Fachhochschule Graubünden
University of Applied Sciences

Churer Schriften
zur Informationswissenschaft

Herausgegeben von
Wolfgang Semar, Bernard Bekavac, Ivo Macek, Armando Schär

Arbeitsbereich
MAS Information Science

Schrift 159

From ISAD(G) to Records in Contexts – A new era

Josip Spec

Chur 2023

Churer Schriften zur Informationswissenschaft

Herausgegeben von Wolfgang Semar,
Bernard Bekavac, Ivo Macek, Armando Schär

Schrift 159

From ISAD(G) to Records in Contexts – A new era

Josip Spec

ORCID 0000-0001-8732-9850

Diese Publikation entstand im Rahmen einer Thesis zum MAS Information Science.

Referent: Dr. Tobias Wildi

Korreferent: Michael Aschwanden

Verlag: Fachhochschule Graubünden

ISSN: 1660-945X

Ort, Datum: Chur, März 2023

Abstract

The current international archival description standard ISAD(G) is after being for almost 30 years in use about to be replaced by the next generation standard Records in Contexts.

The new standard introduces a wide variety of changes and developments in comparison to the current one. These novelties relate equally to the content as well as the underlying technology, and as such Records in Contexts heralds a new era. Description and use of data for archives and other institutions alike is being aligned with concepts of Web 3.0 and Linked Data.

Currently, version 0.2 of the standard is published with the final version 1.0 expected to be published still in 2022. This thesis explores various dimensions of questions that might arise with the introduction of the new standard. The main goals are to explore and execute transformation of existing description to the new standard, presentation of possible use cases and benefits, and on that basis the creation of recommendations for institutions.

The necessary steps to achieve these goals are taken on basis of a sub-set of catalogue data of multiple GLAM institutions from a defined geographical area, the Swiss Canton of Zug.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIS	Archival Information System
BibZug	Bibliothek Zug [City and Cantonal Library of Zug]
CSV	Comma-separated values [File Format]
DB	Database
EAD	Encoded Archival Description
ETL	Extract-Transfer-Load
GLAM	Galleries, Libraries, Archives and Museums
HLS	Historisches Lexikon der Schweiz [Historical Dictionary of Switzerland]
ICA	International Council on Archives
IRI	International Resource Identifier
ISAD(G)	International Standard Archival Description (General)
ISIL	International Standard Identifier for Libraries and Related Organizations
MBZ	Museum Burg Zug
RDA	Resource Description and Access
RDB	Relational Data Base
RDF	Resource Description Framework
RDFS	RDF Schema [A data modelling vocabulary for RDF]
RiC	Records in Contexts
RiC-A	RiC – Attribute
RiC-CM	RiC – Conceptual Model
RiC-E	RiC – Entity
RiC-IAD	RiC – Introduction to Archival Description
RiC-O	RiC – Ontology
RiC-R	RiC – Relation
RiC-AG	Application Guidelines

R2RML	RDB to RDF Mapping Language
SPARQL	SPARQL Protocol and RDF Query Language [sic] (recursive acronym)
SQL	Structured Query Language
StAZG	Staatsarchiv Zug [City Archive of Zug]
SAZG	Stadtarchiv Zug [Cantonal Archive of Zug]
URI	Unique Resource Identifier
WD	Wikidata
XLS	Excel Spreadsheet [File Format]
XML	eXtensible Markup Language [sic]

1 Introduction

Preserving human and institutional memory has always been the primary purpose of archival records keeping[1]. The notion of what the best way would be to go about, naturally changed over time. Up until the eighteenth century the most widespread modus operandi was to sort archival records by topics and subjects. An approach that is even nowadays known to every child from the context of public libraries; history books are kept separately from fiction, and cookbooks have again their own shelves.

While this so-called approach by pertinence still works for libraries, it has long ago proven not to be suitable for most archives. In the nineteenth century the creation of nation-states and subsequently national archives coincides with modern historical methods and new approaches in archiving practice. It was then recognized that the re-arranging of records according to pre-defined topics and subjects ultimately destroys the inherent interrelations and interdependences of records. The way to solve this problem, was the shift from the Principle of Pertinence to the Principle of Provenance; to organize archival records by origin rather than by topic. The Principle of Provenance applies the understanding, that a set of record that was created by a specific actor, should be kept together as a coherent whole. In addition, that whole record set has an order within itself, and that original order should also be respected and not broken up. This approach has subsequently spread, and became and still is the foundation of modern archival practice[2]. While thus in the Principle of Pertinence the content and subject of a record is central, with the Principle of Provenance the creation context of a record set as meta-information is recognized and becomes vital to archival practice.

Nowadays, physical records most often aren't archived any longer by topic, but rather by provenance. In the archive of an exemplary administrative body, archival records of the civil registries' office are recorded separately from the police department ones, and the social welfare office has its own section as well. While this approach seems intuitive, it nonetheless has its flaws. Let's take as an example a set of records regarding a larger building project, like a central train station. Naturally, the records regarding that train station will not have only one provenance, but rather multiple ones, and will stem from creators like the construction department, the railway administration and the department of finance. Respecting the Principle of Provenance, all records contributed by the various creators will be recorded only hierarchically under that single administrative body in

[1] International Council on Archives (ICA) (2021): Records in Contexts – Introduction to Archival Description v0.2; 3

[2] ICA (2021): Records in Contexts – Introduction to Archival Description v0.2; 3 - 4

leading charge of the project. To record the complex reality with multiple provenances under which these records were created becomes impossible. The laws of the analogue world dictate, that a record can only exist at the same time in one place.

Based on these traditions and physical limitations, the nowadays globally used strict hierarchical "archival tectonics" by Principle of Provenance were developed. Now, one benefit of the digital world is, that various limitations of the analogue world can be transcended. A rigid archival hierarchy is certainly one of these limitations to be overcome! What has been, is being, and could be digitally done to further develop archival description shall be explored on the next pages.

1.1 From ISAD(G) to Records in Contexts

The main most straight-forward purpose of archival description is to provide access to records. To produce uniform and consistent result, it certainly helps to have standards. Such standards can describe the data structure (from a technical point of view), the data content, or data values. A data content standard is what one would call a description or cataloguing standard. It helps if a standard is used within a single institution consistently, but if a broad range of institutions are implementing the same standard, only then it's larger potential is used. Until the late 20th century, various standards have been developed, which unfortunately, remained limited to local institutions, and didn't transcend national borders[3].

A working group of the International Council on Archives (ICA) set out in 1990[4] to develop a standard for archival description, and published the 1st version in 1994 as **General International Standard Archival Description [ISAD(G)]**. A 2nd amended version was published in 2000. As an "umbrella" standard for archival description, ISAD(G) was intended to facilitate international exchange. Where no national standards were in place, ISAD(G) was meant to serve as a baseline. Existing national standards nonetheless weren't meant to be replaced, but rather supplemented[5]. Since the publication of the first version of ISAD(G), it has been gradually adopted and implemented in various national[6] and

[3] Weber (1989): Archival Description Standards: Concepts, Principles, and Methodologies; p. 505 - 510

[4] International Council on Archives, Brien Borthman (1992): ISAD(G): General International Standard Archival Description

[5] Brüning et al. (2006): Internationale Grundsätze für die archivische Verzeichnung. p.2

[6] Verein Schweizerische Archivarinnen und Archivare (2009): Schweizerische Richtlinie für die Umsetzung von ISAD(G)

regional[7] contexts, thus being nowadays the global de-facto standard for archival description.

ISAD(G) has different characteristics, out of which only a few shall be mentioned here. The standard is based on a multi-level description within a hierarchical framework. It provides general rules for archival description, that may be applied to any form or medium of archival material. A total of 26 elements within 7 categories are offered, out of which only few are mandatory, but all are available for use to constitute the description of an archival entity[8]. Even though the standard was developed in the early stages of the "digital age", the characteristics described above hint to limitations within.

ISAD(G) is based on, and restricted by principles of the material logic world, its underlying predecessors, and analogous finding aids. The above mentioned strict hierarchical framework maps simple subordination and superordination relationships[9], as they occur in an analogue archive. A document belongs to a single dossier, that dossier belongs to a unique series, which itself belong to a single fond. ISAD(G) only depicts that reality of physical storage of an archive, but lacks the functionalities to describe the much more detailed world and its contexts a record stems from. While ISAD(G) certainly helped to establish archival description in the digital world, thus it unfortunately doesn't make full use of the potential of computers, databases, networks and the digital world in general.

Various problems and shortcomings as the one described above have been identified by the ICA and led the institution to work on the development of a next generation standard[10]: **Records in Contexts (RiC)**. RiC has been under development and review since 2016, while the latest round of comments just recently closed on January 31, 2022[11]. (The release of the final version 1.0 was planned for late 2021, but its launch has been postponed to this year). RiC is the overarching standard, which includes for now 3 parts:

- **Records in Contexts - Conceptual Model (RiC-CM)**: *"[…] a high-level model that focuses on intellectually identifying and describing records, the agents that created, used, or are documented in them, and the activities pursued by the agents that the records both facilitate and document."*[12]

[7] Tropper, C. und Zeloth, T. (2010): Kärntner Landesarchiv - Richtlinien für die Ordnung und Verzeichnung von Archivgut

[8] International Council on Archives (2000): ISAD(G): General International Standard Archival Description; p. 7-9

[9] Messner, Philippe (2017): Records in Contexts – vom Baum zum Netz

[10] Merzaghi & Ohnesorge (2020): Neue Wege in der archivischen Erschliessung – Scopriamoli assieme

[11] International Council on Archives (2021): Records in Contexts – Conceptual Model

[12] ICA (2021): RiC-CM, v0.2, p. 1

- **Records in Contexts - Introduction to Archival Description (RiC-IAD)**[13]: A general and broad introduction to the principles and purpose of archival description
- **Records in Contexts - Ontology (RiC-O)**[14]: A definition of the formal representation of the standard. One could call it the dictionary or vocabulary of the standard, that makes it machine-processable.

