Combining Knowledge about Document Types and Structures for Enhanced Content Curation

Karolina Zaczynska[0000−0002−5395−5463], Florian Kintzel[0000−0003−2423−1260], Julián Moreno-Schneider[0000−0003−1418−9935], and Georg Rehm[0000−0002−7800−1893]

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
Alt-Moabit 91c, 10559 Berlin
firstname.lastname@dfki.de

Abstract. We present the conceptual design of a language technology (LT) system that enables enhanced document curation and processing of different documents types by providing customized NLP workflows that respond and adapt to the extracted characteristics of the input documents. To optimize document and text understanding, the processing steps will not only incorporate textual features but also layout and document type related features like document structure, and the communicative function of specific parts or constituents of a document (e. g., header, subtitle, paragraph, footer). We tackle the lack of standardized representation formats for many of these document features by presenting the first draft of an ontology (QOntology) we plan to incorporate into the overall workflow manager. Since the work is still in progress, we present the theoretical background and conceptual design decisions of the approach which will be the basis of experiments in future work.

Keywords: Document Curation · Content Ontology · Workflow Manager.

1 Introduction

Organizing, structuring, processing and understanding a vast number of heterogeneous documents poses not only a challenge for humans but also for NLP platforms, which often only provide a one-size-fits-all solution for documents without taking into account the characteristics of different content and document types. A *newspaper article* should be treated differently than a *scientific article* and it would be beneficial to have access to knowledge regarding the structure of a document because it is often linked to 'standardized' communicative functions that each part of a text exhibits. Taking the afore mentioned example, newspaper articles are often written according to the inverted pyramid where the most important information is in the beginning (title, first paragraph) and the least important at the end, whereas scientific articles most often include a title, an abstract, the first section titled *Introduction*, and so forth. Using the information different document types provide and – for document images – making use of layout information (e. g., bold indicating importance or a title, italics indicating proper names), we can enhance document curation. The idea of document curation – in the context of the project and corresponding platform QURATOR [25][1] – is to support content curators in

[1] https://qurator.ai

their professional activities, e. g., by classifying and structuring documents, by simplifying processes and workflows or by recommending important content. Text processing in this procedure can be improved utilizing customized language models trained on the respective domain of the document (e. g., journalism vs. science). A summarization service [2] could focus on the text areas that by default contain the most important content. This paper presents an approach for adaptive document processing taking into account the type, domain, and also layout features of a document, which could lead to a major improvement over the typical “one size fits all” approach typically used for document processing. The approach enables intelligent content processing with customized and customizable workflows that respond and adapt to the characteristics of the input [16].

The terms *text type* and *text genre* are used differently in different linguistic articles [10]. Biber, e. g., characterises genres in terms of the author’s or speaker’s communicative purpose, while text types classify texts on the basis of text-internal criteria [5]. We decided to use a pragmatically motivated term for this paper, *document type*, which we define as a group of documents with similar standardized textual and layout features, which we can utilize to optimize NLP workflows. The subject field, or *domain*, of a text is a specific area or use of language such as legal, journalism, or science.

To have a machine-readable vocabulary for all the necessary document characteristics we want to extract, we designed a specific ontology, QOntology, which is used to provide a structured vocabulary for indexing and workflow control. We present the first draft of the ontology in which we make use of existing ontologies we deem useful for our approach and describe how we adjusted them to the requirements of our system. The main goal of the ontology is to be able to describe documents so precisely in a semantic way that the curation technologies that process them can adapt their behavior based on its features. Our approach is developed in the context of the QURATOR project [25], a technology transfer project funded by the German federal ministry of education and research in which a consortium of research institutes and industry partners collaborate to develop a platform for the curation of digital information.

The remainder of this paper is structured as follows: Section 2 presents related work in the field. Section 3 describes the ontology in detail and its planned usage with the workflow manager. Section 4 summarizes our vision of an intelligent workflow manager which makes use of the ontology and describes an example workflow for a document. Finally, Section 5 concludes the paper.

