

Modelling Archival Hierarchies in Practice: Key Aspects and Lessons Learned

Mahsa Vafaie^{1,2}, Oleksandra Bruns^{1,2}, Nastasja Pilz³, Danilo Dessi^{1,2} and Harald Sack^{1,2}

¹FLZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany

²Karlsruhe Institute of Technology, Institute AIFB, Germany

³Landesarchiv Baden-Württemberg, Germany

Abstract

An increasing number of archival institutions aim to provide public access to historical documents. Ontologies have been designed, developed and utilised to model the archival description of historical documents and to enable interoperability between different information sources. However, due to the heterogeneous nature of archives and archival systems, current ontologies for the representation of archival content do not always cover all existing structural organisation forms equally well. After briefly contextualising the heterogeneity in the hierarchical structure of German archives, this paper describes and evaluates differences between two archival ontologies, ArDO and RiC-O, and their approaches to modelling hierarchy levels and archive dynamics.

1. Introduction

Online access to archival records as an important source of information has been the centre of interest for many researchers across different fields since the 1990s [1]. Digitisation of archival records increases accessibility and usability of archival data, by disseminating archival information to a wider group of people [2]. Today, through the use of collaborative knowledge bases and Linked Open Data datasets, cultural heritage institutions and archives can enrich their own collections or even foster the creation of new, authoritative and sustainable subject-specific datasets to increase public engagement and understanding of archives [3]. One of the main challenges in modelling archival information arises from the fact that archival practices vary at international and national levels, in spite of existing standards for archival description. These standards lack clarity for their use [4], thus leading several institutions to design and adopt their distinct and dissimilar models, reducing the possibility to link archives together and, as a consequence, limiting the chances to discover new knowledge. On the other hand, most archival records regarded as Cultural Heritage today were accumulated and structured before the standards were created.


HistoInformatics 2021 – 6th International Workshop on Computational History, September 30, 2021, online

✉ mahsa.vafaie@fiz-karlsruhe.de (M. Vafaie); oleksandra.bruns@fiz-karlsruhe.de (O. Bruns); nastasja.pilz@la-bw.de (N. Pilz); danilo.dessi@fiz-karlsruhe.de (D. Dessi); harald.sack@fiz-karlsruhe.de (H. Sack)

🆔 0000-0002-7706-8340 (M. Vafaie); 0000-0002-8501-6700 (O. Bruns); 0000-0003-3843-3285 (D. Dessi); 0000-0001-7069-9804 (H. Sack)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Despite the heterogeneity of archival practices in organisation of records, the classification scheme and hierarchical structure of the archives are of great significance when it comes to capturing interrelatedness of archival data and facilitating description, search and navigation of archival records. In fact, by properly modelling the structure of archives, they can be easily leveraged to discover new knowledge via supporting both practitioners and inexperienced users. As an example, a proper model enables archival records to be semantically described and linked to external resources such as The Virtual International Authority File (VIAF) [5] or Wikidata¹, opening the doors of archives to the Linked Open Data (LOD) cloud. However, in order to achieve this, archival data has to be modelled according to formal representations. Due to the variety of modelling approaches and lack of a comprehensive analysis, modelling archives with well-defined ontologies is still a challenge. This paper focuses on two different approaches for modelling archival hierarchy and archive dynamics, by providing information on case studies, using two different archival ontologies: the International Council on Archives Records in Contexts Ontology² (ICA RiC-O v0.2³) and the Archive Dynamics Ontology (ArDO) [6]. More specifically, the differences in the hierarchical structure of German archives and the representation of the hierarchy levels using these ontologies are discussed, pointing out the advantages and drawbacks of each model, and yielding insights into modelling archival hierarchy levels.

