ArDO: An Ontology to Describe the Dynamics of Multimedia Archival Records

Oleksandra Vsesviatska
oleksandra.vsesviatska@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Karlsruhe, Germany

Tabea Tietz
tabea.tietz@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Karlsruhe, Germany

Fabian Hoppe
fabian.hoppe@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Karlsruhe, Germany

Mirjam Sprau
m.sprau@bundesarchiv.de
German Federal Archives
Koblenz, Germany

Nils Meyer
nils.meyer@la-bw.de
Baden-Wurttemberg State Archives
Deutsche Digitale Bibliothek
Stuttgart, Germany

Danilo Dessì
danilo.dessi@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Karlsruhe, Germany

Harald Sack
harald.sack@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

Cultural heritage institutions store and digitize large amounts of multimedia data inside archives to make archival records findable by archivists, scientists, and general public. Cataloging standards vary from archive to archive and, therefore, the sharing and use of this data are limited. To solve this issue, linked open data (LOD) is rising as an essential paradigm to open and provide access to the archival resources. Archives which are opened to the world knowledge benefit from external connections by enabling the application of automated approaches to process archival records, helping all stakeholders to gain valuable insights. In this paper, we present the Archive Dynamics Ontology (ArDO) - an ontology designed for describing the hierarchical nature of archival multimedia data, as well as its application on the example of archival resources about the Weimar Republic. Furthermore, ArDO semantically organizes multimedia archival resources in form of texts, images, audios, and videos by representing the dynamics related to their classification over time. ArDO tracks the changes of a specific hierarchical classification schema referred to as systematics adopted to organize archival resources under semantically defined keywords.

SAC '21, March 22–26, 2021, Virtual Event, Republic of Korea
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8104-8/21/03…$15.00
https://doi.org/10.1145/3412841.3442057

CCS CONCEPTS

• **Information systems → Ontologies**; *Document topic models*; **Web Ontology Language (OWL)**; *Resource Description Framework (RDF).*

KEYWORDS

Multimedia Archive, Ontology Design, Ontology Dynamics

ACM Reference Format:
Oleksandra Vsesviatska, Tabea Tietz, Fabian Hoppe, Mirjam Sprau, Nils Meyer, Danilo Dessì, and Harald Sack. 2021. ArDO: An Ontology to Describe the Dynamics of Multimedia Archival Records. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21), March 22–26, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3412841.3442057

1 INTRODUCTION

Since the very beginning of Linked Open Data (LOD) efforts, the idea of making archival data more findable and accessible for users has been discussed within the cultural heritage community [19]. However, until today many cultural heritage institutions keep on generating silos – data buried inside archival information systems. Taking into account that the archival storing and cataloguing standards vary drastically from archive to archive, it is often impossible to adapt and share existing ontologies that are able to reflect all the metadata standards. This limits the sharing, interconnection, and enrichment of data, thus reducing the awareness and impact of cultural heritage resources on society [5]. In recent years, archives in Germany have been storing and digitizing large amounts of cultural heritage multimedia data. To help archivists, historians and the general public to access and explore these resources, web-based platforms are being created to provide access to rich archive records. One of these platforms is Archivportal-D[1], the German Archives Portal, which offers a sector-specific access to the data of the German Digital Library (Deutsche Digitale Bibliothek, DDB)[2] enabling access to information on archive institutions from all over Germany. The project "*Subject Related Points of Access within Archivportal-D on Example of the subject area Weimar Republic*" is based on the Archivportal-D platform with the goal to make a specific selection of multimedia archival records, which are related to the Weimar Republic – the first German democracy, findable and accessible.

In the platform, various core conditions, such as the variety of content descriptors and different users' needs, must be considered to exploit the cultural heritage data. Within the project, a key requirement was to establish a linked data model to represent the organizational semantics behind archives. In fact, arranging the data in a structured semantic model allows users to explore the archive via sophisticated semantic search. For example, through external links the information represented in the archive can be enriched with additional external information. Moreover, since data includes thousands of archival records, the modelling of data as linked data allows its automatic processing, speeding up the archive exploration and analysis. One more requirement for delivering archival resources to the public involves their annotation with semantic metadata that might also help non-experts to find the appropriate content, e.g., via semantic search and faceted browsing. In order to accomplish this requirement, a new hierarchical subject classification further on referred to as *systematics* has been defined and implemented by archivists to facilitate subject based annotations for archival records. One of the major challenges of systematics is its dynamics due to structural variations triggered by newly available archival content for which no appropriate annotation vocabulary exists so far. Structural changes within systematics lead to relevant semantic transformations which must be taken into account within the archive linked data model.

This paper presents the Archive Dynamics Ontology[3] (ArDO), a novel ontology with 12 classes and 19 relationships designed to represent the evolving semantics of multimedia archives. In detail the contribution of this paper is twofold:

- An ontology component to describe subject related access to multimedia archives within the Archivportal-D platform, as well as to depict the hierarchical structure of the archival records.
- A novel ontology component to capture and track changes over time within a hierarchical classification scheme (systematics) employed by archivists.

The remainder of this paper is organized as follows. Section 2 presents the related work and points out the existing gaps. Section 3 depicts the target scenario. In Section 4, the details of the ArDO ontology are presented. Section 5 discusses the proposed design and its use to deal with competency questions. Finally, Section 6 concludes the paper and highlights future work.

2 RELATED WORK

Shared and linked data is surely a common need to adhere to interoperability standards for preserving the richness of data and making use of archival resources [25]. In fact, in literature ontologies and linked data have already proven to be key technologies for supporting practitioners to explore and consume archival records [15, 16]. To name an example within the cultural heritage domain, a recent linked data resource is Linked Stage Graph [23], a Knowledge Graph (KG) on the foundation of historical data released by the Baden-Württemberg State Archives about the Stuttgart State Theatres[4]. However, the need to deal with multimedia archival resources for specific purposes still persists (e.g., discovering potential semantic links that might exist between multimedia archival resources to increase findability and facilitate intuitive exploration [9]). Examples of existing semantic technologies that among others can be leveraged in this context are well-known data models for cultural heritage [6] as well as ontologies such as Bibo[5], FaBiO [21], RiC-O[6], ArCo [4], and Arkivo [20].