2 Related Work

The idea of being able to adjust the parsing and further processing of documents according to document characteristics is not new. Elaborating approaches for document understanding was a goal in the 1980s, abandoned in the mid 1990s and now reappearing. The MUC (Message Understanding Conference) [8] competitions in the 1990s with tasks on information extraction and (since the sixth conference in 1995 [29]) named entity recognition mainly focused on narrow sub-domains. One interesting attempt tackling this task was the MUSE project for cross-domain named entity recognition aiming to identify the parameters relevant for the processing across different document formats and domains [15].

Various types of NLP platforms are specialized in a specific domain or document type or focus on task-specific processing. The language technology platform *Common Round* is focusing on semantic enrichment and argument mining on large-scale web debates [30]. *Canary* is an NLP platform for repositories of unstructured clinical data [14]. Another example is OCR-D, a project initially concentrating on the transformation of German prints from the 16th until the 19th century into digital texts. The project now also provides user-specific modular tools and a workflow manager for the digitization of various document types [18]. While OCR-D provides rudimentary workflows and approaches for automatic document structure and text recognition, we additionally aim to analyze and classify incoming documents regarding their textual and structural characteristics and, based on those results, to initiate adjusted workflows.

We recently observed an increasing number of approaches that emphasize the importance of layout features and their semantic meaning in combination with text features for improving document and text processing. The language model LayoutLM, for example, was trained jointly with text and layout information across scans of documents and reveals good results after fine-tuning for different tasks, such as form understanding, receipt understanding, and document image classification [28]. We make use of the latter in our system for a service that can classify different document types.

Those language models are often based on datasets with annotated layout information including semantic information (such as title, paragraph, footnote, etc.) for some highly standardized text types, like scientific papers. For the creation of these datasets weak supervision is often used based on the fact that scientific publishing platforms often provide metadata with semantic layout annotations (in XML or LaTeX format) as well as the corresponding document images as PDF with compiled layout features [11,32]. In future work, we want to use these datasets to train models for text structure recognition and include the corresponding text regions in our ontology. The PRIMA Research Lab provides the PAGE (Page Analysis and Ground-Truth Elements) annotation scheme which also includes layout information, but the schema is focusing on fine-grained annotation to support individual stages within an entire sequence of document image analysis methods (from document image enhancement to layout analysis to OCR) and their evaluation [20]. Therefore, we decided to use a new annotation scheme based on the ontology we designed for the needs of our workflow manager.

3 QOntology: An Ontology and Classes for Document Curation

In this section, we describe the classes and sketch the usage of the *QOntology*. Our main motivation for developing a new ontology, considering all the available options, was to address the lack of semantic representation for several document features that we consider relevant for the document curation processes in the project QURATOR, such as the domain, document type and sections. The ontology is a machine-readable taxonomy that specifies different types of content, including specific metadata and characteristics. The extracted metadata enables us to initiate document processing depending on the features and types or classes of incoming documents, which means subsequently to channel incoming content into content-type-specific processing workflows (see Section 4). While designing the ontology, for which we used Protégé[2], we made use of relevant, existing ontologies, not only to avoid re-definition of entities but to make our ontology able to interact with other applications. The imported ontologies already come in OWL/RDF format[3], so that transformation processes were not necessary. Next, we describe the QOntology and the used ontologies and how we adapt them to our needs. We also give a short description of the planned usage of the ontology for our document curation platform, also see Section 4.3.