The main evolution from ISAD(G) towards RiC is an evolution in the underlying logic and depth of description. In ISAD(G), the 26 available elements can each be an attribute of a single unit of description (entity). The ISAD(G) logic *unit of description-elements* follows the same *entity-attribute* logic of relational databases (RDB), which is why understandably current ISAD(G) archival information systems (AIS) run on RDBs. All archival description is described by two data pieces as *entity-attribute.* A simple illustrative example for the description of an entity/record with 3 attributes[15] is given below in Table 1:

Entity	*Swiss Federal Constitution of 1848*		
Attribute	Reference code	Title	Creation period
Attribute value	*E22#2.1.1*	*Bundesverfassung von 1848*	*1847 - 1858*

Table 1: Example record to illustrate entity-attribute relationship

How does that relate to what RiC has to offer? While ISAD(G) offers 1 entity/record (unit of description) with a maximum of 26 elements/attributes each, RiC is much broader.

RiC offers 22 distinct entities (RiC-E), that are the main objects of interest[16]. Such entities are among others “Record Resource”, “Agent”, “Event”, or “Place”. The standard offers then in addition 47 attributes (RiC-A), as characteristics of these entities. Not all attributes are applicable to all of the entities though. Such attributes range from the known ones like, “Conditions of Access” or “Scope and Content” to new ones such as “Production Technique” or “Legal Status”. Besides being quite broad itself, RiC is also extensible[17], meaning if it is missing an entity or attribute to fulfil a certain description need, any other ontologies like Resource Description Framework Schema (RDFS) or Dublin Core Terms[18] can be integrated as well.

[13] ICA (2021): RiC-IAD

[14] ICA (2021): RiC-O

[15] Bundesarchiv (2022): Bundesverfassung von 1848

[16] ICA (2021): RiC-CM, v0.2, p. 17

[17] ICA (2021): RiC-CM, v0.2, p. 2

[18] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

But the most crucial change is not only that there are more entities and more attributes available. At the core of the transition to the new standard, is that any ISAD(G) statement with two data pieces (*entity-attribute)* is expanded to three linked data pieces for any statement according to RiC. The three linked data pieces are normally labelled either *subject-predicate-object,* or *entity-relationship-entity.* These expressions of information with 3 elements are called “triples” (or semantic triples) and are the underlying statements of Resource Description Framework (more in the next chapter). What RiC offers among other things, are standardized elements of these semantic triples. These triples in comparison to classic RDBs and its tables, add a multitude of additional dimensions to archival description and open up various possibilities for advanced description, search queries and linked (open) data within and between institutions. And where RiC defines entities (RiC-E) and their attributes (RiC-A) analogous to ISAD(G), the standard also defines predicates, or relations (RiC-R), which themselves also can have attributes (RiC-RA).

The new standard will therefore allow, as an extension to existing single subordination and superordination relationships, the possibility of (1) horizontal and (2) multiple/plural relationships between entities[19]. This means that no longer a multitude of tables will be connected as in RDB, but information can be used to form networks and graphs. Practically speaking, RiC allows to overcome limitations of current archival description based on physical limitations as they were described above, and allows to make full use of the potential of digital databases.

How an implementation of the new standard might look and what benefits it brings to archivists and customers alike shall be explored in the next chapters. Given the possibilities that RiC presents, it would certainly be beneficial to the whole archival community, for it to be adapted as widely as ISAD(G).

Since RiC is by design a deeper and more ambitious model than ISAD(G), implementation of the standard will naturally take its time to “arrive” in the archival world. A whole generation of archivists was educated on the intellectual basis RDBs function with (maybe even without understanding the underlying technical logic). Thus, the RiC-CM even acknowledges in the model itself[20], that *“mastering the intellectual and technological complexity of RiC-CM by archivists, records managers, and the developers of systems that support their work will take time”.* In order to facilitate later discussion and exploration of mapping to the new standard and working with it, a technical introduction to the standard

[19] Messner, Philippe (2017): Records in Contexts – vom Baum zum Netz

[20] ICA (2021): RiC-CM, v0.2, p. 13

and its underlying technology (at least on a very general level) is necessary. The understanding of the underlying technology of archivists generally speaking, and not only computer scientists or developers, will most certainly also be crucial for the successful transition of institutions to the new standard.

The relevance of the topic is evident. RiC is published by the ICA, the world-wide umbrella organisation of all archives. The goal is to replace the most widespread standard ISAD(G), and to lift archival description into a new era. While it is not feasible that this goal will be achieved immediately, it is feasible to see RiC as the de-facto standard in the mid-term.

This thesis will focus on the challenges in moving from ISAD(G) to RiC, in moving from RDB to RDF, and on various benefits the new standard offers to the end-customer. The aim, briefly summarizing, is to map existing archival description to the new standard, and to explore various possibilities that the new standard offers. The underlying research questions and methods are described in more detail in the respective following chapters. Before that, for a better understanding a brief excursion into technical implications and linked data, and its connection to Records in Context is necessary as well.

1.2 From RDB to RDF and Linked Data

As laid out above, moving from ISAD(G) to RiC is not only an improvement in description standard fields, but involves also a change in underlying technology. Data conforming to ISAD(G) is normally stored in RDBs. In ISAD(G), data is described in “tuples”, an *entity-attribute* statement. The Resource Description Framework (RDF) doesn't use "tuples" any more, but “triples”. Every triple makes a statement about resources, and the 3 elements of that statement are always the same[21]: *subject – predicate – object.* As such, RDF is the foundation of the semantic web[22].

The databases (DB) to store such triples are called graph DBs. This means a move from ISAD(G) with its tree structure and from relational DBs and their underlying table structures, to graph DBs with underlying network structures.

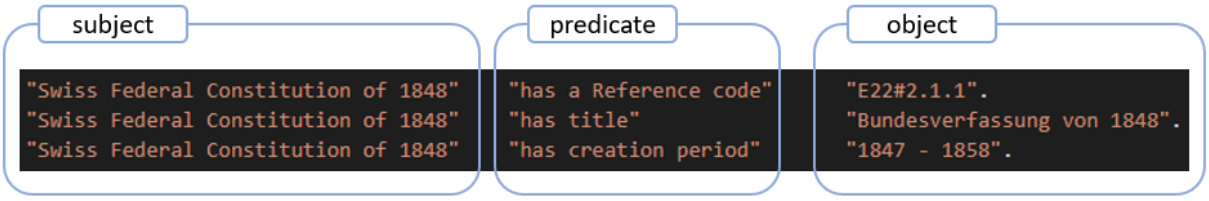
RDF is only a data model, that defines and standardizes elements of triples. How these triples are stored technically, is described by the term serialization. There are multiple ways and formats of serialization of RDF, such as RDF/XML, JSON-LD, RDFa and

[21] World Wide Web Consortium (2014): RDF 1.1 Primer

[22] Gutierrez et al. (2011): Foundations of Semantic Web Databases, in: Journal of Computer and System Sciences 77(3), p. 520

others[23], each with advantages and disadvantages[24]. Because it is a format that is easy to work with, and because it is also quite easily human-readable and editable, I will stick throughout this paper with the Turtle (.ttl) format.

The difference between RDB and RDF, and the use of Linked Data, is easily explained through a simple and illustrative example. Let's take the Swiss Federal Constitution of 1848, recorded as RDB dataset as shown in the chapter above. While it was stored originally in table form, transforming every *entity-attribute* statement into an RDF-*triple* could look like this, where every row is a triple statement:



Expressions with quotation marks (" ") like above, describe literal values. So far, the original RDB-entry has only been translated to RDF, but there isn't yet any additional value to it.

In order to be able to link data, one would want to use (you guessed it) links, instead of literal values. Links remove ambiguity, provide extra information, add standardization, and are normally designated with angled brackets (<>)[25]. In this example-triple, the subject could be instead of the literal "Swiss Federal Constitution of 1848" a unique identifier of that record[26] (the corresponding entry in Wikidata [https://www.wikidata.org/wiki/Q61946269] was chosen for a unique identifier). The predicates then could use any standardized ontology (like RiC-O) to provide a link (e.g. *<rico:hasOrHadName>*) instead of a literal description. When using a prefix to create links, like *rico:* in this example, the element is treated as link, even if it isn't designated with angled brackets in the triple.

```
@prefix rico: <https://www.ica.org/standards/RiC/ontology#> .

<https://www.wikidata.org/wiki/Q61946269> rico:hasOrHadIdentifier   "E22#2.1.1".
<https://www.wikidata.org/wiki/Q61946269> rico:hasOrHadName         "Bundesverfassung von 1848"@de.
<https://www.wikidata.org/wiki/Q61946269> rico:hasOrHadName         "Federal Constitution of 1848"@en.
<https://www.wikidata.org/wiki/Q61946269> rico:creationDate         "1847 - 1858".
```

[23] World Wide Web Consortium (2014): RDF 1.1 Primer

[24] Meindertma, Joseph (2019): What's the best RDF serialization format?

[25] Meindertma, Joseph (2018): A brief introduction to linked data

[26] https://www.wikidata.org/wiki/Q61946269

Also, a subject could have the same predicate multiple times, as shown with *rico:hasOrHadName* for the German and the English name of the described record. With the @de and @en after the objects, the language of the literal is specified. After the described amendments, the subject in every statement is now a clearly identified unique entity, and the predicates are also links defined by and conforming to RiC-O. The subjects of any statement must be links; the predicate and the object of a statement should, but don't necessarily have to.

In the given case, to replace the existing objects-literals with links wouldn't necessarily make sense. But one might need other objects to describe the record, e.g. that the content type of the record is <text>[27], and that the language in the record is <German>[28]. RiC-O offers predicates to make these statements, and as subject here a link instead of a literal makes sense, and the new statement might look like this:

```
@prefix rico: <https://www.ica.org/standards/RiC/ontology#> .

<https://www.wikidata.org/wiki/Q61946269>   rico:hasOrHadIdentifier     "E22#2.1.1";
                                            rico:hasOrHadName           "Bundesverfassung von 1848"@de;
                                            rico:hasOrHadName           "Federal Constitution of 1848"@en;
                                            rico:creationDate           "1847 - 1858";
                                            rico:contentType            <https://www.wikidata.org/wiki/Q234460>;
                                            rico:hasOrHadLanguage       <https://www.wikidata.org/wiki/Q188>.
```

Now the additional objects in the triple are also links, just as the subject of these statements. These objects thus can be also objects of other triple-statements, or become subjects themselves for any other statement, and thus are the first step to expanding the information network and linking other Resources with the same entities.