CIDOC Conceptual Reference Model (CRM)[7] is a widely used cultural heritage model with the goal to provide information integration and exchange between heterogeneous resources. However, one of its limitations is caused by the fact that it is not implemented for specific use cases which often require to model application-dependent aspects. Another relevant model in this context is the Europeana Data Model (EDM) [7] which integrates various standards to facilitate data interoperability between various cultural heritage institutions, and provides a common model to deliver resources to citizens through the Europeana portal[8]. For our target scenario, CIDOC CRM and EDM models were currently considered only as guidelines since there are specific application requirements (e.g., the versioning of archival records arrangement under keywords) that strongly depend on our target application.

Due to the similarities between archival resources and published textual documents, ontologies describing document organization and their annotations can be found in the editorial domain. In fact, tasks performed to catalog resources, such as locating a resource under certain categories, are performed by both librarians and archivists. In the editorial domain, Bibo is an example of an ontology targeted to represent documents in RDF. It can be exploited both as a document classification and citation ontology, and is particularly suited for describing bibliographic references. It has *bibo:Document* as a core class, and models documents in narrower classes such as *bibo:AcademicArticle*, *bibo:Journal*, *bibo:Book* to describe its elements. Another relevant resource is FaBiO, an ontology developed according to the Functional Requirements for Bibliographic Records (FRBR) for describing published textual resources and their references with the goal of supporting the semantic publishing task. Resources can vary from academic papers, books, newspapers, vocabularies, and so on. It also allows to model archival resources.

[1] https://www.archivportal-d.de/
[2] https://www.deutsche-digitale-bibliothek.de/
[3] The reader can find the ontology at the link: https://github.com/ISE-FIZKarlsruhe/ArDO. A full documentation according to https://www.w3.org/TR/swbp-vocab-pub/ is provided in the repository. Please also note that ArDO as well as a Knowledge Graph describing DDB archival resources will be further deployed and publicly available within the Archivportal-D platform.
[4] http://slod.fiz-karlsruhe.de/
[5] https://www.dublincore.org/specifications/bibo/bibo/bibo.rdf.xml
[6] https://www.ica.org/standards/RiC/ontology.html
[7] http://www.cidoc-crm.org/
[8] https://www.europeana.eu/en

Among its classes, FaBiO describes archival records. However, it does not accommodate specific domain requirements, i.e., various types of archival records and their relations are not taken into account. This mostly depends on the fact that archival records are not intended to be published. Therefore, Bibo and FaBiO can only partially be exploited to describe archive peculiarities.

This gap has recently been reduced by further developed cultural heritage ontologies RiC-O, ArCo and Arkivo. RiC-O is an ontology designed to describe archival records. It takes into account the need of producing a generalized description of archives. For example, it provides a single class for representing any kind of records, or agents, i.e., people who store or utilize the records. However, as it is stated in the documentation it might not satisfy specific requirements made by an archival institution. Specifically to our use case, RiC-O is limited in representing the hierarchical structure of our archive as well as the dynamics behind the adopted classification schema. ArCo is an evolving resource in form of a KG that connects various ontologies about Italian document collections and artifacts, and provides ontology patterns to connect people and locations, about cultural heritage events. Arkivo was developed in 2018 and provides classes to model the structure of archives as well as the historical events. However, its class hierarchy for describing archival records can only partially fit our target scenario where an archival record can belong to multiple other archival records. In addition, as a requirement in our ontology we need to provide provenance information about the arrangement of archival records under semantic keywords, which can be subject to changes due to new analyses performed on the underlying multimedia archival record contents.

In the beginning of the digitization process of archival resources, a classification scheme, i.e., a systematics, is usually introduced. However, it can change overtime. Some of the concepts may become obsolete and are deleted from the schema, though they have been already used for record annotations. Therefore, the additional contribution of the presented paper is the introduction of a dynamic storage scheme which supports time-varying instances of the ontology. In contrast to ontology evolution, when only one ontology version is maintained and the old versions cannot be retrieved, some domains require to keep track of the various instances across all versions. The semantic versioning of a systematics presented in this paper is inspired by the medical [10] and e-Government [11] domains, which contain dynamic information such as patients symptoms that change overtime and the division of people according to specific groups - unemployed, self-employed, employed, respectively. In this paper, the dynamics of a systematics in the context of archival records are addressed, however the proposed method may further be applied to any domain with a non-static classification scheme, e.g. blog posts, recipes. In Section 4.2 the semantic versioning method that manages changes of systematics concepts and their semantic relationships are described.

3 USE CASE: ARCHIVAL RECORDS ON WEIMAR REPUBLIC

ArDO was created to describe the dynamics of multimedia archival resources. Archives have the goal to collect historical records, store these records long-term and make these records accessible to researchers as well as the general public [22]. In order to fulfill these goals, archive resources have to be categorized to provide a structured and intuitive access for search and exploration to their users. In recent years, archival platforms (e.g. *Archivportal Thüringen*[9], *Archives Portal Europe*[10], *Archive in Nordrhein-Westfalen*[11]) have been created with the intent to provide access to a specific subdomain of archival resources. For instance, this might refer to a certain time period, geographical region or a specific person or organization of interest. This topic based access often also combines archival records of multiple archives in one platform. Archival records are special within the cultural heritage domain in two ways: (1) They were authored at a point in history without the original intent of being published and read by the general public. (2) Each record is unique and is stored by a single archive. Therefore, new classification schemes have to be created for subject specific entries depending on the topic of the archival records. In contrast, libraries are able to reuse more general already existing and established classification schemes and allow to categorize (at least to some extent) shared records and topics. Furthermore, whenever another archive provides access to additional records within the same subject specific platform, and whenever new records are digitized, the classification scheme has to be accordingly adapted, i.e., it is highly dynamic. When creating an ontology that models the classification of these topic based archival resources, these dynamics have to be taken into account.

