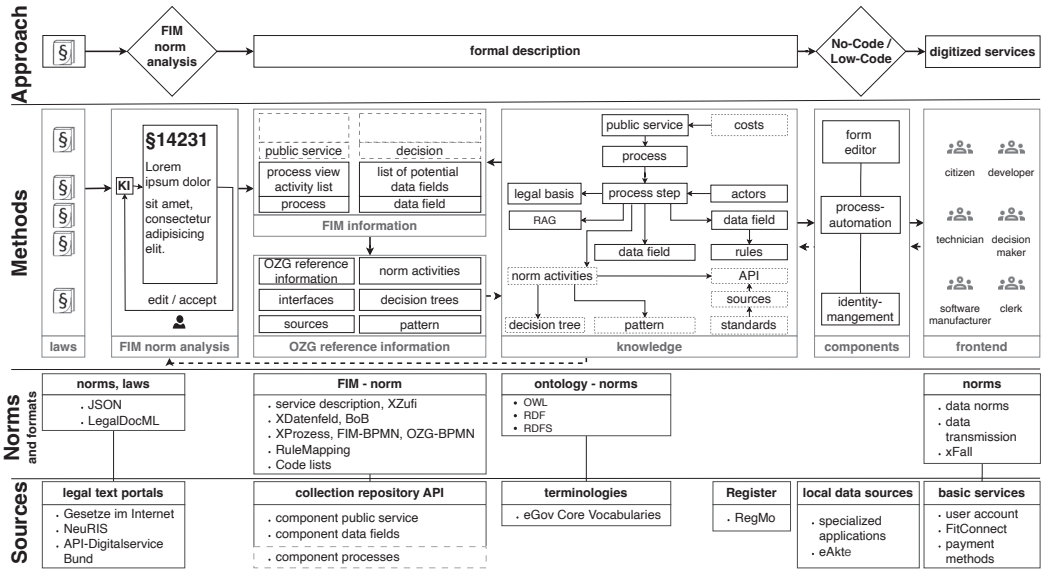

Figure 1: From legal text to digitized services in public administration.

3.1. From legal norms to formal descriptions

To support FIM norm analysis, we aim to automatically generate suggestions that assign categories (10 categories in total) to relevant terms/sentences, allowing users to review them (accept, edit, or delete) and capturing corrections to improve the system (Human-in-the-Loop[33]) continuously. This leverages techniques for NER for detecting categories in the text, thus allowing machines to understand legal texts better. We investigate and compare two NER approaches: Rule-based[34] and machine learning-based, including LLMs. We aim to combine the different approaches and select the best-performing alternative for each category.

Rules are generally helpful if the data to detect follows clear patterns that can be translated into rules. Specifically, we customize the rule-based matcher from the spaCy library[35]. However, even if no training data is needed, rules are not suitable for detecting categories with changing structures or for generalizing to new patterns. In our case, we aim to detect categories with different complexity degrees. For example, while the category[36] "legal basis (Handlungsgrundlage)" can be detected using rules (e.g., BGB §4 and §5 Abs. 3, 4 and 5), the category "main actor (Hauptakteur)" (e.g., Bundesamt für Sicherheit in der Informationstechnik) is difficult to address with rules without the availability of a constantly maintained list of possible instances.

[33] Zhao/Liu, Human-in-the-Loop Based Named Entity Recognition, 2021 International Conference on Big Data Engineering and Education (BDEE), Guiyang, China, 2021, pp. 170–176, doi: 10.1109/BDEE52938.2021.00037.

[34] Eftimov/Koroušić/Korošec, A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations, PLoS ONE 12(6): e0179488, 2017, doi: 10.1371/journal.pone.0179488.

[35] https://spacy.io/usage/rule-based-matching.

[36] Definitions: https://github.com/fusion-jena/GerPS-onto/blob/main/docs/term-definitions.md.

This is why we also investigate supervised machine-learning approaches. We first look for existing pre-trained classifiers that aim to detect either the same or related categories. For example, we evaluated the "flair/ner-german-legal"[37] model that is trained to detect, among others, "laws" and "institutions"[38] over German court decisions. This can be used to detect "legal basis" and "main actor" categories. However, the nature of the training data is different from the law texts analyzed to create processes for public services. Only a subset of categories is covered, and the reused categories may not perfectly match our definitions.