This is in a nutshell what RDF and Linked Data is about, and how RiC is part of it. Even with such a short introduction, there are a lot new words and concepts, which besides computer scientist few people are very familiar with, and can be confusing at first. But in order to work adequately with RiC, and ultimately with RDF and linked data, it will require archival experts to master these new concepts and terminologies.

Some clarification of terminology regarding linked data is necessary at this point. So far, I've mentioned links, and a type of links one might be most familiar with is a URL. Besides URLs, other types like IRI, URI and URN exist. The differences are of technical nature, and don't need to be explored here in detail. In summary, one can say that, IRI is a superset of URI, URI is a superset of URL and URN, and URL and URN are disjoint[29]. Thus,

[27] https://www.wikidata.org/wiki/Q234460

[28] https://www.wikidata.org/wiki/Q188

[29] Keil, Jan Martin (2016): IRI, URI, URL, URN and their differences

every type of link is ultimately a subset of IRI, RiC-O uses IRIs, and therefore I will throughout this thesis also consistently use IRIs.

1.3 State of research and practice

RDF, the technical foundation of RiC, has first been adopted in 1999[30]. The latest version of the standard was published in 2014[31]. As such, the standard has various use cases, such as knowledge management in enterprises[32], in criminal intelligence, urban transport assistance, and many others[33]. Naturally, also various DB management systems exist, such as Neo4j, GraphDB or Microsoft Azure Cosmos DB[34]. But since RiC is currently (July 2022) only published in version 0.2, and thus a very new description standard, there is currently no software based on RDF and conforming to RiC available, that is as an AIS explicitly tailored to archival requirements. Existing description and data can nonetheless be transformed to RDF and loaded into a graph DB.

So far, the first steps towards RiC in the archival community were taken. First proof of concepts and prototypes using RiC have been built so far[35]. Two examples, that have incorporated components of RiC, are "Memobase"[36] and the "Social Networks and Archival Context Cooperative"[37] (SNAC). Memobase is a portal for Swiss audio-visual archives. The newest version of Memobase (released 2021) aggregates metadata from various institutions, and uses RiC as a basis for its own metadata model. SNAC on the other hand is in the planning phase of incorporating certain elements of RiC-O. An example of a mapping from ISAD(G) to RiC is the State Archive of Basel-Stadt[38], which maps and publishes its catalogue to an RDF-model based on RiC-O. The institution nonetheless still is describing its holdings with an ISAD(G) conforming tool, and the RDF mapping is just regularly updated downstream copy of the "original" DB. The National Archives of the UK are currently working on "Project Omega". "Project Omega" has as goal to determine a sustainable data model for a so-called pan-archival catalogue, that integrates more than

[30] World Wide Web Consortium (1999): Resource Description Framework (RDF) Schema Specification

[31] World Wide Web Consortium (2014): RDF 1.1 Primer

[32] Kasenchack, B. et al. (2021): Use Case: Ontologies and RDF-star for Knowledge Managemen

[33] World Wide Web Consortium (2015): Use cases

[34] DB-Engine: (2022): DB-Engines Ranking of Graph DBMS

[35] International Council on Archives (2021): RiC-O projects and tools.

[36] https://memobase.ch

[37] https://snaccooperative.org/

[38] https://data.bs.ch/explore/dataset/100177/information/

10 DBs currently being used by the National Archives. The project determined in its proof of concept, that RiC-CM is the adequate Model and that a graph database has to be used for that pan-catalogue. The National Archives are on basis of these findings developing a pilot platform, an API and an end user interface as a demonstration, with the possibility to perform selected archival core tasks and workflows. The ultimate goal is then to replace the two-decades old catalogue management system[39]. All of the above-described projects are steps in the right directions, but RiC hasn't yet been fully implemented anywhere, and with few implementations also little research and documentation is existent.

Currently, software for archival description according to RiC is for example being developed by the companies "Artefactual Systems", "docuteam" and "Zazuko". Appealing and well-functioning software will be at the core of a successful and sustainable roll-out of RiC, and I am curious about and looking forward to future developments.

[39] The National Archives (2022): Project Omega

2 Research Question

At the time of writing of this thesis (mid-2022), the RiC-CM 0.2 is the most recent version of the standard, having been published in July 2021. The Expert Group on Archival Description of the ICA called for comments on that version until end of January 2022, and after received feedback the release of a consolidated version 1.0 as a normative ICA recommendation is still planned in 2022[40]. Since it is a fairly new standard, and the release of the final version still eagerly anticipated by the archival community, only a few explorative use-cases have been carried out as outlined above.

With the whole standard being still under development and in early stages, this thesis has the aim to conduct a dry run of implementing the standard, and identifying various possible pitfalls, possible solutions to these challenges and implementation opportunities. Thus, the research questions are aligned with the process of an exemplary mapping and transformation of existing description to the new standard, and the use of new opportunities RDF and linked data offers. Therefore, the guiding research questions in these 2 categories to be explored within this thesis are:

Mapping and transformation

1a How can existing ISAD(G) archival description to be transformed to RiC?

1b How can non-ISAD(G) description (e.g. RDA description) be transformed to RiC?

2 How can ISAD(G) and non-ISAD(G) description be combined within RiC?

New opportunities

3 What are potential RiC-expansions of mapped ISAD(G) datasets?

4 What are potential use cases and benefits for internal users and customers?

5 Which recommendations regarding the implementation of ISAD(G) can be given to institutions?

In order to give possible answers to these questions, a sample of datasets from multiple institutions within a defined geographic space will serve as a test object, as described in the following chapter.

[40] Expert Group on Archival Description (2021): Call for comments on ICA RiC-CM 0.2

3 Methodology

The following methodology describes (1) the general practical procedure, as well as fundamentals such as (2) the data model to be used, and (3) sources for IRIs and ways of creating them where needed.

3.1 Procedure

The main idea is, simply put, (1) to map existing datasets to RiC, and (2) to improve the datasets. The whole process of mapping to RiC and enriching the datasets is carried out on the basis of a few selected holdings of multiple institutions. One could reasonably assume, that geographically close institutions might hold similar items relating to local topics. Based on that assumption, the geographical area of the Canton of Zug in Switzerland was chosen, for its manageable size. Within this municipality, the following institutions were identified as holders of relevant records and items:

Institution	Abbreviation	ISIL	Description
Bibliothek Zug	BibZug	CH-000063-8	Library of City and Canton of Zug
Staatsarchiv Zug	StAZG	CH-000073-4	Archive of the Canton of Zug
Museum Burg Zug	MBZ	CH-000153-7	Zug Castle - Historic Museum of Zug
Stadtarchiv Zug	SAZG	CH-001402-9	Archive of the City of Zug

Table 2: Overview of institutions to be examined

Each of these institutions will be separately contacted regarding their respective DBs and exports. After receiving the exports from the current catalogues/databases, the approach can be split in 2 main tasks, divided according to the same logic as the research questions are grouped.

Firstly, existing archival description and other records have to be transferred to RiC. And secondly, after the mapping, the datasets have to be expanded by means made possible by RiC. Both main steps come with various sub-steps which are described briefly below. A more detailed description of the approach follows in the respective sub-chapters under 4 and 5.

Step 1: Mapping to RiC (chapter 4.)

- *4.1 Identifying suitable topics*
  - For this purpose, only a set of the (archival) DB, and not the whole DB with its entirety of records shall be mapped to Ric. Various holdings will be hand-picked by pertinence, based on potential representation of different use-cases of clients
- 4.2 *Identifying records and exporting the relevant datasets*
  - The relevant records and datasets relating to the respective topics will be identified through manual research in the respective DBs, and exported in a suitable format
- *4.3 Transforming and mapping*
  - The exports from the relevant DB will undergo a manual screening and will be cleaned and/or split where necessary, to be better suited for mapping to RiC
  - The cleaned datasets will be mapped to RiC-triplestores in Turtle-format. For this step the tool OntoRefine implemented within GraphDB will be used.
- *4.4 Using the graph DB*
  - A local installation of the DB will be installed, and the necessary preparation of a repository with adequate settings will be prepared
  - The mapped data will be imported step-by-step into the DB, and settings of the repository will be adjusted where necessary

Step 2: RiC use cases (chapter 5.)

- *Enriching data*
  - Various data enriching and/or additional archival description will be undertaken along the RiC-entities in order to facilitate inter-connection between the fonds of different institutions, and expand information networks outside of the existing description
- *Knowledge graphs*
  - Various knowledge graphs based on potential use-cases of clients will be extracted through adequate DB- queries, and challenges and possibilities will be explored.

After the mapping, transformation, and practical discovery, the lessons learned along the process will be translated and formulated into recommendations for institutions that might implement RiC in one way or the other (chapter 6.).

3.2 Data model – RiC «light»

RiC offers a very broad conceptual model and ontology, that is intended to be able to capture a wide variety of contexts occurring in reality. Existing description based on I-SAD(G) (or most bibliographic records) don't hold information in that depth. A mere mapping of an existing archival DB won't make full use of RiC. In addition, the goal of this thesis is not to fully map existing description and use as much as possible elements of RiC, but rather to explore certain use cases in-depth. Thus, only a set of elements offered by RiC will be used in the mapping; one might call it RiC "light". Nonetheless, it makes sense to think about which RiC entities will be newly created, which existing attributes shall become which elements of RiC, how the entities link among each other, and how RiC attributes describe them. A helpful tool for visualizing such planning is a schema of a data model, as shown below.