2. Standards and Archival Ontologies

Most German archives adopt the theoretical General International Standard of Archival Description ISAD(G) [7], when modelling their archival hierarchy. ISAD(G) introduces principles describing the type of information each hierarchical level should contain. Aside from theoretical standards, i.e., ISAD(G), the Encoded Archival Description (EAD) [8] standard provides an xml-format representation of multi-level archival descriptive information. It is able to reflect structures and relations of different information pieces and thus, is helpful for illustrating the hierarchy in archival records. Also, the so-called “principle of provenance”, stating that records should be maintained in organic units in which they are accumulated, has gained universal acceptance in the archival profession⁴.

Provenance-based multi-level hierarchies are helpful for both archivists and users: by maintaining an archival record in its original context it is easier to prove its authenticity and to understand the content of a record. Archival hierarchies reflect the logical relation between documents on higher and lower levels. Archivists and users need the hierarchy to track down records even if there is no topic-related or person-related index – which is the case for most German archives. Moreover, the hierarchy according to provenance is distinct, unambiguous, and not subject to the archivist’s personal preferences of collecting and arranging files.

During the last ten years, a new standard for archival description has been widely discussed.

¹https://www.wikidata.org/wiki/Wikidata:Main_Page

²ICA RiC-O, <https://www.ica.org/standards/RiC/ontology.html>

³Referred to as RiC-O in this paper

⁴<https://www.archives.gov/research/alic/reference/archives-resources/principles-of-arrangement.html>, accessed on 05.08.2021

Records in Context-Conceptual Model (RiC-CM) ⁵ aims at harmonising existing standards and is intended to form a complete model for archival description, taking advantage of developments in communication technologies and thus, representing archival description with semantic web technologies. Aside from the vertical hierarchy representing the principle of provenance, horizontal and plural relations between information elements as modelled in RiC-CM, offer new opportunities to standardise, structure and search for archival information. However, because of its complexity, RiC-CM has only rarely and selectively been adopted by German archives.

In spite of the existence of the aforementioned standards and principles, the division of archival records into subgroups (fonds, series, collections, etc.) and the level of granularity in the description always depends on the material and the archivist handling it. Even when the principle of provenance and ISAD(G) are applied, the hierarchical structure might consist of different number of levels. Archives might have different ways of naming and structuring record series and record groups, especially on the higher levels. There is neither a fixed terminology nor a fixed number of levels, but the standards offer a framework.

As a result of the growing interest in digitisation of archives, as well as the introduction of LOD in the past decades, several archival ontologies have been designed and developed by researchers in the field. The Europeana Data Model (EDM) [9], developed in the context of the Europeana portal ⁶, integrates various standards to facilitate data interoperability between different cultural heritage institutions, and provides a common model to deliver resources to scholars. Arkivo [4], developed in 2018, provides classes to model the structure of archives as well as the historical events. ArCo [10] is an evolving resource that includes a knowledge graph consisting of a network of ontologies, modelling the Italian CH domain and an LOD dataset that describes cultural properties and their catalogue records.

Two of the most recent developments in archival description are the Archive Dynamics Ontology (ArDO) and the Records in Context Ontology (RiC-O). ArDO is an ontology specifically designed for describing the hierarchical nature of archival data. RiC-O on the other hand, is a generic domain ontology that aims to produce a generalised description of archives, based on RiC-CM. A more detailed description of these two ontologies with focus on how they model the hierarchical structure of archives can be found in Section 4.

3. ArDO and RiC-O

ArDO. The Archive Dynamics Ontology (ArDO) is an ontology designed for describing the dynamic hierarchical nature of archival records [6]. ArDO reflects the hierarchical structure of the archive via classes. The core class of a single archival unit *ardo:Record* has been further extended by sub-classes *ardo:Portfolio*, *ardo:Chapter*, *ardo:Volume*, *ardo:Archive* and *ardo:Dossier*. They are disjoint classes and connect with each other via the object property *ardo:consists_of*. Such connection allows for different hierarchical allocations of archival records, e.g., an administrative record of class *ardo:Dossier* may consist of a record from class *ardo:Archive*, and vice versa. Digitising archives is a dynamic process, as it usually covers records piecemeal. Thus, some of the archives open parts of their holdings before the digitisation is completely finished.