For this reason, we fine-tune different existing machine learning models (e.g., BiLSTM-CRF[39], XLM-RoBERTa[40], or distilbert-base-multilingual-cased[41]) on our own gold standard corpus for NER. We also build a new model by pre-training from scratch on a collection of different types[42] of unannotated law texts (e.g., laws, administrative provisions, etc.) gathered via web crawling of different relevant web sources (e.g., gesetze-im-internet.de) and then fine-tuning it using our gold standard corpus.

Moreover, we investigate data augmentation[43] techniques to enlarge the amount our gold standard corpus (e.g., synonym replacement) and then fine-tune the same previously mentioned models. In addition, we evaluate existing LLMs and select a subset using predefined inclusion/exclusion criteria. These models do not need a training corpus but only a few examples and a collection of instructions provided as a prompt to solve the task. We investigate the influence of different prompt parts and examples on the performance of legal norm analysis.

3.2. From formal descriptions to digitized services

The knowledge in this environment is huge, distributed, and often unavailable or inaccessible. For smaller startups that want to work on the digitization of administration, acquiring this knowledge in a reasonable amount of time is impossible. Citizen developers need well-documented knowledge that provides a compass through the federal IT-architecture and Low-Code/No-Code platforms with common, basic functions such as authentication or payment services that are easy to use and fit into the federal IT landscape. Platforms that already have a modular structure were sought for the research and implementation of an intelligent prototype service that obtains its knowledge from semantically described knowledge based on the underlying legal text. The project began with thorough research into the environment, standards, and important components of the federal IT architecture as well as existing Low-Code/No-Code platforms and their use in the administration. The basis for further scientific work was laid by analyzing the current processes, data, and bases for action when applying for unemployment benefits and planning the processes to be implemented for working with the No-Code platform Logos for use in the jenarbeit[44] in the city of Jena. The formal descriptions – the FIM information – described previously now form the basis for further investigations. The process steps with formal, bound, or discretionary decisions must now be described in more detail as OZG reference processes with the help of standardized, typified tasks from FIM. Knowledge graphs are a central component. These graphs can map the rules that the checks follow; they also describe where discretion must be exercised by humans and thus lay the foundation for automated checks and the preparation of decisions. Rulemapping is used to describe the rules and decisions. This method, which can be used intuitively by lawyers and other specialist users not trained in software technology, is already being used successfully in various federal ministries. A proposal for the formulation of graphs, their use, and their connection to FIM information for the specialized mapping of rule-based administrative decisions is being developed in the project. This proposal

[37] https://huggingface.co/flair/ner-german-legal.

[38] Definitions: https://github.com/elenanereiss/Legal-Entity-Recognition/blob/master/docs/Annotationsrichtlinien.pdf.

[39] Huang/Xu/Yu, Bidirectional LSTM-CRF models for sequence tagging, preprint, arXiv:1508.01991, 2015.

[40] Conneau/Khandelwal et al, Unsupervised Cross-lingual Representation Learning at Scale, ACL 2020, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[41] https://huggingface.co/distilbert-base-multilingual-cased.

[42] https://www.xrepository.de/details/urn:xoev-de:fim:codeliste:handlungsgrundlagenart.

[43] Erd/Feddoul/Lachenmaier/Mauch, Evaluation of Data Augmentation for Named Entity Recognition in the German Legal Domain, AI4LEGAL/KGSum@ISWC 2022, vol. 3257, pp. 62–72.

[44] https://www.jenarbeit.de/en.

is intended to prepare the development of a XöV standard, "XRule", supplementing the three existing XöV standards. For this purpose, very good quality FIM information must first be created manually for the unemployment benefits service, as this information is not available in this quality. MODULO and LIMO are used to create OZG reference information with the help of standardized, typed tasks in FIM. Decisions are modeled with decision trees. By combining these two standards, FIM and Rulemapping, we want to formally describe the associated services so they can be digitally implemented with the help of Low-Code/No-Code platforms and the process management platforms contained therein based on semantically described knowledge. With the help of this prototypical implementation in science and the more general implementation at jenarbeit itself, we can investigate how such platforms can be used intelligently in an administrative context and derive a reference architecture proposal that will make it possible for smaller startups to participate in the development of future digital services.