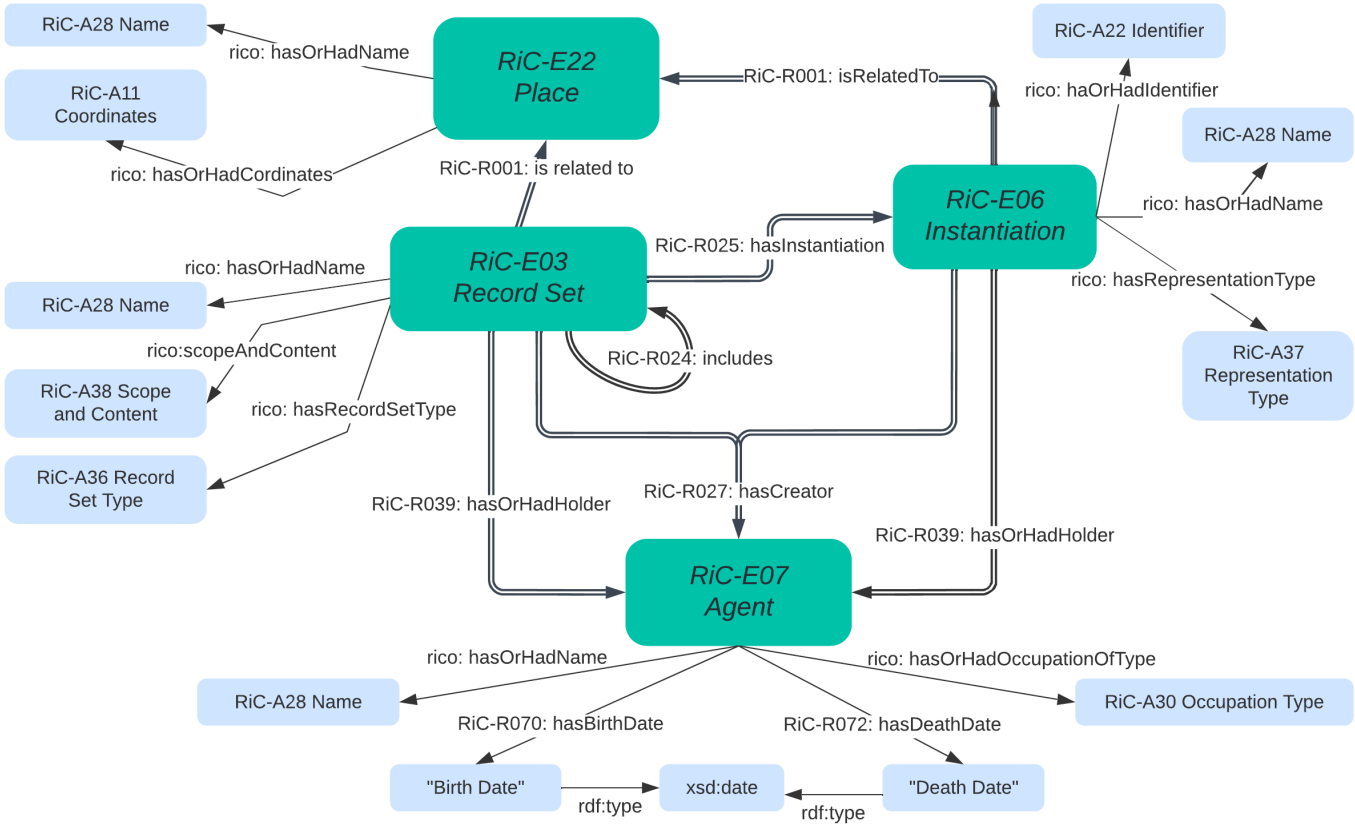


Figure 1: A simplified schema of the used data mode

All archival records/collection objects/bibliographic records will be treated either as RiC-E03 “Record Set” or as (physical or digital) RiC-E06 “Instantiation”. RiC-E03 alone is sufficient to be able to describe the existing archival hierarchy, and RiC-E04 Record and RiC-E05 Record Part will in this case not be introduced for simplicity reasons. In addition, the later defined research topics will focus on certain RiC-E07 Agents (RiC-E08 Persons and RiC-E11 Corporate Bodies), RiC-E22 Places, and RiC-E14 Events, which is why these entities are at the core of this data model.

Furthermore, dates and times tend to be based on different formats, standards, tradition and description depth messy. RiC offers possible ways of handling these issues and offers an own entity RiC-E18 Date, but cleaning/normalizing of date formats is extensive and exhaustive. For this reason, existing dates will simply be mapped as literal values, and assigned the XML Schema datatype "*xsd:date*".

Some possible attributes of the entities that are to be expected are included in the data model, like RiC-A28 Scope and Content for RiC-E03 Record Set, or RiC-A22 Identifier for RiC-E06 Instantiation. Nonetheless, the data model depicted is not exhaustive. During the mapping additional information pieces might occur, which are more appropriately represented by other entities or attributes not depicted in the data-model above. In cases where such pieces of information occur, additional attributes or entities will be created during the mapping.

And ultimately, the entities are linked among each other in the data model with double arrows. This information is only recorded implicitly (if at all) in the existing DBs, and will need to be represented explicitly during the mapping to RiC/RDF.

This data-model will serve as the target-model, into which later on the exported data-sets shall be mapped into. As a disclaimer I would like to state, that I try to make sure that the mapping is correct and senseful, but I do not raise the claim, that it is complete or 100% correct. This is an explorative thesis and so far, only v0.2 of RiC-O, RiC-CM and RiC-IAD have been published. After publication of the final version 1.0 of RiC, which should also include in addition RiC "Application Guidelines" (RiC-AG), a revisited and probably corrected mapping would make definitely sense.

3.3 Ontologies and IRIs / disambiguation

This chapter serves the purpose to define, which ontologies will be used in the course of this thesis, and which IRIs will be used for entities, or how IRIs will be created where not existent.

Since RiC is expandable, any RDF ontology could be used in addition to RiC. Since the RiC-O already offers a quite broad set of elements, I will for the sake of focus and simplicity refrain from integrated and too wide variety of ontologies. In the following, the used ontologies and prefixes are described.

The focus is RiC-O, and RiC offers besides that general ontology also two sub-ontologies RiC-RecordSetType (`ric-rst`) and RiC-DocumentaryFormType (`ric-dft`), that will be used were adequate. In addition, there exist various ontologies/namespaces (OWL; RDFS; RDF; SKOS; XSD), which RiC itself uses by design. These are the basic RDF vocabulary for e.g. defining classes of entities, its extension RDF schema, the XSD schema mainly used in this context for defining datatypes, as well as the OWL Web Ontology Language. Besides that, an additional prefix will be used for the integration of Wikidata-IRIs (wd:):

Prefix Name	Prefix IRI
owl:	http://www.w3.org/2002/07/owl#
rico:	https://www.ica.org/standards/RiC/ontology#
ric-rst:	https://www.ica.org/standards/RiC/vocabularies/recordSetTypes#
ric-dft:	https://www.ica.org/standards/RiC/vocabularies/documentaryFormTypes#
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs:	http://www.w3.org/2000/01/rdf-schema#
wd:	https://www.wikidata.org/wiki/
xsd:	http://www.w3.org/2001/XMLSchema#

Since I will work with exports from DBs that only contain literal values, I do not expect any IRIs in the exported data. Therefore, I will have to assign IRIs for entities to be created. For some I will be able to use existing IRIs, as e.g. entries for Persons or Institutions in Wikidata. For others that don't have a corresponding IRI somewhere on the Internet, I will have to manually assign IRIs, e.g. for DB-entries pertaining to holdings within an individual institution.

Depending on the type of entities, various sources for IRIs in various hierarchical order will be used. Exports of entries in archival or library software normally contain a specific identifier for the entry. Such an identifier (e.g. "413398") will be used to create with a prefix RecordSet and Instantiation entries, as shown in the example below:

Entity-type	Source	Example IRI
RiC-E03 RecordSet	Locally created IRI	Source Software ID: "413398" Target GraphDB IRI: <E03_413398>
RiC-E06 Instantiation	Locally created IRI	Source Software ID: "3314" Target GraphDB IRI: <E06_001_3314>

Such a procedure wouldn't be acceptable in a real-world scenario. The exported identifier stem from multiple different DBs, and there is a possibility that with such a procedure non-unique IRIs would be created, and another way of creating unique IRIs would be better suited. Since the sample set is quite small, I will nonetheless proceed in this manner, and after merging all exports check for duplicates IRI, and correct them if needed.

Other IRIs will be needed for Agent-entities (RiC-E07). As a first priority, I will use Wikidata entries for Persons and Corporate Bodies where possible, because it is for my purpose the vastest source of possible IRIs, and also could be manually expanded. If no Wikidata entry exists, as a second priority I will use Links to the respective entry in the HLS. As a third priority, Foto-CH as a database for Swiss photographers and Swiss photographic institutions will act as a source for Swiss photographers. And if an Agent-entity is not represented in any of these Sources, I will create a local IRI with "http://example.org/base/" as the basis and the name of the entity as a suffix, as shown in the example below:

Entity-type	Source	Example IRI
RiC-E07 Agent	Wikidata	wd:Q98931764
	HLS	https://hls-dhs-dss.ch/de/articles/042168/2014-11-26/
	Foto-CH	https://www.foto-ch.ch/?a=fotograph&id=29555
	Example.org/base	http://example.org/base/HenryHabegger

A further entity types that needs IRI creation, is Place (RiC-E22). I will here again as a first priority use, if possible, a Wikidata entry. Where not applicable, I will refer to the Linked Data Portal of the Swiss Government https://geo.ld.admin.ch/; which offers the possibility of using IRIs for certain postal addresses. Strictly speaking, the IRIs offered describe the (postal) address that a building has, and not the physical building itself. The physical address and the building referred to itself don't necessarily have a 100% overlap

and also aren't persistent. A specific building might have multiple main entries and therefore multiple addresses, or the address of a building might also change over time. Thus, strictly speaking, using that IRI of the address to identify the building logically wouldn't be correct. A much better IRI would e.g. be a unique identifier of a building of the respective administrative body that building is located in. For illustrative purposes and the lack of better options, I will nonetheless go with the option described above. If neither source offers an IRI, I will here again use a fictive IRI as basis.