⁵<https://www.ica.org/en/records-in-contexts-conceptual-model>

⁶<https://www.europeana.eu/en>

The nature of archival records as well as information contained in these documents may not always be predictable beforehand; however, unlike other ontologies, ArDO has a dynamic component that enables keeping track of changes during the digitisation process, e.g., changes in hierarchy of records or in archival classification schemes used for their semantic annotations. This is achieved by implementing a versioning mechanism that connects dynamic entities to their version via the property *pav:hasVersion*⁷.

RiC-O. RiC-O is a formal representation of RiC-CM, developed by the Expert Group on Archival Description (EGAD). A core entity in RiC-O is *rico:RecordResource*, with three sub-classes, *rico:RecordSet*, *rico:Record* and *rico:RecordPart*. According to the documentation, “determining when an information object is a Record, Record Part, or Record Set is based on perspective and judgement exercised in a particular context”. The classification of archival record groups in RiC-O is modelled with named individuals which are members of the class *rico:RecordSetType*, and are connected to *rico:RecordSet* through the property *rico:hasRecordSetType*. At the time of writing, the class *rico:RecordSetType* has four individuals in RiC-O, namely, *ric-rst:Collection*, *ric-rst:File*, *ric-rst:Fonds* and *ric-rst:Series*. Each of these named individuals in RiC-O are defined by the ICA ISAD(G) standards. The approach proposed by RiC-O provides the means to loosely link different types of Record Set through the property *rico:includesOrIncluded*, and accurately model the archival hierarchy according to the particular use case. For example, some archives allow Series to be part of Fonds and vice versa, whereas in other archives this is not allowed. Additionally, this modelling provides the possibility to introduce and define new members for the class *rico:RecordSetType* based on the use case and regardless of the number of levels in the hierarchical structure of the archives.

4. Use Case Modelling

4.1. Weimar Republic

Within the project “Subject Related Points of Access within Archivportal-D on Example of the subject area Weimar Republic”⁸, the German Federal Archives and the Baden-Württemberg State Archives have compiled 21,043 archival records that describe political and economic events, social and everyday life of German citizens from the period after the World War I and until the takeover of power by the Nazi Regime. In the future, the portal is to be supplemented by the digitised archival material from archives from all over Germany.

Due to the nature of archival records, they are stored in a hierarchical manner in a file system: *Bestand (Fonds/Portfolio)* is a collection of archives of one provenance; *Gliederung (File/Chapter)* groups archival records based on their topic; *Serie (Series/Volume)* arranges documents chronologically; *Archivale (Item/File/Archive)* denotes a complete file that may be extended by one or more *Vorgänge (Administrative record/Dossier)*. Figure 1 demonstrates the hierarchy of archival records based on the example of record 1abw-4-4012311 and its context—the list of ancestors of the archival record by traversing up the the file system hierarchy. RiC-O (see Figure 1a) assigns each of the archival units to the top class *rico:RecordResource*. To depict