As initial example ArDO is utilized for the subject specific access of the Weimar Republic created during the project "*Subject Related Points of Access within Archivportal-D on Example of the subject area Weimar Republic*". Due to the 100 year anniversary of the Weimar Republic in 2018 a noticeable demand for historical resources of that time evolved from historical researchers as well as from the German general public. On that account, this sub-domain has been considered well suited for a subject specific access. Two digitization projects by the German Federal Archives and the Baden-Württemberg State Archives have compiled a large number of relevant archival records from ministries, public institutions, corporate bodies and noteworthy individuals from this particular period to be digitized and described [13]. These 21,043 records provide the foundation for the subject specific archival platform. Consequently, it is the data basis for ArDO as well. It covers all aspects of politics, economy, society and everyday life in Germany from 1918 to 1933. For instance, there are records related to the Versailles peace negotiations, election campaign posters, food provisions and monetary inflation, handwritten letters from former nobility and monarchs, and many more. The digitized records consist of descriptive metadata and digitized multimedia files. Archivists create metadata to describe the content of archival documents to make it findable and accessible. Each document is given a title describing the content in a concise manner and if necessary an abstract containing more details about the content of each record. Due to the special nature of archival records, they are stored in a hierarchical manner in a file system. This creates a context, i.e., the list of ancestors of an archival record by traversing up the the file system hierarchy, which

[9] https://www.archive-in-thueringen.de/de/
[10] https://www.archivesportaleurope.net/
[11] https://www.archive.nrw.de/

denotes a substantial part of the content and has to be considered for any classification or retrieval.

The necessary classification scheme has been developed by domain experts, i.e., archivists and historical researchers. Based on their expert knowledge, the available data of the Archivportal-D and existing classification systems from LeoBW[12], Wikipedia[13] and IPTC NewsCodes[14], 881 specific subject keywords have been devised. In order to further structure these highly specialized keywords 17 categories and 121 subcategories have been introduced, as, e.g., "government and administration", "foreign affairs", "society" and "media". Keywords are linked to these broader categories, which enable the retrieval of fine-grained keywords by narrowing down their more general topic. This goal in mind, the hierarchical subject classification allows to link a keyword to multiple categories. Therefore, the systematics differs from a strict hierarchical structure being obeyed in most taxonomies, which assign each keyword only one higher level category. Aside from the subject classification an additional geographical classification has been introduced due to the regional context of many documents. It follows the same structure and principles as the subject classification.

4 ARCHIVE DYNAMICS ONTOLOGY DESIGN

The Archive Dynamics Ontology (ArDO) is an ontology for multimedia archival resources focusing on the dynamics of archives. The ontology was developed based on historical documents of Weimar Republic that are being digitized by the German Federal Archive and the Baden-Württemberg State Archives. The aim of ArDO is to structure the metadata information obtained from the archives, capture historical knowledge through a dynamic logical conceptual framework which is designed to classify archival resources, and expand this knowledge by enriching steps involving external resources. ArDO was created as one of the final contributions of the project to make the archival resources and their metadata available as LOD. More specifically, the ontology design phase started after the domain experts had created the systematics, and continued to adapt to the changes within the semantic model. Thus, a well established methodology (e.g., as defined in [1]) of ontology construction was not fully applied. The current version of ArDO includes 12 classes and 19 semantic relationships. It is available online and builds on SKOS[15] [17], PAV[16], Web Annotation Ontology[17], FOAF[18] [3], Bibo[19] and OWL[20]. ArDO was designed with two core components:

- *MultiArch* which organizes the archival resources and makes sense of the archive structure.
- *DynSyst* which allows the management of updates within the systematics while preserving versioning information.

The reader will find further details about ArDO's core components in the subsequent sections.

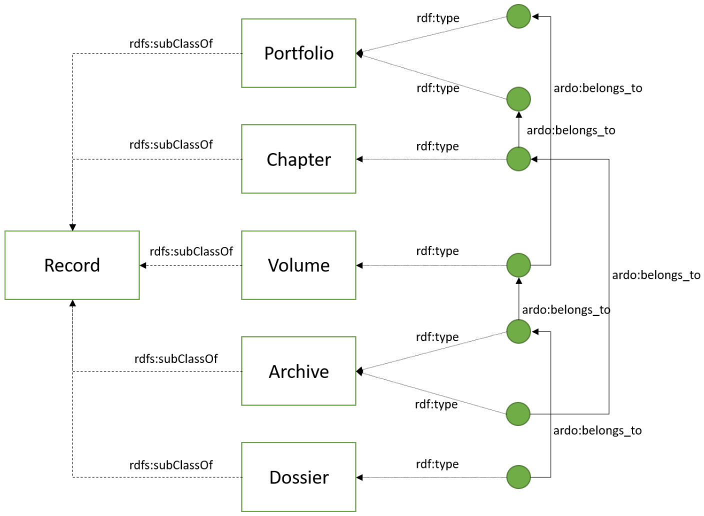

Figure 1: Hierarchical structure of archival documents.

4.1 MultiArch

Records of the presented use case are unique and stored by a single archive. Because of this peculiarity, a new class `ardo:Record` is first defined which differs from existing classes to model archival records. For example, the class `rico:Record` allows multiple instantiations of the same record which can be considered different based on the context or function it serves. This does not apply to our archive. Archival resources typically are organized in a hierarchical file system. To reflect this hierarchical structure in the ontology, the class `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:belongs_to` (see example in Figure 1). The class `ardo:Portfolio` is a collection of archives of one provenance and is always the top level element of the hierarchy. Archives and dossiers are the smallest units of a portfolio, their allocation may vary depending on an archival resource (i.e., various paths between an archive or dossier and a portfolio can exist). For example, the archive resource "*Novemberrevolution 1918*" belongs to chapter "*Kriegszielbewegung im 1. Weltkrieg*" which in turn belongs to chapter "*Einzelne Aktionsbereiche*", and then to a portfolio "*Alldeutscher Verband*". While archive "*Soldatenrat der Art. Prüfungskommission Schießplatz Kummersdorf (Krs. Zossen/Bez. Potsdam)*" is assigned directly to portfolio "*Arbeiter- und Soldatenräte in Deutschland*".