Entity-type	**Source**	**Example IRI**
RiC-E22 Place	Wikidata	wd:Q68144
	Linked Data Service of the Swiss Government	https://geo.ld.admin.ch/location/address/101239133
	Example.org/base	http://example.org/base/Kolinplatz

And ultimately, if for any other type of entity besides the 4 entities defined in the data model, an IRI outside of the described ontologies and sources for IRIs would become necessary, I will in that case again use "http://example.org/base/" as a basis to create a local IRI.

As addition, the exports are expected to render various type of identifiers, analogue reference codes as well as software identifiers for example. Where triple-statements about a RecordSet or Instantiation are made to link the object with the predicate `rico:hasOrHadIdentifier` to the respective literal identifier, two types of identifiers (`localSoftwareID` / `localHolderRefCode`) are defined to distinguish the identifiers.

4 Step 1: Mapping to RiC

The existing records and objects in the institutions are recorded either in relational databases or in tables. There are 2 main ways to transform data from RDBs to RDF, either through direct mapping or through the R2RML (RDB to RDF Mapping Language)[41]. While R2RML is better suitable for the mapping of more sophisticated DBs, in our case direct mapping is sufficient (more in chapter 4.4). Since the goal is to work with records from different institutions, i.e. to combine datasets from multiple DBs, the process to be undertaken shall follow the ETL concept for merging data from different sources: extract-transform-load[42]. Since the goal is not to map the full DBs to RiC, there is one additional previous step, the identification of use-case topics and research topics (4.1). After identifying which entries are of interest, the entries shall be extracted from the respective DB (4.2). Since the data is heterogenous it shall be transformed (4.2) in two ways: cleaning the data and homogenizing it in order to comply with RiC-O, and transforming it from RDB-statements to triples for graph DB. And ultimately after the transformation, the improved data shall be loaded and combined into a graph DB (4.3). The detailed description of each step follows in the subsequent sub-chapters. Sub-chapter 4.3 especially is written with a focus on the initial research questions "*1a. How can existing ISAD(G) archival description to be transformed to RiC?"* and *"1b. How can non-ISAD(G) description (e.g. RDA description) be transformed to RiC?".* The chapter then brings together data from different sources into one single Repository in 4.4.2., offering a possible solution to research question "*2. How can ISAD(G) and non-ISAD(G) description be combined within RiC?".*

4.1 Use-case research topics

As described previously, Records in Contexts offers a wide variety of entities for the description of contexts. These entities, such as “`person`”, “`corporate body`” or “`event`” are the logical core of RiC, and can also be at the core of the research interest of potential users. An average is thus oftentimes interested in a specific topic or a person, and would approach a heritage institution with the according “topic”. That person might research and approach an institution regarding a specific person, a building or some kind of other entity, and be less interested in diving into the depths of archival hierarchy according to provenance.

[41] Malhotra, A. (2012): Mapping Relational Data to RDF

[42] Ashraf, S. (2020): ETL is Not Dead. It is Still Crucial for Business Success

Therefore, when approaching the covered institutions, such possible topics of interest for users were chosen as starting point, that might occur in multiple institutions, as well as in different fonds and sub-fonds. The underlying assumption is, that with the possibilities of RiC and the linking of these objects, the needs of users might be better serviced. Thus, the potential research topics were pre-defined as use-cases for the Canton of Zug, that will serve as examples are:

Category	Research topic	RiC class / entity
Person	“Weiss, Katharina” (photographer, 1834 - 1911)	Person (E08)
Person	“Keiser, Dagobert” junior (architect, 1879 - 1959)	Person (E08)
Person	“Keiser, Dagobert” senior (architect, 1847 – 1906)	Person (E08)
Person	“Etter, Philipp” (federal chancellor, 1891 – 1977)	Person (E08)
Building	“Schulhaus Neustadt”	Place (E22)
Building	St. Oswalds-Gasse 21, 6300 Zug	Place (E22)
Event	Zuger “Seegfrörni” 1895 / 1929 / 1963	Event (E14)
Company	“Metallwarenfabrik Zug”	Corporate Body (E11)
Company	“Crypto AG”	Corporate Body (E11)
Institution	“Theater- und Musikgesellschaft Zug” (tmgz)	Corporate Body (E11)

Table 3: Overview of potential research topics of clients, to be examined In this master’s thesis

These research topics have been used to approach the different institutions and to request exports of relevant entries of their respective DBs. The exports are described in the following chapter.

4.2 Extracting data

The second step after defining the use-case topics above, is to identify relevant holdings in the various institutions, and to export them from their respective DBs for later mapping. The different institutions use different standards and catalogues, and thus a pragmatic approach to exporting was taken. The lowest common denominator is a row/record in a DB, pertaining to a single unit of description. Speaking in ISAD(G) terminology, such a described unit could be a fonds/sub-fonds, a series/sub-series, a file or an item[43]. Since a Museum and a Library are included as well, the units of description will also go beyond

[43] International Council on Archives (2000): p. 16

what ISAD(G) offers. The goal was to create from the different software and DB-exports as similar as possible, to facilitate later joining and cleaning of the data. Comma-separated values (CSV) file-format is the optimal choice for that purpose, because it is widely used, easy to work with and manageable by almost any software.

The relevant holdings in the institutions were identified by manual research. The goal wasn't to export the full DB or all holdings, but only a representative selection. The reason is that the goal of the thesis is to explore various use-cases, and not to map full DBs to RiC. For such a purpose, a smaller sample with hand-picked holdings might be more beneficial than a huge DB-extract. For all institutions various relevant DB-entries with a varying depth of description and attributes were identified, as shown in the table below:

Institution / Holding	Attributes / Columns	Records / Rows	Software
Stadtarchiv Zug	10	64	CMI Archiv
Museum Burg Zug	10	31	Plus Ria
Bibliothek Zug – library catalogue	9	109	Quria
Bibliothek Zug - archival catalogue	12	178	CMI Archiv
Staatsarchiv Zug	-	-	CMI Archiv
Total		**382**	

Table 4: Overview of data-set extent extracted from the relevant institutions

In the following sections, the respective holdings, units of description and the exporting itself is described in more detail. The attributes/columns all originally are in German, and their consolidation, translation, and the data-cleaning process of the data-exports is described in detail in chapter 4.3. All exports can be accessed under https://github.com/specjo/ISADGtoRiC .

Stadtarchiv Zug (SAZG)

The archival holdings of the SAZG are managed with an AIS called "CMI Archiv". This AIS is among the few options available on the Swiss market, and therefore one of the most widespread AIS in Switzerland. CMI Archiv is the internal client for employees managing archives or collections. The tool also offers for external clients an integration in a publicly researchable web-client[44].

In order to export data, one has to either access directly the underlying SQL DB, or access through the internal client. The only option in my case is to use the client, since I don't have direct access to the DB. The client offers exports either as XML-file in the EAD-standard, or with a customizable report creation functionality (to extract data from the DB) in various other formats, including CSV. Since the export-target is CSV, I've opted to go with the report functionality.

I've created such a report, exporting 10 distinct attribute values for each unit of description. Unfortunately, I wasn't able to export any kind of "reference key" to link subordinate with superordinate (e.g. linking a dossier to its parent fonds) objects, and thus to extract the hierarchy within the software[45]. Luckily the logic of reference codes of objects in archives most of the times implicitly maps the hierarchical connections. Therefore, the hierarchy of the archival holdings isn't explicitly present in the exports, but can be reconstructed in the later step of data cleaning with the help of reference codes.

I've researched manually in the catalogue relevant units of description, and ran the custom report on the search results. Ultimately, I've exported from the SAZG holdings a data-set with 78 rows/units and 10 attributes/columns each.

Museum Burg Zug (MBZ)

The MBZ doesn't have a catalogue publicly available[46], like archives or libraries do. It keeps like most museums, only an internal inventory of its objects. I've approached the MBZ with a request for an inventory-excerpt for holdings pertaining to the pre-defined research topics. As an export, I received 6 separates .xls files (1 for each search topic, that rendered results in the inventory) with 12 distinct columns each. The files were combined with the excel Power Query functionality to a CSV-file, and the columns "Bibliografie" and "Literatur" were dropped right away for lack of values. For the MBZ, 31 distinct

[44] https://mobile.cmistar.ch/webclients/zgstadtarchiv/#/

[45] I've also spoken to the deputy head of the SAZG, who is administrator for the software. He also wasn't able to export any hierarchy-representing keys.

[46] Nonetheless, some of the holdings are accessible digitally: https://www.burgzug.ch/page/de/sammlung/sammlung-online

recorded units / rows were identified with 10 separate columns (see Appendix 1). There is no hierarchy among the holdings for MBZ; every row equals one object, and is directly held by the MBZ.

Bibliothek Zug - library catalogue

The cultural heritage items at the Bibliothek Zug are recorded in multiple places: a library catalogue, an archival catalogue, and others (e.g. finding aids for separate holdings, that haven't yet been integrated into the digital catalogue). Since the catalogues follow different internal logics and offer different attributes of the holdings, 2 separate exports have to be taken for the library and for the archival catalogue of BibZug.

The library catalogue follows the Resource Description and Access (RDA) standard for descriptive cataloguing, and is the main catalogue of BibZug. These holdings are recorded in a library service platform called Quria. This software is the main catalogue, offering multiple functionalities in one place: holdings and borrowing management, an online catalogue and website management[47], as well as client or event management. The StAZG keeps as an archive also a small library but doesn't have a library catalogue itself. On basis of a cooperation agreement, the library holdings of the StAZG are also recorded in the catalogue of BibZug, and could therefore be included in the export of BibZug as well.

For the storage and presentation of bibliographic records, Quria uses the IFLA Bibliographic Concept Models "work", "expression", "manifestation" and "item"[48]. The relation between these elements of the model is hierarchical, in the way as described in the Figure below. The cardinalities between the different levels are similar to ISAD(G), in such a way that each element might have multiple subordinate elements of the same type. For example, a work (e.g. "1984" by George Orwell) can be realized in 3 different expressions (the original in English; a German translation; and a graphic novel of the story); any of these expressions might be embodied in 3 distinct manifestations (the story published in 1949, in 1956, and in 1967); and a manifestation might be exemplified in a library by existing as 4 physical books.