⁷<https://pav-ontology.github.io/pav/>

⁸<https://www.archivportal-d.de/themenportale/weimarer-republik>

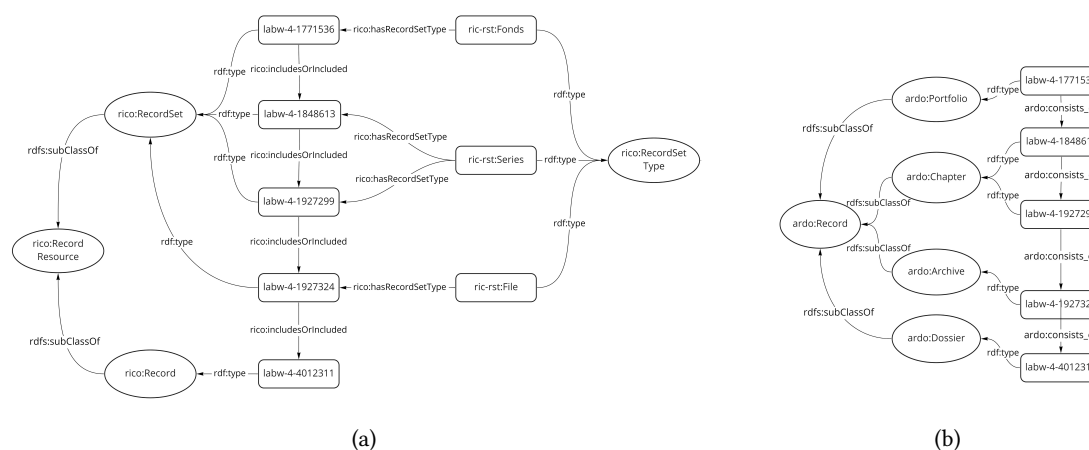


Figure 1: Modelling of the archival hierarchy within Weimar use case with (a) RiC-O ontology & (b) ArDO ontology

the hierarchical allocation the records are further divided into single records and record sets—records that physically reside together. Via the property *rico:includesOrIncluded*, records are connected with each other in a top-down manner, e.g., 1abw-4-1771536 with title “Margravia/Grand Ducal family archive” is the top element of the hierarchy that is further categorised into smaller sets, e.g., 1abw-4-1848613 entitled “Education, schools”. The smallest unit of the archival hierarchy 1abw-4-4012311 (“Bible quotes”) is a single record that may not include any further complete records. To make sense of the archival structure within record sets, the class *rico:RecordSetType* is inserted. An alternative modelling of the hierarchy of Weimar Republic archival records with ArDO ontology is depicted in Figure 1b. Similar to RiC-O, ArDO assigns any type of archival records to the top class *ardo:Record*. However, it addresses the hierarchical levels within the archive via classes, e.g., *ardo:Chapter* and *ardo:Archive*. The context is modelled from the top record down via the property *ardo:consists_of*.

4.2. Wiedergutmachung

“Transformation der Wiedergutmachung”⁹ is a pilot project issued by the German Federal Ministry of Finance. It is centred around archival data from the reparation process and reparation cases filed after World War II and the fall of the Nazi regime. Wiedergutmachung archival records originate from the offices of reparation installed by the German government in every state after the war. After the files had been closed and the offices were dissolved, the records have been collected by all nine West German state archives. The State of Baden-Württemberg alone holds 120,000 Wiedergutmachung case files in its departments in Freiburg, Sigmaringen, Ludwigsburg, and Karlsruhe.

The objects within these fonds are classified by a multi-level hierarchical system in accordance with ISAD(G) and EAD(DDB). The archive presents records at various levels which are labeled by level codes from *A0* to *J0*, reflecting the top-down hierarchy. Content information of each

⁹<https://www.fiz-karlsruhe.de/en/forschung/wiedergutmachung>

compensation file is provided on level F0. On this level, all documents contained in a case file relate to the application process of an individual applying for compensation. Each case file consists of several documents, starting with an application form filled in by the applicant and ending with a notice of compensation by the court. The case files include other letters and documents such as proof of evidence to the office of reparation as well. Many forms and documents in the case files contain a mix of stamps, machine-printed and handwritten text. The information content in each of these inscription forms corresponds to various creators and requires different text recognition technologies. Therefore, each record is further subdivided into its constituent parts, i.e., machine-printed, handwritten and stamp.