Besides the most common metadata used for describing archival resources such as title, language of the document, introduction, provenance, persons and time period, the ontology provides properties to denote the location of a physical archival resource, the number of samples available in archive, material and size of a physical portfolio, as well as the date of the document creation.

In addition to archival text resources ArDO describes multimedia records such as images, audio, video, films. The property `ardo:type_of_media` points out the media type of an archive (e.g., TEXT, IMAGE, AUDIO). Similar to textual documents other

[12] https://www.leo-bw.de/
[13] https://de.wikipedia.org/wiki/Kategorie:Weimarer_Republik
[14] https://iptc.org/standards/newscodes/
[15] https://www.w3.org/2004/02/skos/
[16] https://pav-ontology.github.io/pav/
[17] https://www.w3.org/ns/oa
[18] http://xmlns.com/foaf/spec/
[19] https://www.dublincore.org/specifications/bibo/bibo/bibo.rdf.xml
[20] https://www.w3.org/OWL/

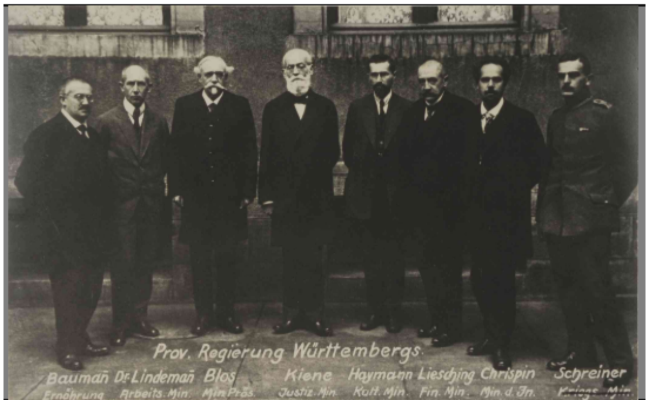

Figure 2: An example of graphical content associated to the archive "*Provisorische Regierung Württembergs von 1918*".

multimedia records may serve as full-fledged archival resources with their own title, provenance information, year, event or depicted persons (the reader can see an example in Figure 2 which shows a multimedia archival resource publicly accessible at the URL[21]). In case a resource provides additional information to a textual archive, the class `bibo:AudioVisualDocument` is used for storing films, audio files, and videos, and `bibo:Image` for storing still-images. For instance to address photos related to a textual archive a sub-class `ardo:Archival_Photo` has been defined. The photos can be mapped to the class `ardo:Record` via property `ardo:provides_multimedia`. Since by the time of writing in our use case the majority of archival text resources do not provide a transcribed version, it is useful to map the available metadata with at least the digitized content information as depicted in Figure 3, where the archival object located at the URL[22] has not been transcribed yet. For storing such resources a sub-class `ardo:Scan` has been introduced. Following the same logic subclasses for `bibo:AudioVisualDocument` have been implemented: `ardo:Film` to store professional documentary videos that may relate to a specific archival text document; `ardo:Video` for amateur videos; and `ardo:Audio` to preserve original audio recordings such as radio recordings or telephone conversations.

MultiArch provides the following semantic relations:

- `ardo:belongs_to` indicates that an archival object resides under another archival object. To maximize the inference capability, transitivity and reflexivity characteristics have been inserted to the property, and the inverse property `ardo:consists_of` has been implemented.
- `ardo:located_in` (inverse of `ardo:location_of`) and `ardo:findsystem` relate an archival resource to its physical and digital locations respectively.
- `ardo:provides_multimedia` (inverse of `ardo:multimedia_of`) is a relation between an archival record and the additional multimedia information that is related to a textual archive.
- `ardo:mentions` (inverse of `ardo:mentioned_in`) relates an archival object to a person that is mentioned/depicted in it.

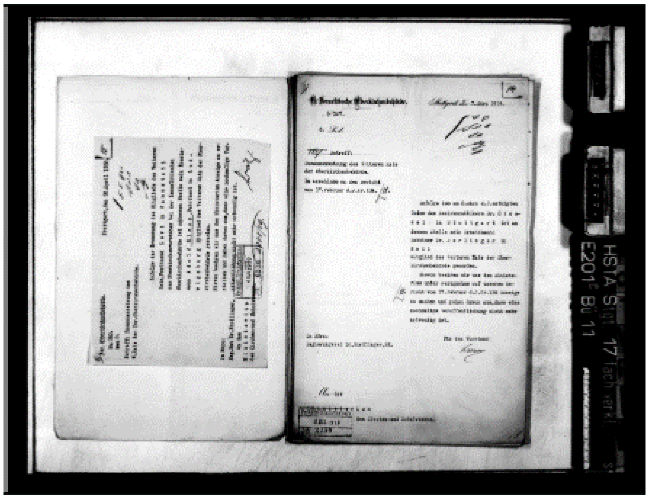
HSTA Stgt E 201 c Bü 11 17

Figure 3: A digitized page from archive "*Israelitische Oberkirchenbehörde: Wahl der Mitglieder des weiteren Rats*".

Finally, the MultiArch ontology component connects to the DynSyst component via the relation `ardo:tagged_with` that connects an archival object to semantic keywords it was annotated with.