3.1 Description of Document Types and Components

To define workflows according to specific document features for each document type, we first need to define the different types of documents. Categorizing and cataloging textual entities is a task typically done by librarians, so we decided to build upon an ontology designed to define bibliographic records. FaBiO (Functional Requirements for Bibliographic Record[4], FRBR-aligned Bibliographic Ontology) is an ontology for representing and publishing bibliographic records of scholarly endeavors [19]. It focuses upon describing scholarly articles and text-based publications, such as *books*, *magazines*, and *newspapers* but also contains other text types like *Email*, Web Page, and *Dataset*. It includes classes and properties for collecting important metadata about the document, like the title-name, date, identifier and URL. FaBiO divides bibliographic records fourfold, as abstract entity (class *Work*), the realization mode of a work (*Expression*), the physical embodiment of a work (*Manifestation*), and as the medium of a single exemplar of the physical embodiment, which can be analog or digital (*Item*). Because of the large number of different types of documents it includes, we found FaBiO very useful as a fundamental set of labels and categories for our document type classes. In future work, we will reduce the classes depending on whether they express known layout and textual criteria that we can use for customising workflows and whether they are interesting for the user of the curation platform.

For the annotation of document sections, we adapted the Document Components Ontology (DoCO) which provides a structured vocabulary of document elements [6]. DoCO is a combination of several ontologies to describe (1) the structural patterns like text vs. non-text (based on the Patterns Ontology[5]), and (2) classes for document sections according to their communicative function (Discourse Elements Ontology[6]), while we are more interested in the latter[7]. The communicative function of a *Title*, for example, is to attract the reader's attention and to transport the most important information, the *NavigationBar*, a section appearing in the document type *WebPage*, enables the user to navigate and offers an overview of the website's structure. Generally, DoCO focuses on the content of mainly scientific and other scholarly texts. Therefore, we extended QOntology with more section classes to enable annotations for other document

[2] https://protege.stanford.edu

[3] https://www.w3.org/TR/owl-features/

[4] http://www.sparontologies.net/ontologies/frbr

[5] https://sparontologies.github.io/po/current/po.html

[6] https://sparontologies.github.io/deo/current/deo.html

[7] The Pattern Ontology is based on the idea of classifying different types of XML tags in HTML-documents and has, therefore, a different scope, also see [22,23,24].

types as well. These sections will be used for the Semantic Layout Identification tool in our system (see Section 4.2). We started to link the document type and document component classes by defining which document type can or must contain specific document sections (a news article has a title and at least one paragraph, but not necessarily a footnote; a scientific article hast one abstract, etc.). This is work in progress.

3.2 Other Classes

As the ontologies we presented do not provide all classes we deemed important for our document curation requirements, we included the following classes from other resources. For the description of web pages, we utilized parts of schema.org[8], like the class *WebPageElement* including sub-classes for site navigation, sidebar, advertisement, etc. Sometimes it is necessary to not only annotate the document itself but also the different processing stages and tools. An example is a paper document and its scanned version, where it would make sense to record and to annotate both documents as different entities. For this purpose, we included the Provenance Ontology, PROV-O [4]. PROV-O provides an annotation scheme to distinguish the relationship between two documents and additionally allows to describe the origin, production, modifications and responsible entities.

3.3 Using the Ontology in the Content Curation Platform

For computational document processing, various language technology tools exist enabling partial processing and information extraction concerning distinct aspects of a document. From a technical point of view, to build an document curation platform combining these tools, it becomes necessary to use information from previous steps (e. g., document type and language) to control which specific tasks to perform afterwards (e. g., which pipeline to use regarding the characteristics of a document). Similarly, the same metadata can be viewed as of special interest for indexing the documents after processing for later search and retrieval. We use the ontology to provide machine-readable vocabulary and rules which defines the underlying database structure of our platform and triggers the processing workflow. In more detail, the content curation platform will make use of the ontology in four ways:

1. It acts as the blueprint for a database schema as a direct mapping of the ontology.
2. When a document arrives in the system, this data structure is (semi-)automatically populated based on the incoming metadata of the document and the results of several analysis processes, e. g., language identification or document structure analysis.
3. The ontology describes the classes used to classify documents.
4. The NLP workflow is dynamically adapted based on the document metadata, i. e., the specific pipeline step to call next is determined by the metadata.

The ontology, which is currently work in progress, will be applied for NLP workflow configuration, as described in Section 4.