3.3. Knowledge linking and sharing to foster transparency and traceability.

Textual descriptions of public administration services, federal architectures, APIs, registers, and data formats are spread across various platforms and hardly accessible to humans and machines. Often, they are packed with legal terms and phrases that are hard to understand for developers and citizens. Documentation and supplements for standards are often provided in non-machine-readable formats like PDF, e.g., in XRepository. For successful end-to-end digitization, both humans and machines need appropriate access to this important domain knowledge. We model the necessary information in a semantic format, resulting in a so-called knowledge graph (KG). First, we identify what topics and sub-domains are relevant for the end-to-end digitization of a public service for unemployment benefits. The service was analyzed in detail, including a description of all the actors, activities, norms, standards, and basic services involved, such as the eID[45]. Here, user stories were collected and analyzed as part of several workshops. Further topics and categories that are possibly not specific to the selected service but relevant to the entire e-Government domain were identified. In addition, existing terminologies will be investigated to reuse and link to them to identify structured information that is not available in a semantic format or to identify missing terminologies. Further objectives are quality testing, the visualization of KGs to increase readability and comprehensibility, and the experience gained to identify requirements for a future platform for the development, maintenance, and publication of KGs.

4. Preliminary Results

The compass of the federal IT infrastructure[46] documents the current state of the German IT landscape for service providers and employees and gives an overview of the basic functioning of the public sector, the main guiding principles behind decisions, laws. Further it outlines actors and offers an in-depth analysis of existing and planned components in accordance with the Federal IT Cooperation (FITKO). (cf. Section 3.2).

4.1. From legal norms to formal descriptions

Data collection. We collect regulations that are used as a basis for the creation of German public services. For this purpose, we used a list of services provided by the FIM portal[4] and crawl legal texts[47] corresponding to different services. Each collection of sources related to a specific service is stored in a separate document and identified using the service ID. We considered creating a balanced corpus with respect to the types of services[48] (160 in total, e.g., health) by selecting a fixed number of services from each type. We collected 1020 documents from 141 service types. Code[49] and collected data[50] are publicly available.

[45] eID, https://www.personalausweisportal.de/.
[46] AG openDVA, https://docs.fitko.de/kompass/.
[47] We limit ourselves to those having a publicly available text via https://www.gesetze-im-internet.de/.
[48] https://www.xrepository.de/details/urn:de:fim:leika:leistungsgruppierung/.
[49] Bachinger/Lachenmaier/Feddoul; Data-collection-FIM-laws, doi: 10.5281/zenodo.7875287.
[50] Bachinger/Lachenmaier/Feddoul; Corpus-FIM-laws, doi: 10.5281/zenodo.7875387.

Knowledge Graph (Concept model). We precisely defined and extended the original norm analysis categories provided by the FIM standard[32]. We then semantically described them by developing a KG that models processes in the German public sector reusing existing KGs (BBO[51]) and the e-Government core vocabularies[26]. The KG, together with code for its automatic population with an example public service by parsing XML-based descriptions of BPMN processes (XProzess[52]) and data fields (XDatenfelder[53]), is publicly available[54].

Annotation. The project started with an initial annotation of 10 documents to iteratively create an annotation guideline for a larger annotation campaign. The inter-annotator agreement between the three annotators was calculated, followed by an adjudication step to solve mismatches and discussions of open issues with domain experts. The larger annotation campaign started with a pilot phase and a small fraction of remaining documents to train three new annotators[55] for consistent results. The actual annotation phase with one annotator per document is in progress. We plan to extend the training corpus and iteratively fine-tune and evaluate the previously mentioned models. The best-performing models will be integrated into a tool for human-in-the-loop norm analysis to automatically detect relevant categories in legal texts, to transform the detected information into an editable initial draft process, to export to BPMN format, and to provide a list of data fields for web forms.

4.2. From formal descriptions to digitized services

Use case for digitization and research. In collaboration with jenarbeit[44] we have chosen the use case of applying for the unemployment benefits for a German citizen with no additional income or assets. For a first prototype implementation, we chose a different smaller profile[56].

OZG-Reference information. In collaboration with experts from jenarbeit, the processes and data fields to be digitized were modeled using MODULO and Limo[57]. The process management platform PICTURE[58] was used by the city administration of Jena to model in as much detail as possible individual process steps as they currently take place[59]. The legal basis was important here. Both the process and data model for applying for unemployment benefits have been created and adapted to the system logic for automation.

Digitized Service. For jenarbeit, the case of applying for benefits for a school trip and unemployment benefits was digitized using the Logos Rulemapping platform. All decision trees for unemployment benefits were created. The service will be evaluated, published, and used by the authority.

In the Study on the limits of Low-Code/No-Code in public administration, we investigated the possibilities and limitations of Low-Code/No-Code approaches with stakeholders from the German public administration. We analyzed Low-Code/No-Code platform components, the role of a citizen developer, relevant frameworks, standards regarding the digitized administrative processes, and security of Low-Code/No-Code platforms[60].