4.2 DynSyst

One of the most important milestones in digitizing archival material is to make it available for general users who may not be aware of specific metadata such as the correct title of an archive. Thus, domain experts usually work on presenting semantic linchpins between what people search and the multimedia resource that is appropriate to fulfill this specific information need. The usual unit of semantic information available for an archival object are keywords assigned to a document. To store such units in ArDO the class `ardo:Keyword` has been integrated. The object property `ardo:tag_of` (inverse of `ardo:tagged_with`) represents the relation between an archive and a keyword. A new class is defined to represent keywords within the archival domain because the process used for defining a new keyword and for assigning it to archival resources complies with well-defined procedures followed by archivists. Therefore, keywords defined for archives semantically differ from those already used in other domains (e.g., those used for phylogenetic studies [8]) and need new definitions.

In addition, since the keyword-based annotation may vary from the introduced keyword vocabulary, e.g., instead of keywords from the vocabulary the domain, synonyms or hypernyms are applied, the set of annotations might become inconsistent and difficult to use. To address this problem, domain experts provide a specific hierarchical classification system, i.e., a *systematics* containing semantic concepts of different hierarchy levels. Such classification frameworks enable users to approach the more specific concepts via generally understandable subject blocks, and subsequently to narrow down

[21] https://www.deutsche-digitale-bibliothek.de/item/GNRBEXSU54B7XEC5M4PNGQLMROSPEY3S

[22] https://www2.landesarchiv-bw.de/ofs21/bild_zoom/zoom.php?bestand=4209&id=65716&gewaehlteSeite=01_0000074846_0002_1-74846-2.jpg&leo=1&screenbreite=1280&screenhoehe=720

their own area of interest. However, it is worth to note that this classification scheme is highly dynamic, strongly depends on the target archive, and must be accordingly adapted to provide access to additional new records within the same subject specific platform. Hence, in the design process it was agreed with the experts to depict this categorical hierarchy with a depth of three: the most specific semantic annotation `ardo:Keyword` is related to a general class `ardo:Subcategory` via the object property `ardo:is_keyword_of`, while `ardo:Subcategory` is linked to a root class `ardo:Category` via `ardo:is_subcategory_of`. Both `ardo:is_keyword_of` and `ardo:is_subcategory_of` are sub-properties of `skos:broader`. Similarly, the two new properties provide a direct hierarchical link between two concepts, however their semantics refined by range restrictions is more expressive for our use case.

ArDO allows multiple assignments of keywords. For example, the keyword *Abortion* may be examined from health, legal or social perspective, and accordingly be asserted to subcategories such as *Crime and Criminal Justice* with the upper category *Justice and Law Enforcement*, subcategory *Health Care* and upper category *Health*, and finally to *Family and Marriage* from *Society and Social Issues*.

Digitization is a dynamic process, as it usually covers documents piecemeal. As information contained in these documents may not always be predictable in the beginning of the digitization process, we argue that categorical systematics should not be regarded as 'static'. Unlike other ontologies, ArDO enables to keep track of changes in a systematics by connecting its entities with the version of the systematics. DynSyst covers the following aspects of a systematics refinement:

- A new concept is added to the systematics (classes `Keyword, Subcategory, Category`).
- A concept is deleted from the systematics.
- A concept changes its hierarchical level, for example, a keyword becomes a subcategory.

DynSyst covers all changes of the systematics by applying two atomic operations: insertion and deletion.

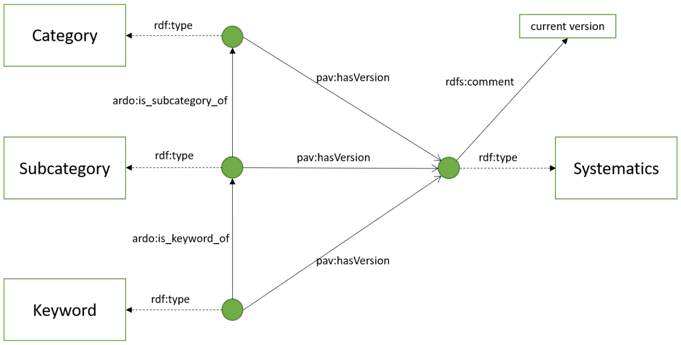

Figure 4: A sample systematics. The reader can see how keywords, subcategories, and categories are linked among each other and to the systematics individuals.

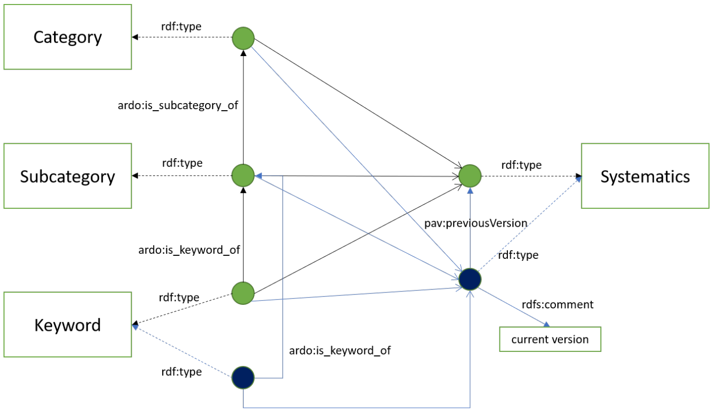

Figure 5: The illustration of an insertion operation exemplified by the case of adding a new keyword to a systematics. Open arrowheads illustrate the relation `pav:hasVersion`.

An initial systematics is depicted in Figure 4, where all individuals that exist in the current classification scheme are connected to an individual of the class `Systematics` (e.g., *"systematics1.0"*) via property `pav:hasVersion`. Since with every subsequent version a new individual is added to the class `Systematics`, a pointer denotes the latest version. This is indicated by adding the RDF triple, e.g. `<ardo:systematics1.0, rdfs:comment, "current version"@en>`. The following subsections describe the primitive operations that are applied to monitor changes in the systematics.