[8] https://schema.org

4 Adaptive Workflow Manager for Document Processing

The ontology helps to classify incoming documents and to gain metadata in order to trigger particular processing workflows, but this is not enough to define the parameters and procedures for each processing step to be performed on the document. For this, we designed a workflow manager (WM), which is able to execute certain tasks depending on the annotations a document entails. These annotations are continuously updated through the analysis performed by each NLP module, so the WM will adapt its functionality to each modification of the annotations. We can describe the challenge as the metadata-driven combination of different task-specific NLP services. Below we describe the technical details of the WM and give an overview of the workflow components necessary for comprehensive document curation.

4.1 Functionality

The communication between different language technology tools – where the output of one tool serves as input for the following tool – requires interoperability regarding the used vocabulary and annotation format used across the NLP tools. The former is guaranteed by the ontology (Section 3), for the latter, we use an annotation format that enables the flexible orchestration of NLP services, which is based on the NLP Interchange Format (NIF) [9]. NIF can be serialized in RDF-XML and serves as the communication language between the input and output of the services. Using NIF, we can organize the interoperability of different NLP services in our WM and allow Linked Data compatibility. The architectural and technical details of the WM used for document processing are described in a previous paper [16].

The workflow uses a JSON-based language format to define the processing steps for the documents. The WM can adapt the functionality of a workflow previously defined based on the metadata and other information of a document through a set of rules, which are defined manually for each workflow component and are composed of three elements (see Equation 1): the respective property (P_s); the value that the property must have or the condition that it must meet (v); and the action it has to perform (a).

$$R_1 = [P, v, a] \tag{1}$$

For example, if we want to define the rules so that the named entity recognition module uses a specific model depending on the document's language, the rules would be: $nif : language$ is the property to be analyzed, EN/DE is the value the property must have, and the modification of $model\_name$ is the targeted action (see Equation 2). Depending on the language of the input document (provided as metadata or annotated by a language identification module), the module will adapt its functionality using a differently trained model.

$$\begin{aligned} R_{NER}(EN) &= [nif : language, EN, model\_name = BERT\_EN] \\ R_{NER}(DE) &= [nif : language, DE, model\_name = BERT\_DE] \end{aligned} \tag{2}$$

4.2 Workflow Components

This section describes the components which extract information and metadata about the incoming document the workflow afterwards uses to control and adapt the workflow execution for the following steps. The most important components are as follows.

- **Language Identification** The language identification service langid [13] annotates the document or parts of the document with the property $qont : language$.
- **Document Type Classification** For this task we implement a fine-tuned version of LayoutLM [31]. We used the RVL-CDIP dataset[9] for fine-tuning the model on the document type classification task. The model can distinguish 16 discrete document types with an accuracy of 0.84. We will expand the initial experiments so we can classify different document types according to the classes defined in the QOntology ($qont : documentType$).
- **Optical Character and Layout Recognition (OCR, OLR) and Semantic Layout Identification (SLI)** If the input document is an image, we need to first perform OCR on the image, which is the automatic conversion of document images (scans, photos, etc.) into machine-readable text. OCR is a complex process, including several steps in addition to character recognition, like preprocessing (image optimization and binarization), layout analysis (recognition and classification of structural features), and eventually post-processing (error correction). We use Tesseract[10] which supports various output formats (plain text, hOCR, PDF, invisible-text-only PDF, TSV). Layout recognition includes text line recognition, text vs. non-text recognition, region segmentation and classification, and document-level structural analysis [3]. After preprocessing the documents with Tesseract we use GROBID (GeneRation Of BIbliographic Data), which annotates the document sections according to their communicative function, which we call Semantic Layout Identification (SLI) [1] [12]. For future work, the classes for the SLI service will be aligned to the QOntology definition for the document components.

In addition to the fully automatic detection of relevant parameters to drive the workflow, metadata gained from other sources will also be utilized. We distinguish between these sources: (1) automated extraction of metadata through NLP tools (see above); (2) manual input of document metadata by the user; (3) existing metadata that accompany a document; and (4) metadata determined by the input channel (e. g., twitter, web-crawling, user-provided uploads).