Figure 2 depicts two ways of modelling the hierarchy levels in Wiedergutmachung documents, based on RiC-O and ArDO, using an example of one case file 4-1583314. Adopting RiC-O (Fig. 2a), every archival resource in Wiedergutmachung is modeled as a *rico:RecordResource*. Three subclasses of *rico:RecordResource*, namely, *rico:RecordSet*, *rico:Record*, and *rico:RecordPart*, are used to model the hierarchy levels in each archive, starting from fonds, down to series, files, records, and record parts. The property *rico:includesOrIncluded* connects the higher levels of hierarchy to the lower ones. In this example, the individual 4-1835078 refers to all the files that start with the letter “A”. This subgroup in the classification scheme is then broken down into a smaller subgroup 4-1545483, with files starting from “Aa” to “Ad”. The smallest item in *rico:RecordSet* is the file, which is a collection of records constituting the case file of one person applying for reparations. Single records in this file are individuals of type *RiC-O:Record*. The property *rico:hasOrHadConstituent* connects a record to a record part that is a component of that Record. Different archival levels such as “fonds”, “series”, and “file” are modelled as individuals of the class *rico:RecordSetType* in RiC-O. The property *rico:isRecordSetTypeOf* connects these named individuals to individuals of type *rico:RecordSet*. The modelling of Wiedergutmachung records with ArDO can be seen in Fig. 2b. Here the hierarchy level is modelled using three classes, *ardo:Portfolio*, *ardo:Volume*, *ardo:Chapter*, and the class *ardo:Archive* is used to model single records in each case file. All these classes are sub-classes of *ardo:Record*. The property *ardo:consists_of* connects the higher levels of the hierarchy to the lower ones.

5. Advantages, Drawbacks and Insights

Hierarchy modelling via classes vs. named individuals. RiC-O and ArDO present two different approaches to modelling the hierarchy of archives. RiC-O provides general and flexible classes (i.e., *rico:RecordResource*, *rico:RecordSet*, *rico:Record*, and *rico:RecordPart*), the adoption of which is based on the use of archival resources within their specific context. A hierarchy is expressed by bounding archival resources to a set of individuals to describe which hierarchical level the archival resources belong to. Conversely, ArDO presents a more strict class hierarchy which combines the type of archival resources (e.g., single records or record groups) and hierarchical levels where archival resources are classified and kept at. This might considerably restrict the adoption of ArDO to scenarios which differ from the one used for its development. Different scenarios might demand the design of new classes and object properties, as a result of which, a drastic change in the original definitions of ArDO classes might be required. On the other hand, RiC-O is not bound to a specific archive and can be adopted independently, from

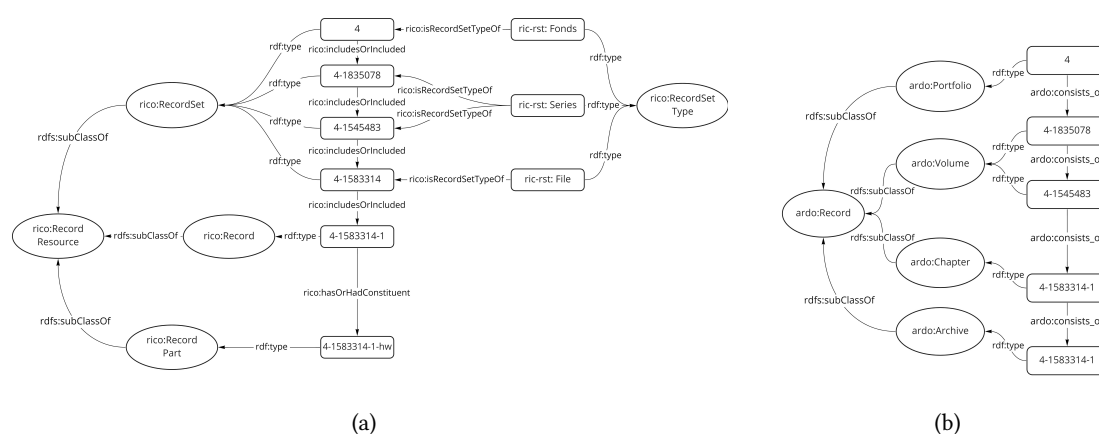


Figure 2: Modelling of the archival hierarchy within the Wiedergutmachung project with (a) RiC-O ontology & (b) ArDO ontology