Toolbox. A collection of microservices provides services for managing the basic components and services, for using methods and filling knowledge about the formal representations in KG based on legal text analysis.

Intelligent prototype service. Due to its modular structure, the formsflow platform is suitable for an exemplary implementation of the version of the unemployment benefits restricted by a special profile of a citizen

[51] ANNANE / AUSSENAC-GILLES / KAMEL / BBO, BPMN 2.0 based ontology for business process representation, ECKM 2019, Lisbon, Portugal, 2019, Vol. 1, pp. 49–59.

[52] https://www.xrepository.de/details/urn:xoev-de:mv:em:standard:xprozess.

[53] https://www.xrepository.de/details/urn:xoev-de:fim:standard:xdatenfelder.

[54] https://github.com/fusion-jena/GerPS-onto.

[55] One from the public administration sector and the two others with a background in law.

[56] A pregnant woman with a migration background, without additional income and further assets.

[57] Bornheimer/Löbel, citizen-income-LIMO-MODULO, doi: 10.5281/zenodo.8047555.

[58] https://www.picture-gmbh.de.

[59] Erhardt, Example-Public-administration-Process, doi: 10.5281/zenodo.8047811.

[60] Bodenstein/Bornheimer/Raupach/Brust/Schuh/Albertin/Krumpe/Erhardt/Aldubosh/Frank Löffler/Mauch/ et.al., low-Code/noCode-IT-architecture, 2023, doi: 10.5281/zenodo.8055954.

with basic services used, such as eID or payment transaction components. It uses the toolbox, creates processes and matching forms based on machine-interpretable knowledge from our formal representations in the KG.

FIM-Information for unemployment benefits. Our FIM expert is currently preparing the FIM information to also be used digitally. Then the already created OZG reference information has to be adapted to match this new FIM information. Further, it can be shown which RAG with which typed tasks have been described more precisely with the help of MODULO and LIMO. The next step is to examine how the RAG‘s “Formal decision”, “Decision without leeway”, and “Decision with leeway” can be modeled using which decision tree.

4.3. Knowledge linking and sharing.

User Stories. Collecting public services in user stories allows us to identify further broad topics to restructure modeled classes and relationships if necessary and extend the developed high-level ontology (Section 4.1).

The use case of unemployment benefit was analyzed from multiple perspectives to gain detailed insights into relevant terms and relationships from the citizen benefit to legal, process, and IT development perspectives. For each use case, we extract and classify important entities from legal texts, process or technical descriptions.

Competency questions and quality criteria were created for terminology development or extension. Competency questions are necessary to describe the purpose of a modeled domain and validate the final terminology. Quality criteria were established to ensure that terminologies are usable, interlinked, and maintainable.

5. Conclusion & Vision

In this paper, we outlined our vision for digitization of public administration. We discussed different approaches to derive formal, semantically annotated process descriptions starting either from legal norms or the experiences of practitioners. The result is a well-documented landscape of service requirements and existing base services. As a first use, we describe Low-Code/No-Code platforms to enable citizen developers to create their own digitized processes, largely without the support of software engineers. The developed knowledge base can also serve as a foundation for a range of advanced applications. A rather straightforward use case is a semantic search for a variety of stakeholders. Another application is the automated tracing of changes. Rules and regulations are often changed and adapted to new circumstances or requirements. A formal representation allows for an easy comparison between two versions and changes to the resulting workflows are thus easily possible. Finally, we imagine formal descriptions as the source for generating textual descriptions for human use. While the legal texts defining a workflow are often geared towards precise and unambiguous descriptions, they are rather hard to comprehend for common citizens. However, structured representations of workflows also allow the generation of texts geared to laymen or even in simple language. This may omit details necessary for legal scholars but would greatly improve the situation for large shares of the general population.

In a more distant future, we imagine reversing the entire process: Instead of deriving formal descriptions from legal texts, the law-defining process would create the formal description, from which legal texts would then be derived in a similar way as other human-readable descriptions. This would not only remove any inaccuracies in the aforementioned transformation process but also very likely reduce the number of ambiguous wordings to be interpreted later by the court system.

6. Acknowledgements

The research projects Canaréno, simpLEX, and KollOM-Fit of the working group openDVA were funded by the Federal Ministry of the Interior and Community, FITKO, and Thuringian Ministry of Finance in Germany in the scope of the OZG implementation as well as funding from the FITKO from the digitization budget. We would like to thank all employees, project partners, and supporters of the openDVA working group, who could not be mentioned here by name, for their great support, helpful comments, discussions, and good cooperation.