4.2.1 *Insertion.*

Insertion is used if a new concept (i.e., a keyword, subcategory, or category) is added to the systematics. DynSyst introduces five steps to display the process in the ontology:

1. A new individual (e.g., the keyword *gendarmerie*) is defined.
2. Relations between the new individual and the concepts of other hierarchical levels via `is_keyword_of` and/or `is_subcategory_of` are created (e.g., the keyword *gendarmerie* is linked to the subcategory *police*).
3. A new individual of class `ardo:Systematics` is created (e.g., *systematics1.1*).
4. All already existing individuals of classes `ardo:Keyword`, `ardo:Subcategory`, and `ardo:Category` that also exist in the new systematics are linked to it via `pav:hasVersion` (*police*, *gendarmerie*, etc. are connected to *systematics1.1*).
5. The pointer "*current version*" is deleted from the previous systematics individual and added to the latest systematics individual.

Figure 5 illustrates the workflow of systematics refinement using the example of a new keyword that was added to a systematics.

4.2.2 *Deletion.*

During the digitization process some concepts introduced in the beginning may become unrepresentative or absorbed by other concepts, and thus deleted from a systematics. Figure 6 demonstrates the deletion procedure in DynSyst which corresponds to steps 3-5 in Section 4.2.1.

In case a concept changes its hierarchical position in a systematics, for example, the keyword *political party* becomes a subcategory which then covers more specific keywords such as *German People's*

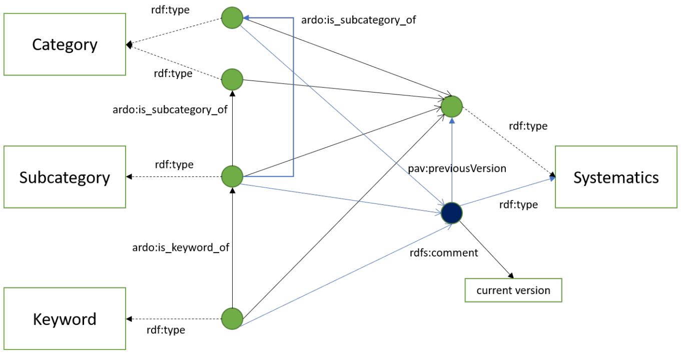

Figure 6: Deletion procedure exemplified by the case of deleting a root category from a systematics. Open arrowheads illustrate the relation `pav:hasVersion`. The reader notices that an individual of class `ardo:Category` is not linked to the new `ardo:Systematics` individual.

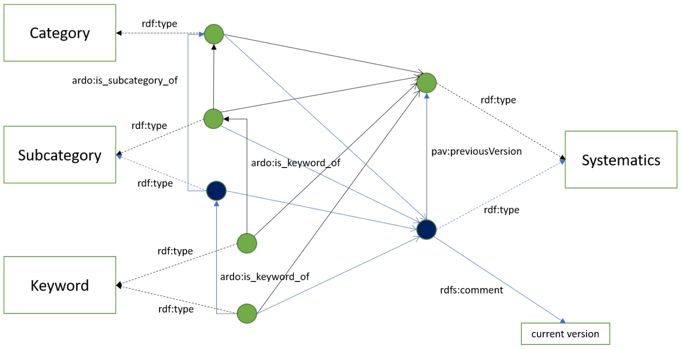

Figure 7: Deletion and insertion procedures when a keyword shifts its position to a subcategory. Open arrowheads illustrate the relation `pav:hasVersion`.

Party and *National Socialist Freedom Movement*, the union of deletion and insertion procedures is applied. Referring to our example, the keyword *political party* follows the steps of deletion, while the subcategory *political party* along with keywords *German People's Party* and *National Socialist Freedom Movement* are newly inserted into the systematics. Figure 7 depicts the workflow of systematics refinement using the example of a hierarchical level change.

4.3 Mapping to External Resources

ArDO supports the mapping of concepts in a systematics to external resources, e.g., entities of general KGs like Wikidata [24] or Integrated Authority File (GND) [2] via `owl:sameAs`. Wikidata was launched in order to create a shared knowledge base and to connect structured data between all Wikipedia projects, while GND was developed to integrate the content of Name Authority File (PND)[23], Corporate Bodies Authority File (GKD)[24], Subject Headings Authority File (SWD)[25] and Uniform Title File of the Deutsches Musikarchiv (DMA-EST)[26]. Today both are used as central hubs to lift concepts from a closed to a public environment [18]. As already mentioned in Section 3, systematics and taxonomies for document labeling vary depending on the domain they are used for. Thus, if searching for more general concepts of a keyword, e.g., "*espionage*", the result will depend on the specific domain the systematics was introduced in. For example, in an encyclopedic or bibliographic systematics "*crime*" might be the appropriate classification, while in a systematics designed for archival resources the appropriate classification might be "*institutions of foreign policy and sanctions*".

Despite the different hierarchical assignment, mapping concepts from a local classification scheme to an external resource is fundamental for expanding the knowledge of these concepts. Such mapping provides access to information that is not available from archival records and allows to find data interconnections, which can be used to improve search and retrieval. For instance, it enables users to search an archival resource not only by a specific entity mentioned in a systematics, but also by concepts related to it (e.g., accessing archives tagged with "*Genoa Conference (1922)*" by searching for information about "*Walther Rathenau*"). Furthermore, it highly supports the automation of archive document classification, as e.g., described in [14], where only limited descriptive metadata is available for semantic document classification. Since the concepts represented in archival systematics are often highly domain specific, in particular regarding geographic location and time, general KGs such as Wikidata might not contain appropriate identical mapping entities. To solve this issue, mapping has been relaxed by utilizing SKOS mapping vocabulary, in particular `skos:broadMatch`, `skos:narrowMatch`, `skos:closeMatch`, `skos:relatedMatch` to denote a more relaxed mapping for systematics concepts without exact matching Wikidata counterpart.

5 ONTOLOGY VALIDATION

In this section, the validity and correctness of the proposed ontology design are discussed.

5.1 Verification and Error Provocation

During ontology development the reasoner HermiT [27] (version 1.4.3.456) was regularly run in order to detect incoherence in the ontology design. In addition, ArDO was tested by manually creating inconsistencies in the ontology. In particular, individuals were injected as belonging to disjoint classes (e.g., `ardo:Volume` and `ardo:Archive`). It was evaluated whether the reasoner is able to spot introduced errors. Potential detected issues have been considered and fixed in the subsequent design step.