archive to archive. However, potential issues might be raised due to the lack of well-defined relationships that can be expressed between archival hierarchical levels; thus, allowing to create inconsistencies when representing them. An example can be observed in Fig. 1a, where an *Archivale* is described through *rico:RecordSet*. In its context (i.e., “Weimar Republic”), *Archivale* cannot contain archival resources such as fonds or series. However, there are no constraints in the model which prevent its creation, and thus, a user of RiC-O has to correctly connect the archival resources and cannot rely on a reasoner to find inconsistencies. From a broad perspective, this implies that the semantics behind archival hierarchies cannot be made fully explicit.

Modelling record parts. As discussed in 4.2, the current version of ArDO does not allow modelling of entities smaller than a single archival record (*ardo:Archive*), which might limit the exploration of an archive. On the contrary, RiC-O provides the class *rico:RecordPart* which is defined to model objects constituting a record, containing “discrete information content”. Based on the definition, parts such as headings, stamps, graphic attachments, etc., can be designated a *rico:RecordPart* and therefore, RiC-O is more suitable for use cases in which bounded information regions within a single archival record are also of interest. This level of granularity in the modelling is more practical for usability reasons and matches the lowest level of hierarchy in ISAD(G). Moreover, this low-level description of archival resources enables modelling of the content, besides the structural modelling at the level of record groups.

Archival classification dynamics. The main goal of archival digitisation projects is to provide access to archival material across institutions for exploration and interoperability, as well as to make it understandable and searchable to the general public [11], who usually does not have the expertise in the hierarchical organisation of archives. To address the latter, during the digitisation process, domain experts develop a semantic classification scheme that assigns keywords to archival records based on their content [12]. In RiC-O, archival records can only be associated with dates, places, persons or events, while ArDO allows for assignment of any kind of linchpins that are stored in classes *ardo:Keyword*, *ardo:Subcategory* and *ardo:Category*.

Since the specific content of archival records is often unpredictable in the beginning of the digitisation, the development of a classification scheme is a dynamic process: the scheme is being continuously adapted while new records are discovered. Unlike RiC-O, ArDO enables to keep track of changes in the classification scheme by connecting every linchpin to its version. **Content modelling.** It should also be borne in mind when modelling archives that there is more than the hierarchy of archival resources that needs to be structured for a better exploration and understanding. In fact, archives might present information in the content of the material which require proper modelling. For example, archival resources from the “Weimar Republic” scenario might present different types of content such as images and texts which are interesting for an archive’s user, and thus, need to be properly represented within the model. For doing so, ArDO provides specific classes e.g., *ardo:Image* to define the content of an archival resource. Conversely, RiC-O proposes a modelling by providing the ontology users with a class (i.e., *rico:ContentType*) which makes it possible to define individuals and link them to archival resources. However, these individuals are not defined in the current version of RiC-O. As a result, archives modelled with RiC-O might have different individuals representing the same content type, and therefore reducing interoperability and understanding of different archives.

6. Conclusion

Working towards a comprehensive archival ontology requires a collective effort. To this day, there is no particular hierarchy specification with properties and classes that can accurately model all German archives of all times. Moreover, different institutes make use of varying archival terms, which are often not in accordance with the existing standards [13]. Such inconsistencies create a need for modelling strategies that can be modified according to the use case. In this paper, two approaches for modelling archival hierarchy based on two ontologies, ArDO and RiC-O, are presented. The modelling approaches are exemplified and illustrated with examples from “Weimar Republic” and “Wiedergutmachung” archival resources. The difference in the archival hierarchy in these two use cases is leveraged to point out the heterogeneity of hierarchical structures in German archives. Furthermore, the advantages and drawbacks of each of the approaches are discussed, providing insights into modelling archival hierarchy and archive dynamics.