[47] https://www.bibliothekzug.ch/

[48] Axiell (2022): Quria [Accessed: 30.05.2022]

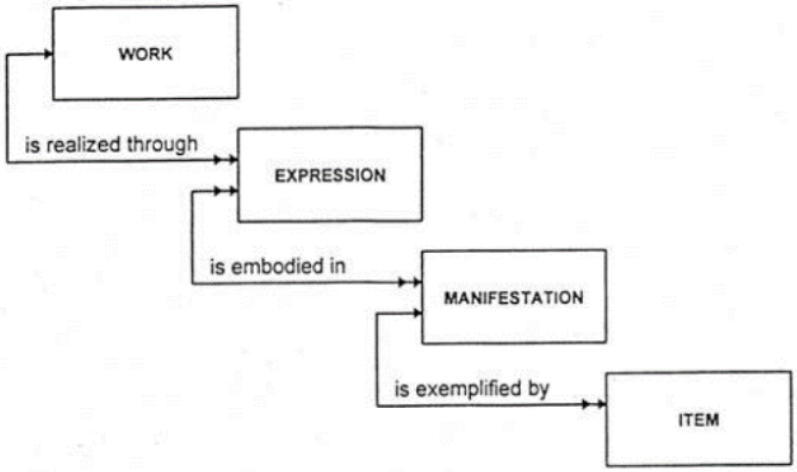


Figure 2: Figure 1: RDA "entities" and relationships1

During the initial set-up of the newly commissioned Quria software at BibZug in 2019, the bibliographic records from the previous DB were imported as MARC datasets. The software stores both, the original MARC-imports, as well as up-to-date bibliographic records according to RDA. Unfortunately, both data-sets are “view-only” as raw-data for internal users. Regrettably, the software doesn’t offer any functionality of exporting neither the RDA-based nor the MARC data as full datasets[49].

The only export functionality of bibliographic records is offered by the software on the hierarchically lowest RDA-level “item”. Given these technical limitations, and because mapping of RDA to RiC is not the topic of this thesis, a pragmatic decision was taken. For the bibliographic records, the DB entries on level “item” were not treated as components within the hierarchy RDA, but rather as single object / record resource. Thus, any item isn't part of a hierarchy, but is directly a "child-object" of the institution “BibZug”.

Here again, for each research topic a manual search query was run in Quria, and relevant bibliographic entries were exported to separate .csv files for each query. The files were merged via a Windows command prompt command *“copy *.csv quria_merged_exports.csv”* in the respective folder.

As a multi-purpose tool, Quria doesn’t only export bibliographic data-fields, but also administrative and technical. A total of 20 of these columns were dropped / deleted, because they were either blank or irrelevant for this purpose. Such columns included for example: “mode of acquisition”, “last activity”, “count of borrowing”, “due date of borrowing”, etc. After removing the unneeded administrative and technical columns, the dataset for the

[49] According to eMail-response by Quria Customer Support

bibliographic catalogue of BibZug remained containing 108 rows with 9 columns (see Appendix 1) each.

Bibliothek Zug - archival catalogue

Besides the described library catalogue, BibZug operates also an AIS. The reason why BibZug uses an archival catalogue as second inventory is twofold. Through the AIS, BibZug accesses its digital long-term archive and maintains the corresponding catalogue. The digital-long term preservation infrastructure is used for (1) retro-digitized collections (e.g. historical photography or digitized audio-visuals) as well as (2) the “Zuger Dokumentation”. The "Zuger Dokumentation" is a contemporary documentation, which was taken over in physical form from a now defunct association, and is since 2021 continued digitally. The internal hierarchy of this documentation was similar to the hierarchical logic of ISAD(G), and thus the decision was taken, to continue that documentation in an AIS instead of mapping it to a bibliographic catalogue, and thus breaking up its inherent hierarchy.

BibZug is part of the same administrative body, the City of Zug, and through cooperation both institutions jointly use the same AIS. Therefore, I was able to run the same custom exports as in SAZG also on the manually researched catalogue entries of BibZug. For every separate research topic, a separate report was run, and every of the 18 iterations led to a separate report in CSV. These were ultimately merged with the Power Query functionality of Excel, and the final exported data-set contains a total of 183 units of description.

Staatsarchiv Zug (StAZG)

The StAZG uses like the SAZG and BibZug the software CMI Archiv for the management of their archival holdings. Here as well, clients have access through a web client[50].

Initially, I've contacted the archival IT specialist, explained the content of this thesis, and asked whether an export of their DB would be possible (analogue to the SAZG and BibZug export). He remarked that exporting with the report functionality is quite cumbersome, and offered to pull an export from the SQL DB directly [51].

I've then manually researched in the web client holdings on different levels of description, relevant to the research topics. After submitting the request of export for datasets of the 95 identified records, with the request to include reference keys to be able to map the

[50] https://staatsarchiv.zg.ch/webclients/zgstaatsarchiv/#/

[51] Wehrle M. (StAZG) (2022a): eMail reply on 29 April 2022

archival hierarchy, I unfortunately received a negative answer. The explanation and apology were, that an export including the necessary IDs to map the hierarchy of the archive, was not possible without considerable additional time investment, which he didn't have at that time. He suggested instead to use the web-client to export some of the data-fields and to recreate the hierarchy using the IDs in the URL of each respective unit of description.[52]

I then decided to omit an export of the StAZG, because manual copy-pasting from an online catalogue isn't timely feasible or reasonable. In addition, with the exports of SAZG and BibZug I have already extracted exports from the software AIS, and the data from StAZG might not even show an added value. If at the end it nonetheless would become useful to include some data from the StAZG, individual data-sets could still be exported manually.

4.3 Transforming and mapping

Even though I have not exported the complete existing databases but only small sample-sets, one might anticipate that we are nonetheless left with a very heterogenous set of data. The exported data comes from 2 cataloguing types (ISAD(G), RDA) as well as a museal inventory, with varying attributes. Even in the three cases where the same standard is used (ISAD(G) in BibZug, SAZG, StAZG), the attribute values are neither homogenous among the institutions, nor within a single institution, but rather depend on the institutional tradition and when, how and by whom the holdings were described.

As with every DB in any field, over time erroneous and incomplete entries, inconsistencies, and other issues accumulate. Here, the popular computer and data science idiom "Garbage In, Garbage Out"[53] applies unfortunately as well. Data cleaning is a process, that is vital for any DB and for useful results in working with it. After exporting the datasets, and before mapping and importing to a new DB, is in this case the right time to perform data cleaning. The goal is to improve data quality and prepare it in such a way, that it is ready for mapping to RiC-triples and loading into an RDF DB.

The process of splitting of the original data-field and transforming it towards RiC happens in multiple steps. Below are the steps described briefly in theory, while in the following sub-chapters the same steps are carried out practically for the data exports:

[52] Wehrle M. (StAZG) (2022b): eMail reply on 09 June 2022

[53] Cambridge Dictionary (2022): Garbage In, Garbage Out

1. Identifying RiC-entities within a catalogue; assigning each entity an IRI (as described in 1.2), and assigning the corresponding attribute-columns to the entities.
2. Splitting columns that contain multiple kind of data
   - E.g. if a data-field/column “Name of Creator” holds the text “Friedrich Dürrenmatt, 1921 – 1990”, it should be split into the columns “name”, “birthDate" and “death-Date” for each type of information
3. Splitting .csv-File of the original export into multiple .csv-tables, one for each RiC entity type and connecting them using the IRIs as "Primary/Foreign keys" (because smaller tables are easier to work with and extended)
4. Mapping of the existing columns to triples through aid tables, in order to depict all transformations of columns to subject or objects, and the corresponding predicates (see Table 5 as an example)
5. With the help of the tool OntoRefine (an open-source data transformation tool based on Open Refine)[54], the columns from the .csv-File exports are imported, and mapped to triples according to the logic laid out in the aid table. The transformed output product is a Turtle File (.ttl).
6. The output-turtle is ultimately imported into a repository.

In the following chapter, this process is described in detail for exemplary catalogue exports and underpinned with the necessary appendices in the annex.

The mapping schemas are accessible as .json under:

https://github.com/specjo/ISADGtoRiC/tree/main/43_mapping_to_RiC ;

the mapped outputs can be accessed as .ttl-files under:

https://github.com/specjo/ISADGtoRiC/tree/main/43_mapping_output .

[54] ontotext (2022): Loading Data using OntoRefine

4.3.1 Mapping ISAD(G) – an SAZG example

The export of the SAZG contains records on all ISAD(G) levels, describes only analogous objects, and comes with 10 different distinct columns (see Appendix 1):

ID	Verzeichnungsstufe	Titel	Signatur	Frist
Form und Inhalt	Überlieferungsform	Zeitraum	Umfang	Archivalienart

The ISAD(G) units of description describe, when thinking in RiC, two things: the entity Record Set (RiC-E03) as intellectual content of a record, as well as the representation thereof, an Instantiation entity (RiC-E06). To make both distinguishable, the ID (assigned by the AIS to every item/unit of description) is used to create a compound IRI for RiC-E03 (`<B_E03_"id">`) as well as for RiC-E06 (`<B_E06_"id"_001>`), as shown in the table below. Each existing attribute (with a green background) in the first column has been relabelled to a node in the triples on the right side of the table (also green). The nodes with a yellow background were created manually, since they didn't exist in the original export.