5.2 Competency Question Verification

An evaluation of the ontology has been performed with respect to its requirements by verifying the appropriateness of the ontology in delivering correct answers to a set of competency questions (CQs) defined by the archival domain experts. Goal of this evaluation was to test whether the CQs could be converted into SPARQL queries

[23] https://en.wikipedia.org/wiki/Name_Authority_File
[24] https://en.wikipedia.org/wiki/Corporate_Bodies_Authority_File
[25] https://en.wikipedia.org/wiki/Subject_Headings_Authority_File
[26] https://portal.dnb.de/opac.htm?method=simpleSearch&reset=true&cqlMode=true&query=partOf%3D01670763X&selectedCategory=any
[27] http://hermit-reasoner.com/

```
PREFIX pav: <http://purl.org/pav/2.3#>
SELECT ?keyword_label ?subcategory_label
WHERE { ardo:S104 rdfs:label ?subcategory_label .
        ?keyword ardo:is_keyword\_of ardo:S104 ;
            pav:hasVersion ardo:systematics1_0 ;
            rdfs:label ?keyword_label . }
```

Figure 8: The SPARQL query for the CQ11 targeting the subcategory with identifier S104.

Table 1: Sample of result of the CQ11 SPARQL query.

keyword_label	subcategory_label
Bürgermeister	Kommunalpolitik
Kommunale Finanzen	Kommunalpolitik
Kommunalpoltik	Kommunalpolitik
Kreisverwaltung	Kommunalpolitik
Landkreis	Kommunalpolitik
Landratsamt	Kommunalpolitik
Städtetag	Kommunalpolitik
Stadtrat	Kommunalpolitik

through the defined ontological vocabulary (cf. Figure 8 as an example). The CQs consider all modeled aspects of the ontology such as the organization of the archival resources, the links between multimedia data representing the same archival resource, the annotations, and finally the possibility to enrich the ontology knowledge through external links. Moreover, CQs have been adopted to analyze the dynamic connections between the elements of the systematics and archival records. In particular, keywords are often deleted if they are too broad or too specific for a set of records, i.e. keywords are substituted if they do not represent the archival records well enough to provide the user with intuitive exploration possibilities. The CQs are listed in the following:

(1) *Given a resource with title X, which other archival resources does it belong to?*
(2) *What are the titles of records labeled with keyword X?*
(3) *Which images are linked to a text document with title X?*
(4) *Which text documents are linked to image X?*
(5) *What locations are associated to the archival resource X?*
(6) *What is the definition of keyword X?*
(7) *What is the description of location X?*
(8) *Are all keywords used to label document X still valid under the latest version of systematics?*
(9) *What keywords are associated with category X in different existing versions of the systematics?*
(10) *What keywords have been added to the latest version of the systematics?*
(11) *What keywords belonged to subcategory X according to systematics Y?*
(12) *What categories belonged to every existing version of the systematics?*

For example, Table 1 shows the result of the SPARQL query in Figure 8 for CQ11. Issues spotted by the testing, for example missing concepts (e.g., classes `bibo:Image` and `bibo:AudioVisualDocument` for multimedia resources that do not serve as separate archival files but complement the textual resources), relations such as `ardo:provides_multimedia`, and property characteristics have been considered in a subsequent design phase, and the ontology has been updated accordingly.

5.3 OntoClean Validation

In addition, ArDO has also been validated according to the *OntoClean* methodology [12]. OntoClean distinguishes four so-called metaproperties: rigidity, identity, uniformity, and dependency, for which explicit inheritence rules must be fulfilled within the evaluated ontology. Considering the limited size of ArDO's subclass hierarchy, this evaluation was restricted to the ontology part depicted in Figure 1.

Rigidity: While `ardo:Record` according to OntoClean terminology must be considered *rigid*, i.e., it is essential for the existence of an individual, its subclasses `ardo:Portfolio`, `ardo:Chapter`, `ardo:Volume`, `ardo:Archive`, `ardo:Dossier` are considered *not rigid properties*, because they are not essential for the existence of an instance, which is still a `Record` although it might not be considered to belong to one of the remaining ordering criteria anymore. All classes are not considered *anti rigid*, because change is not mandatory. OntoClean demands that rigidity and non-rigidity cannot be inherited down the subclass hierarchy, a condition that obviously holds for the design of ArDO.

Identity: For `ardo:Record`, all instances of archival records provide a unique identification criteria. The same holds for `ardo:Portfolio`, `ardo:Chapter`, `ardo:Volume`, `ardo:Archive`, and `ardo:Dossier` which create new identity criteria, fulfilling the inheritance rule of OntoClean for *identity*.

Unity: `ardo:Record` can be considered a whole since we can devise fixed boundaries for each individual. The same holds for `ardo:Portfolio`, `ardo:Chapter`, `ardo:Volume`, `ardo:Archive`, and `ardo:Dossier`, which fulfills the inheritance rule for *unity*.

Dependence: The existence of a `ardo:Record` is not dependent on any external resource. However, instances of its subclasses `ardo:Portfolio`, `ardo:Chapter`, `ardo:Volume`, `ardo:Archive`, and `ardo:Dossier` are dependent of the existence of records. Again, OntoClean's constraint for *depencence* indicating that non-dependent classes cannot be subclasses of dependent subclasses is fulfilled.

6 CONCLUSION

This paper presents ArDO, an ontology developed for managing dynamics of multimedia archival resources. The proposed design allows to describe the semantics behind the target archive, keep track of annotations through a versioning mechanism which enables dynamics of a systematics, and to enrich the current archive through links to external linked data hubs. However, the choices made during the design process and the interaction with archivists have led to an ontology schema that can be adapted to other target scenarios with minor changes. In particular, on the basis of the target application, ArDO might need to be updated with proper metadata about the multimedia archival records. However, these changes do not affect the core infrastructure to model archival records and their relationships. Moreover, the systematics as it is modeled is not limited to the presented use case. Its design with categories, subcategories, and keywords can easily be extended to represent sophisticated deep hierarchies for the classification of not only arbitrary archival records, but also other use cases that require fluent classification schemes.