The modelling strategy adopted by ArDO provides a more strict intensional meaning of its classes and model which might limit its dissemination. However, this enables creation of more homogeneous archival models, providing means for a correct interpretation of the modelled information. Moreover, the dynamic component in ArDO enables the adaption of the model to changes in classification scheme of archives. Alternatively, RiC-O’s conceptualisation of the hierarchy with named individuals, offers a strategy that is more generic, and thus, enables a wide spread of its use across archives of various historical periods and places. However, this flexibility might lead to different uses of the ontology classes, and therefore, creating discrepancies and limiting the advantages of the ontology’s adoption. Furthermore, the inclusion of smaller entities in the hierarchy in RiC-O increases findability and facilitates a more accurate modelling of archival resources. Extending RiC-O with the dynamic component of ArDO would increase the representation power of RiC-O while offering flexibility in the modelling of archival structure.

The challenge of integrating the peculiarities of different archival ontologies is nowadays being addressed by the archivists and Semantic Web communities and further developments might be expected in the next years to increase accessibility and explorability of archives, bringing people closer to their past. Opening up archives after digitisation requires more than just the structural representation of archival hierarchies via ontologies [14]. It is important to further identify and map persons, organisations, locations, and events to external resources and authority files in order to enable content-based and federated semantic search.

References

- [1] W. Duff, Archival mediation, *Currents of archival thinking* (2010) 115–136.
- [2] B. Altman, J. Nemmers, The usability of on-line archival resources: the Polaris Project finding aid, *The American Archivist* 64 (2001) 121–131.
- [3] J. Oomen, M. van Erp, L. Baltussen, Sharing cultural heritage the linked open data way: why you should sign up, in: *Museums and the Web 2012*, 2012.
- [4] L. Pandolfo, L. Pulina, M. Zielinski, ARKIVO: an Ontology for Describing Archival Resources, in: *CILC*, 2018, pp. 112–116.
- [5] R. Bennett, C. Hengel-Dittrich, et al., Vial (virtual international authority file): Linking die deutsche bibliothek and library of congress name authority files, in: *World library and information congress: 72nd IFLA general conference and council*, 2006.
- [6] O. Vsesviatska, T. Tietz, F. Hoppe, M. Sprau, N. Meyer, D. Dessí, H. Sack, ArDO: an ontology to describe the dynamics of multimedia archival records, in: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021, pp. 1855–1863.
- [7] B. Brothman, et al., Isad (g): General international standard archival description, *Archivaria* 34 (1992).
- [8] D. Pitti, Encoded archival description: The development of an encoding standard for archival finding aids, *The American Archivist* 60 (1997) 268–283.
- [9] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, H. Van de Sompel, The europeana data model (edm), in: *World Library and Information Congress: 76th IFLA general conference and assembly*, volume 10, 2010, p. 15.
- [10] V. A. Carriero, A. Gangemi, e. a. Mancinelli, ArCo: The Italian cultural heritage knowledge graph, in: *International Semantic Web Conference*, Springer, 2019, pp. 36–52.
- [11] F. Guernaccini, S. Mazzini, G. Bruno, LOD publication in the archival domain: methods and practices, in: *ODOCH@ CAiSE*, 2019, pp. 15–26.
- [12] F. Hoppe, T. Tietz, D. Dessí, M. Sprau, M. Alam, H. Sack, The challenges of german archival document categorization on insufficient labeled data, in: *Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020)*, 2020.
- [13] R. Brüning, W. Heegewaldt, N. Brübach, ISAD(G) - General International Standard Archival Description, 2nd Edition (DE), Technical Report, International Council for Archives (ICA), 2011.
- [14] O. Bruns, T. Tietz, M. Vafaie, D. Dessí, H. Sack, Towards a representation of temporal data in archival records: Use cases and requirements, in: *Proceedings of the International Workshop on Archives and Linked Data*, 2021.