SAZG AIS export	*subject (entity)*	*predicate (attribute)*	*object (entity)*
	`<B_E03_"id">`	*rdf:type*	rico:RecordSet
ID	`<B_E03_"id">`	*rico:Identifier*	"id"^^localSoftwareID
	`<E06_"id"_001>`	*rdf:type*	rico:Instantiation
	`<B_E03_"id">`	*rico:recordResourceToInstantiationRelation*	`<B_E06_"id"_001>`
Verzeichnungsstufe	`<B_E03_"id">`	*rico:hasRecordSetType*	ric-rst:"levelOfDescription"

Table 5: Aid table for entities RecordSet and Instantiation for the SAZG AIS export

The IRI for RiC-E06 receives a sequential number as suffix, because any single RecordSet could have an infinite number of Instantiations (e.g. analogue original, digitized usage copy, etc.). Both entity types are assigned its respective classes with the predicate *rdf:type,* and the original "ID" remains an identifier of the RiC-E03. Archival records normally have reference codes, which are another type of identifier than the software `"id"`. Thus, `"id"` is assigned a datatype (*^^localSoftwareID*), to keep it distinguishable from other types of identifiers. In this case, the "id" from the software is not mapped as an IRI, but as a literal value with an own datatype, because it isn't guaranteed that the "id" from the AIS-software is in fact unique, but could have the same string as a record from another software or export. The relation, that each Instantiation is an Instantiation of the respective RecordSet, is stated with *rico:recordResourceToInstantiationRelation*. The existing ISAD(G) "Verzeichnungsstufe" is relabelled to "levelOfDescription" within the to the Record Set Type vocabulary (*ric-rst*), and mapped with the predicate *rico:hasRecordSetType*.

Besides the exported columns, an additional one is created manually, since the described Report-export functionality, doesn't allow to export hierarchical relations between 2 items in the AIS. In order to preserve this hierarchy, the column "superordinate" is created on basis of the underlying Signature (reference codes assigned according to the hierarchy of the archive), containing <E03_"ID"> of the superordinate unit of description. The entity <E03_"id"> is then connected with the predicate *rico:isPartOf* to its respective "superordinate" (see Table below). The exported attribute "Signatur" is mapped similarly to <id> above, as *rico:Identifier* with a datatype *(^^localHolderRefCode)*. Furthermore, the column *"Titel"* was exported from the AIS. This *"title"* is mapped as the name of the entity RecordSet (with *rico:hasOrHadName*), and with a prefix *"Instantiation of"* it is also mapped as name of the corresponding Instantiation entity. Both are assigned a suffix @de to denote that they are in German language. Also, both are assigned the respective title through *rdfs:label,* because else the graph DB would use the IRI to label nodes (e.g. https://www.wikidata.org/wiki/Q12519). With the predicate *rdfs:label* a human-readable label (e.g. "Magna Carta") can be assigned to a node, making it user-friendlier.

SAZG AIS export	*subject (entity)*	*predicate (attribute)*	*object (entity)*
superordinate	<B_E03_"id">	*rico:isPartOf*	<superordinate>
Signatur	<B_E06_"id"_001>	*rico:Identifier*	"signature" ^^localHolderRefCode
Titel	<B_E03_"id">	*rico:hasOrHadName*	"title"@de
	<B_E03_"id">	*rdfs:label*	"title"
	<B_E06_001_"id">	*rico:hasOrHadName*	"Instantiation of" & "title"@de
	<B_E06_001_"id">	*rdfs:label*	"Instantiation of" & "title"

The remaining exported columns from the AIS are then also assigned either the entity RecordSet or Instantiation. "Form und Inhalt" ("scopeAndContent"), "Zeitraum" ("creationDate"), "Archivalienart" (<contentType>) and "Frist" (<protectionPeriod>) describe attributes of the RecordSet, irrespective of its physical or digital appearance. The last 2 attributes "Überlieferungsform" (<representationType>) and "Umfang" ("extent") then describe attributes of the physical Instantiation, and are therefore mapped to the entity <E06_"id"_001>, as described below.

SAZG export column	*subject*	*predicate*	*object*
Form und Inhalt	<B_E03_"id">	*rico:scopeAndContent*	"scopeAndContent"
Zeitraum	<B_E03_"id">	*rico:creationDate*	"creationDate"
Archivalienart	<B_E03_"id">	*rico:contentType*	<contentType>
Frist	<B_E03_"id">	*rico:conditionsOfAccess*	<protectionPeriod>
Überlieferungsform	<B_E06_"id"_001>	*rico:hasRepresentationType*	<representationType>
Umfang	<B_E06_"id"_001>	*rico:carrierExtent*	"extent"

This concludes the mapping of the export from the AIS as it is, describing the entities RecordSet and the physical Instantiation, as well as various attributes. Having in mind that later on records from different institutions will be linked and combined in a single DB,

one more statement is missing; who the holder of the respective records is. For that reason, additional triples are stated in the mapping:

SAZG export column		*subject*	*predicate*	*object*
		`<B_E03_"id">`	*rico:hasOrHadHolder*	`<wd:Q27490192>`
		`<B_E06_"id"_001>`	*rico:hasOrHadHolder*	`<wd:Q27490192>`
		`<wd:Q27490192>`	*rdfs:label*	`"Stadtarchiv Zug"`

Both the RecordSet entity as well as the Instantiation entity have SAZG as holder, which is represented as an entity through the IRI based on the Wikidata entry, and that IRI has the *rdfs:label* "Stadtarchiv Zug" as above, that it becomes better human-readable in a knowledge graph.

The mapping is now fully laid out in an aid-table (see Appendix 2), and ready to be applied to the CSV-export. The logic of the mapping is transferred, as explained above in step 5, into OntoRefine.

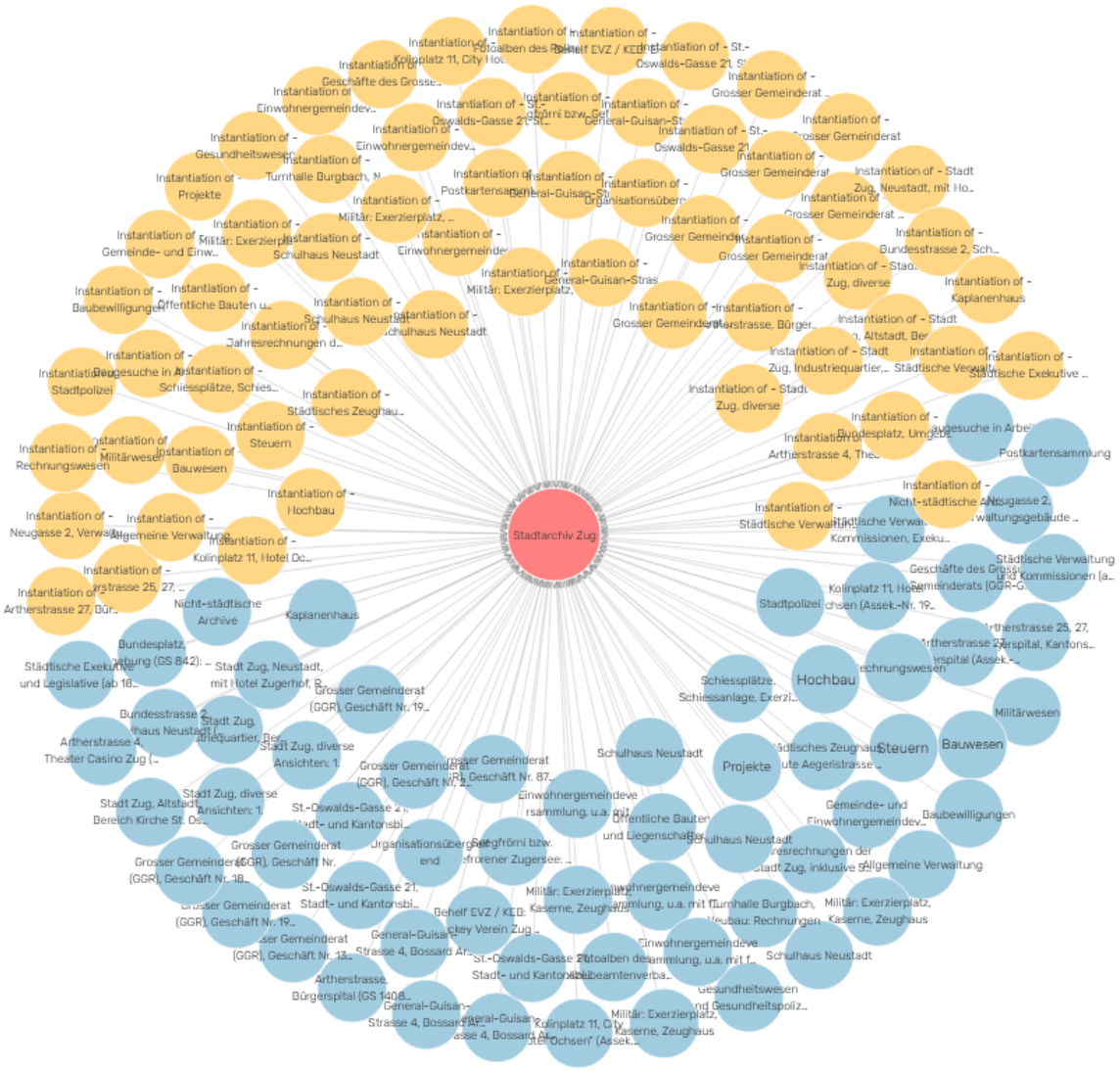


Figure 3: Knowledge graph of partial holdings of the SAZG

The data has now been transformed to RDF, and can be used to create knowledge graphs. For now, only a mapping to RiC-entities RecordSet and Instantiation has been executed. All these entities are held by the SAZG, and these relations would look in a knowledge graph as follows, where the red node is the SAZG, the blue nodes stand for the RecordSet, and the yellow nodes represent Instantiations:

Depicting a too large subset of data might give a first overview, but has its obvious downsides. It quickly becomes unmanageable, and useful statements about contained information can be hardly made. Thus, it makes sense to dive deeper into the data, and have a more focused knowledge graph, as the following example shows.

The datatypes used in this example [`(^^localSoftwareID)` and `(^^localHolderRefCode)]` are minimal, but illustrative examples, of where a defined internal vocabulary for an institution would make sense, if one couldn't make use of an external vocabulary.