Focusing on the use case, the current ontology can be further improved and developed in several ways. An additional effort will include the linking of ArDO to a classification framework that is currently under development to automatically generate the annotations for archival records to support the manual annotation process. This will also include upgrades of the ontology schema since annotations may be originated both from humans and machines. Further investigation will also involve the evolution of the ontology design to represent more fine-grained data of the archival records (e.g., entities mentioned in a text or people depicted in an image). That will help to enrich the knowledge represented by the archive. In the long term ArDO signifies a substantial first step towards the provision of FAIR archival documents supporting all four FAIR principles: findability, accessibility, interoperability, and reuse [25].

ACKNOWLEDGMENTS

This work has been supported by a public grant overseen by the Deutsche Forschungsgemeinschaft (DFG) for the project “Aufbau einer Infrastruktur zur Implementierung sachthematischer Zugänge im Archivportal-D am Beispiel des Themenkomplexes Weimarer Republik (Projektphase 2)" with grant number 396707386 in the DFG LIS programme (Wissenschaftliche Literaturversorgungs- und Informationssysteme).

REFERENCES

[1] Terje Aaberge and Rajendra Akerkar. 2012. Ontology and Ontology Construction: Background and Practices. *IJCSA* 9, 2 (2012), 32–41.

[2] Renate Behrens-Neumann and Barbara Pfeifer. 2011. Die Gemeinsame Normdatei—ein Kooperationsprojekt. *Dialog mit Bibliotheken* 1 (2011), 37–40.

[3] Dan Brickley and Libby Miller. 2007. FOAF vocabulary specification 0.91.

[4] Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Nuzzolese, et al. 2019. ArCo: The Italian cultural heritage knowledge graph. In *International Semantic Web Conference*. Springer, 36–52.

[5] Marilena Daquino, Francesca Mambelli, Silvio Peroni, et al. 2017. Enhancing semantic expressivity in the cultural heritage domain: exposing the Zeri Photo Archive as Linked Open Data. *Journal on Computing and Cultural Heritage (JOCCH)* 10, 4 (2017), 1–21.

[6] Chris Dijkshoorn, Lora Aroyo, Jacco Van Ossenbruggen, and Guus Schreiber. 2018. Modeling cultural heritage data for online publication. *Applied Ontology* 13, 4 (2018), 255–271.

[7] Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, et al. 2010. The europeana data model (edm). In *World Library and Information Congress: 76th IFLA general conference and assembly*, Vol. 10. 15.

[8] Aurona Gerber, Nishal Morar, Thomas Meyer, and Connal Eardley. 2017. Ontology-based support for taxonomic functions. *Ecological Informatics* 41 (2017), 11–23.

[9] Karen F Gracy. 2018. Enriching and enhancing moving images with Linked Data. *Journal of Documentation* (2018).

[10] Fabio Grandi. 2016. Dynamic class hierarchy management for multi-version ontology-based personalization. *J. Comput. System Sci.* 82, 1 (2016), 69–90.

[11] Fabio Grandi, Federica Mandreoli, Riccardo Martoglia, et al. 2009. Ontology-based personalization of e-government services. In *Intelligent User Interfaces: Adaptation and Personalization Systems and Technologies*. IGI Global, 167–187.

[12] Nicola Guarino and Christopher Welty. 2002. Evaluating Ontological Decisions with OntoClean. *Commun. ACM* 45, 2 (Feb. 2002), 61–65. https://doi.org/10.1145/503124.503150

[13] Tobias Herrmann and Vera Zahnhausen. 2016. Auf dem Weg zum Digitalen Lesesaal: Das Projekt 'Weimar – Die erste deutsche Demokratie'. In *Kulturelles Kapital und ökonomisches Potential. Zukunftskonzepte für Archive. 86. Deutscher Archivtag*. Verband deutscher Archivarinnen und Archivare e.V.

[14] Fabian Hoppe, Tabea Tietz, Danilo Dessí, Nils Meyer, et al. 2020. The Challenges of German Archival Document Categorization on Insufficient Labeled Data. In *3rd Workshop on Humanities in the Semantic Web (WHiSe)*.

[15] Eero Hyvönen. 2009. Semantic portals for cultural heritage. In *Handbook on ontologies*. Springer, 757–778.

[16] Antoine Isaac and Bernhard Haslhofer. 2013. Europeana linked open data–data. europeana. eu. *Semantic Web* 4, 3 (2013), 291–297.

[17] Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. 2005. SKOS core: simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*. 3–10.

[18] Joachim Neubert. 2018. Linking Knowledge Organization Systems via Wikidata. In *Proceedings of the 2018 International Conference on Dublin Core and Metadata Applications* (Porto, Portugal) *(DCMI'18)*. Dublin Core Metadata Initiative, 3.

[19] J. Oomen, M.G.J. van Erp, and L. Baltussen. 2012. Sharing cultural heritage the linked open data way: why you should sign up. In *Museums and the Web 2012*.

[20] Laura Pandolfo, Luca Pulina, and Marek Zielinski. 2018. ARKIVO: an Ontology for Describing Archival Resources.. In *CILC*. 112–116.

[21] Silvio Peroni and David Shotton. 2012. FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Journal of Web Semantics* 17 (2012), 33–43.

[22] Joan M Schwartz and Terry Cook. 2002. Archives, records, and power: The making of modern memory. *Archival science* 2, 1-2 (2002), 1–19.

[23] Tabea Tietz, Jörg Waitelonis, Kanran Zhou, et al. 2019. Linked Stage Graph.. In *SEMANTICS Posters&Demos*.

[24] Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on world wide web*. 1063–1064.

[25] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.