4.3 Example Workflow for Enhanced Content Curation

The following example workflow extracts metadata and semantic information from a document image (a scientific paper) to drive additional downstream processing tasks.

In the example (Figure 1), the user provides a scientific article in PDF format in English, including a French paragraph (a quotation). The first workflow components aim to enrich the document's metadata by preprocessing the paper and understanding

[9] https://www.cs.cmu.edu/~aharley/rvl-cdip/

[10] https://github.com/tesseract-ocr/tesseract

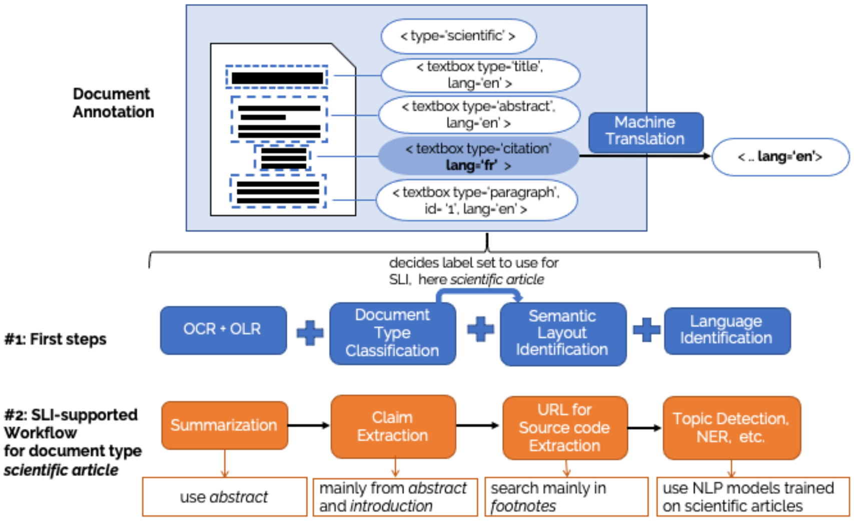
Document Annotation
< type='scientific' >
< textbox type='title', lang='en' >
< textbox type='abstract', lang='en' >
< textbox type='citation' lang='fr' >
< textbox type='paragraph', id= '1', lang='en' >
Machine Translation
< .. lang='en'>
decides label set to use for SLI, here scientific article
#1: First steps
OCR + OLR
Document Type Classification
Semantic Layout Identification
Language Identification
#2: SLI-supported Workflow for document type scientific article
Summarization
Claim Extraction
URL for Source code Extraction
Topic Detection, NER, etc.
use abstract
mainly from abstract and introduction
search mainly in footnotes
use NLP models trained on scientific articles

Fig. 1. Workflow that extracts document type and target language and branches accordingly

the structure of the text. The OCR and OLR service extracts the text and different parts the document is made up of. The next step is document type classification, which then defines the label set used for the document components classification (SLI step). As mentioned, the document types are defined in the ontology as well as the label sets each document type can contain (see Section 3.1). In this example, we obtain the label *Scientific Paper* and section labels such as, among others, *Abstract*, *Introduction* and *RelatedWork* will be used for annotation. Once the document's sections are determined, the language identification module will then assign a language to each section. Because there is a quotation identified as French, a machine translation service can translate this piece of content into English; another possibility would be to define dedicated pipelines for each sections in another languages.

After the first processing steps, the WM is now able to channel the document into a pipeline adapted to the document type *Scientific Article* for the language *English*. Understanding the characteristics of scientific papers and their internal structures can be utilized to weight the importance of different sections for each task in the following steps. If we want to obtain a summary of the document, the WM now can simply extract the section marked in the previous SLI step as *Abstract*, because the abstract is a summary of a paper. Claims are a fundamental unit of scientific discourse, therefore the next step in our workflow is claim extraction. Claims often appear in the *Abstract* or *Introduction*. The model could be instructed to only search for claims in these sections or focus on them, which could improve the overall results. The same applies to the extraction of URLs. In the last step, we make use of the fact that language models are mainly trained in a task-specific (for NER, topic detection, etc.) but also language- and domain-specific way. Based on the metadata, the workflow manager can automatically choose an optimized language model for the input document, which is in this case an

English language model trained on scientific documents for topic detection. An alternative workflow could foresee processing documents from the legal domain, e. g., case reports. NLP services like time extraction, named entity recognition for legal entities, legal argumentation extraction etc. can also be included in such a workflow for this class of documents [27,17].

5 Conclusion

In this paper, we describe our concept of a document processing and curation platform, which uses not only textual features but information about different document types and semantic annotation of text regions in documents. We present the first version and concept of the QOntology that will provide a structured vocabulary for the annotation of document features and to provide these features to the feature-driven processing workflow manager (WM). We describe the functionality of the WM and illustrate our concept with an example workflow for a document curation pipeline adapting to the extracted features of the incoming document but there is still a lack of tools and data sets that support the semantic annotation of document regions for other document types than scientific articles. We are currently also exploring the feasibility of distributed processing workflows that include components made available on multiple platforms [26]. For the annotation of semantic regions in documents, an exchange with scholars from humanities could be beneficial because they are the most familiar ones regarding the documents and their internal logical structure [7]. The Discourse Elements Ontology also reveals its focus on scientific articles. That is why we extended the QOntology with more classes to enable possible annotations for other text genres. Another idea for future work is to make use of the more general ontology for the Penn Discourse Treebank (PDTB) which strives to model discourse structures, particularly coherence relations between abstract entities in the text [21].

Acknowledgments

The research presented in this paper is funded by the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (http://qurator.ai) (Unternehmen Region, Wachstumskern, grant no. 03WKDA1A).

References

1. Grobid. https://github.com/kermitt2/grobid (2008–2021)
2. Aksenov, D., Moreno-Schneider, J., Bourgonje, P., Schwarzenberg, R., Hennig, L., Rehm, G.: Abstractive Text Summarization based on Language Model Conditioning and Locality Modeling. In: Calzolari, N., Béchet, F., Blache, P., Cieri, C., Choukri, K., Declerck, T., Isahara, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020). pp. 6682–6691. European Language Resources Association (ELRA), Marseille, France (2020)

3. Baierer, K., Boenig, M., Engl, E., Neudecker, C., Altenhöner, R., Geyken, A., Mangei, J., Stotzka, R., Dengel, A., Jenckel, M., Gehrke, A., Puppe, F., Weil, S., Sachunsky, R., Schiffer, L.K., Janicki, M., Heyer, G., Fink, F., Schulz, K.U., Weichselbaumer, N., Limbach, S., Seuret, M., Dong, R., Burghardt, M., Christlein, V., Doan, T.H.A., Dogan, Z.M., Panzer, J.H., Schima-Voigt, K., Wieder, P.: OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative. BIBLIOTHEK – Forschung und Praxis (2020). https://doi.org/http://dx.doi.org/10.18452/21548
4. Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: PROV-O: The PROV Ontology. Tech. rep. (2012), http://www.w3.org/TR/prov-o/
5. Biber, D.: Variation across Speech and Writing. Cambridge University Press (1988). https://doi.org/10.1017/CBO9780511621024
6. Constantin, A., Peroni, S., Pettifer, S., Shotton, D., Vitali, F.: The Document Components Ontology (DoCO). Semantic Web **7**(2), 167–181 (2016). https://doi.org/10.3233/SW-150177
7. Engl, E., Baierer, K., Boenig, M., Hartmann, V., Neudecker, C.: Volltexte – die Zukunft alter Drucke:. o-bib. Das offene Bibliotheksjournal / Herausgeber VDB **7**(2), 1–4 (2020). https://doi.org/10.5282/o-bib/5600, https://www.o-bib.de/article/view/5600, number: 2
8. Grishman, R., Sundheim, B.: Message Understanding Conference-6: A Brief History. In: Proceedings of the 16th Conference on Computational Linguistics - Volume 1. p. 466–471. COLING '96, Association for Computational Linguistics, USA (1996). https://doi.org/10.3115/992628.992709, https://doi.org/10.3115/992628.992709
9. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP Using Linked Data. In: Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K. (eds.) The Semantic Web – ISWC 2013. pp. 98–113. Springer, Berlin, Heidelberg (2013)
10. Lee, D.: Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. Language Learning and Technology **5** (01 2002)
11. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: DocBank: A Benchmark Dataset for Document Layout Analysis. CoRR **abs/2006.01038** (2020), https://arxiv.org/abs/2006.01038
12. Lopez, P., Romary, L.: GROBID - Information Extraction from Scientific Publications. ERCIM News **2015**(100) (2015), https://ercim-news.ercim.eu/en100/r-i/grobid-information-extraction-from-scientific-publications
13. Lui, M., Baldwin, T.: langid.py: An Off-the-shelf Language Identification Tool. In: Proceedings of the ACL 2012 System Demonstrations. pp. 25–30. Association for Computational Linguistics, Jeju Island, Korea (2012), https://www.aclweb.org/anthology/P12-3005
14. Malmasi, S., Sandor, N., Hosomura, N., Goldberg, M., Skentzos, S., Turchin, A.: Canary: An NLP Platform for Clinicians and Researchers. Applied Clinical Informatics **8**, 447–453 (05 2017). https://doi.org/10.4338/ACI-2017-01-IE-0018
15. Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y.: Named Entity Recognition from Diverse Text Types. In: Proceedings of the Recent Advances in Natural Language Processing 2001 Conference. pp. 257–274. Tzigov Chark, Bulgaria (2001), http://gate.ac.uk/sale/ranlp2001/maynard-etal.pdf
16. Moreno-Schneider, J., Bourgonje, P., Kintzel, F., Rehm, G.: A Workflow Manager for Complex NLP and Content Curation Pipelines. In: Rehm, G., Bontcheva, K., Choukri, K., Hajic, J., Piperidis, S., Vasiljevs, A. (eds.) Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020). pp. 73–80. Marseille, France (2020), 16 May 2020
17. Moreno-Schneider, J., Rehm, G., Montiel-Ponsoda, E., Rodriguez-Doncel, V., Revenko, A., Karampatakis, S., Khvalchik, M., Sageder, C., Gracia, J., Maganza, F.: Orchestrating NLP Services for the Legal Domain. In: Calzolari, N., Béchet, F., Blache, P., Cieri, C., Choukri, K.,

Declerck, T., Isahara, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020). pp. 2325–2333. European Language Resources Association (ELRA), Marseille, France (2020)

18. Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K.M., Hartmann, V., Herrmann, E.: OCR-D: An end-to-end open source OCR framework for historical printed documents. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019). pp. 53–58. ACM, New York (2019). https://doi.org/10.1145/3322905.3322917, 46.12.02; LK 01
19. Peroni, S., Shotton, D.: FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Journal of Web Semantics **17**, 33 – 43 (2012). https://doi.org/https://doi.org/10.1016/j.websem.2012.08.001, http://www.sciencedirect.com/science/article/pii/S1570826812000790
20. Pletschacher, S., Antonacopoulos, A.: The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. In: 2010 20th International Conference on Pattern Recognition. pp. 257–260 (2010). https://doi.org/10.1109/ICPR.2010.72
21. Prasad, R., Webber, B., Lee, A.: Discourse annotation in the PDTB: The next generation. In: Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation. pp. 87–97. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/W18-4710
22. Rehm, G.: Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage. In: Sprague, R. (ed.) Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35). pp. 1143–1152. IEEE Computer Society, Big Island, Hawaii (1 2002)
23. Rehm, G.: Hypertextsorten: Definition – Struktur – Klassifikation. Ph.D. thesis, Institut für Germanistik, Fachgebiet Angewandte Sprachwissenschaft und Computerlinguistik, Justus-Liebig-Universität Gießen (2005), http://geb.uni-giessen.de/geb/volltexte/2006/2688/, thesis submitted on 16 August 2005 and defended on 23 January 2006.
24. Rehm, G.: Hypertext Types and Markup Languages – The Relationship Between HTML and Web Genres. In: Metzing, D., Witt, A. (eds.) Linguistic Modelling of Information and Markup Languages. Contributions to Language Technology, pp. 143–164. Springer, Dordrecht, Heidelberg, London, New York (2010)
25. Rehm, G., Bourgonje, P., Hegele, S., Kintzel, F., Schneider, J.M., Ostendorff, M., Zaczynska, K., Berger, A., Grill, S., Räuchle, S., Rauenbusch, J., Rutenburg, L., Schmidt, A., Wild, M., Hoffmann, H., Fink, J., Schulz, S., Seva, J., Quantz, J., Böttger, J., Matthey, J., Fricke, R., Thomsen, J., Paschke, A., Qundus, J.A., Hoppe, T., Karam, N., Weichhardt, F., Fillies, C., Neudecker, C., Gerber, M., Labusch, K., Rezanezhad, V., Schaefer, R., Zellhöfer, D., Siewert, D., Bunk, P., Pintscher, L., Aleynikova, E., Heine, F.: QURATOR: Innovative Technologies for Content and Data Curation. In: Paschke, A., Neudecker, C., Rehm, G., Qundus, J.A., Pintscher, L. (eds.) Proceedings of QURATOR 2020 – The conference for intelligent content solutions. Berlin, Germany (2020), cEUR Workshop Proceedings, Volume 2535. 20/21 January 2020
26. Rehm, G., Galanis, D., Labropoulou, P., Piperidis, S., Welß, M., Usbeck, R., Köhler, J., Deligiannis, M., Gkirtzou, K., Fischer, J., Chiarcos, C., Feldhus, N., Moreno-Schneider, J., Kintzel, F., Montiel, E., Doncel, V.R., McCrae, J.P., Laqua, D., Theile, I.P., Dittmar, C., Bontcheva, K., Roberts, I., Vasiljevs, A., Lagzdiņš, A.: Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability. In: Rehm, G., Bontcheva, K., Choukri, K., Hajic, J., Piperidis, S., Vasiljevs, A. (eds.) Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020). pp. 96–107. Marseille, France (2020), 16 May 2020

27. Schneider, J.M., Rehm, G.: Towards a Workflow Manager for Curation Technologies in the Legal Domain. In: Proceedings of the LREC 2018 Workshop on Language Resources and Technologies for the Legal Knowledge Graph. pp. 30–35 (2018)
28. Shobana, D., Phil, M.: Layout Based Information Retrieval from Document Images. IOSR Journal of Computer Engineering **4**, 31–35 (2012), http://www.iosrjournals.org/iosr-jce/papers/Vol4-issue4/E0443135.pdf
29. Sundheim, B. (ed.): Proceedings of the Sixth Message Understanding Conference (MUC-6). ARPA, Morgan Kaufmann, Columbia, MD (1995)
30. Uszkoreit, H., Gabryszak, A., Hennig, L., Steffen, J., Ai, R., Busemann, S., Dehdari, J., van Genabith, J., Heigold, G., Rethmeier, N., Rubino, R., Schmeier, S., Thomas, P., Wang, H., Xu, F.: Common Round: Application of Language Technologies to Large-Scale Web Debates. In: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics. pp. 5–8. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-3002
31. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200 (2020), https://dl.acm.org/doi/10.1145/3394486.3403172
32. Zhong, X., Tang, J., Jimeno-Yepes, A.: PubLayNet: Largest Dataset Ever for Document Layout Analysis. 2019 International Conference on Document Analysis and Recognition (ICDAR) pp. 1015–1022 (2019)