So far, only a mapping of the existing data to RiC has been executed, and the data hasn't been enriched with more meaningful data, which would be quite easily possible with RDF. Nonetheless, depicting a knowledge graph starting from the file / unit of description `"Neugasse 2, Verwaltungsgebäude (Assek.-Nr.914a, GS986)..."`, one can already see quite well the information network within the existing data. Among other thing, it is visible that this specific (former) unit of description is now an entity of type `RecordSet` (as are all other red nodes), has the `RecordSetType` "Dossier" (English ISAD(G) equivalent: "`File`"), and also

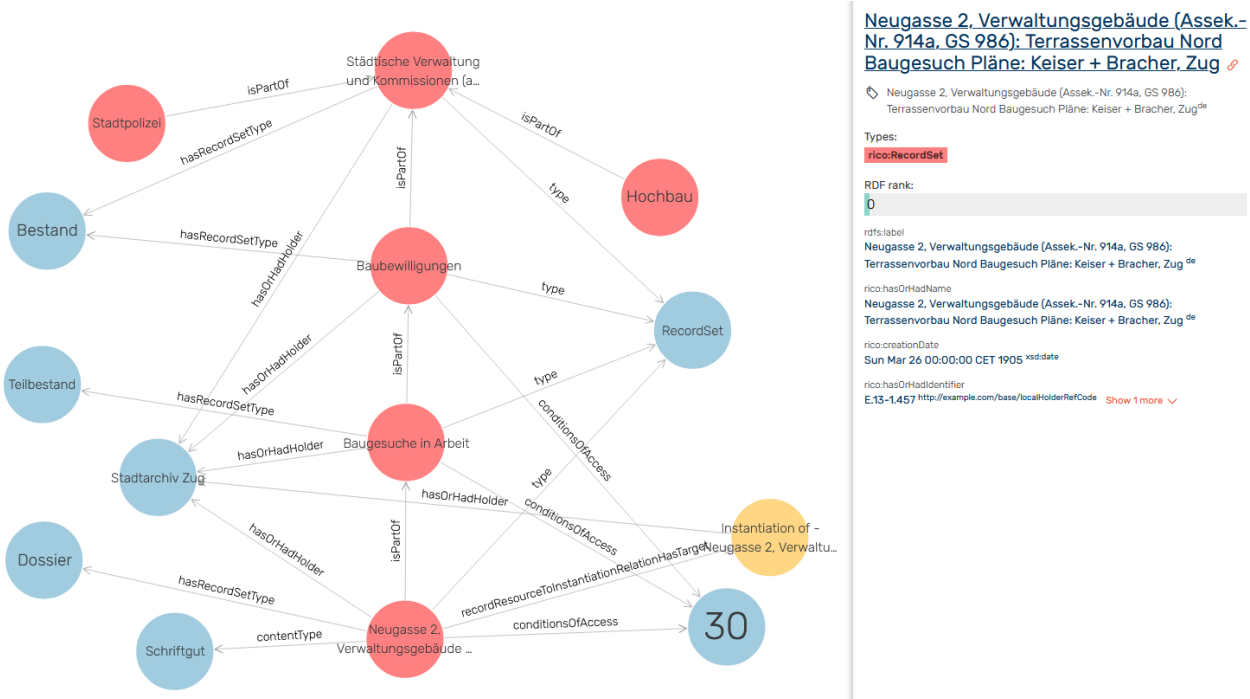


Figure 4: Knowledge graph of an exemplary file from the mapped SAZG export

rico:isPartOf the unit of description "*Baugesuche in Arbeit*"[55]. This entity is itself a *RecordSet* that *hasRecordSetType* "*Teilbestand*" (English: "*Sub-fonds*"), and is also part of a superordinate *RecordSet* "*Baubewilligungen*"[56], that is a "*Bestand*" (English: "*Fonds*").

The RecordSet on the very top of the hierarchy is "*Städtische Verwaltung und Kommissionen*"[57], which contains two other exemplary entities "*Hochbau*"[58] and "S*tadtpolizei*"[59]. And finally, the initial unit of description/file *"Neugasse 2, Verwaltungsgebäude (Assek.-Nr.914a, GS986)…"* has an Instantiation

represented with the yellow node, which has content of the Type "*Schriftgut*"[60], and is described by various literal objects. These literals are not represented as nodes, because they aren't IRIs and therefore can't function as nodes of the network, but are in the example above shown for the chosen *RecordSet* on the right-hand side as literals. These contain such objects as the *rico:creationDate* mapped as datatype "*xsd:date*", or the current archival identifier "*E.13–1.457*", mapped as datatype "*localHolderRefCode*".

As such, the current existing archival hierarchy according to ISAD(G) and its elements have been mapped and reproduced one-to-one to RiC. In addition, the knowledge graph gives an outlook on RDF in general with the visual representation of interlinked triples and networks of information. It also gives an outlook on RiC with the use of first pieces from the RiC-O like classes (the entities RecordSet and Instantiation) or object properties (e.g. *rico:hasOrHadHolder*), and datatype properties (e.g. *rico:creationDate*). But besides the entities E03 RecordSet and E06 Instantiation, no new RiC-entities (except for the holder "Stadtarchiv Zug", which wasn't stated as RiC-E11 CorporateBody) have been introduced. More elements, their relationships and possible implementations are explored in the following sub-chapters.

4.3.2 MBZ transformation – example object

A mapping of the MBZ export is quite useful to explore RiC functionalities, since the museum DB offers additional possible entities besides the ones mapped with the SAZG export above. As described in 4.2, the export of the DB of MBZ offers 31 rows, each for one

[55] English: Building applications in progress

[56] English: Building permissions

[57] English: Municipal administration and commissions

[58] English: Building construction

[59] English: City Police

[60] English: written materials

item in the collection. Every single entity/item has 10 attribute columns. An example description of a single object (a black/white photograph of the former military barracks of Zug, nowadays the library of Zug) is shown in the Table below, and will be used to illustrate the mapping steps.

MBZ collection object [original export]	
attribute	*attribute value*
Objekt/Titel	Fotografie der Ostfassade der heutigen Stadt- und Kantonsbibliothek, ehemals Kornhaus, dann Kaserne
Inv.-Nr.	15054
Pers./Körp.	Josef Maria Weber-Strebel (1842 - 1933), Fotograf/in
Datierung	um 1900 (Schätzung)
Masse	Objektmass (H x B x T): 16.3 x 22.3 cm
Inschr./Sign.	Inschrift: auf RS Stempel: J. M. Weber-Strebel, Zug
Mat./Technik	schw. - weiss Fotografie, auf Karton geklebt
Beschreibung	Das Gebäude in der Funktion der Kaserne, mit Wachtposten davor.
Zustand	Zustand: fleckig
Reprorecht	Museum Burg Zug

Table 6: Example collection object / entity export from MBZ

So, what is happening on a logical level with the data when transforming to RiC? The information held in the attribute value fields doesn't only describe attributes of the object as a single entity. It rather describes according to RiC attributes of different entities. For a better understanding of data transformation, cleaning, and mapping, I will in this example go one by one over what kind of RiC-Entities[61] and RiC-Attributes[62] are in this dataset to be found. The RiC *entities* implicitly identifiable in this record are:

- RiC-E03 Record Set: the description of the intellectual content of the Record Set; in this case the motive of the photograph
- RiC-E06 Instantiation: the description of the representation (in this case physical, might also be digital) of the photograph (this particular example is a positive photo print of the original negative photographic Instantiation, which is not recorded in this context)
- RiC-E07 Agent: Multiple agents are involved; "Josef Maria Weber Strebel" as photograph; "Museum Burg Zug" as holder of the record, and as holder of reproduction rights

[61] RiC-CM: p. 19 - 38

[62] RiC-CM: p. 39 - 62

- RiC-E18 Date: The attribute “Datierung” describes a date (what exactly happened on that date? Was the picture taken then, was the positive photo print created, or changed the picture owner; it isn’t explicitly recorded!)
- RiC-E22 Place: the attribute “Objekte/Titel” describes a physical location, the “Stadt- und Kantonsbibliothek” (city and cantonal library of Zug), formerly “Kornhaus” (granary), and afterwards “Kaserne” (military barrack).

Each single of these entities should be identified with an IRI, and could be described by different attributes (e.g. the profession photographer in data field “Pers./Körp.”). Let's start with the photographer as an agent-entity. An IRI[63] based on its entry in Wikidata is assigned (Row 0 in the Table below). The original single data-field holds information describing different attributes, like the name, the dates of birth and death, as well as the occupation. Each of these attributes is a separate element of information with a different meaning and context, and thus should be stored as separate data-piece. To depict the subject-predicate-object relations and to map the data to the triples, the aid-table below is helpful. Thus, the original single data-attribute is split and mapped into RiC is according to the following table:

	subject *(entity)*	***predicate*** *(relation)*	***object*** *(entity)*
0	wd:Q111594077		
1		*rdf:type*	*rico:person*
2		*rico:hasOrHadName*	"Josef Maria Weber-Strebel"
		rdfs:label	"Josef Maria Weber-Strebel"
3		*rico:hasBirthDate*	"1842" ^^xsd:date
		rico:hasDeathDate	"1933" ^^xsd:date
4		*rico:hasOrHadOccupationOfType*	wd:Q33231
5	wd:Q33231	*rdfs:label*	"Photographer"@en

Table 7: Aid table for cleaning data in original database exports

- 0: The IRI for the agent-entity is defined
- 1: The entity is assigned an entity-type (rdf:type), which is rico:person (RiC-E08 Person)
- 2: The entity is assigned a name (rico:hasOrHadName), which is the literal "Joseph Maria Weber-Strebel".
- 2: The entity is also assigned a label (rdfs:label) "Joseph Maria Weber Strebel", which will be used as a human-readable label on the knowledge graph, instead of a “cryptic” IRI

[63] https://www.wikidata.org/wiki/Q111594077

- 3: The entity is assigned a rico:birthDate "1842" and a rico:deathDate “1933”; for both the datatype is defined (^^xsd:date)
- 4: The entities occupation is defined with rico:hasOrhadOcupationType, where the object is an IRI of the profession photographer from Wikidata64, in order to be able to link to other agent-entities with the same profession/ IRI.
- 5: The object is assigned the label "Photographer", so that it becomes human-readable and doesn’t show the IRI in a graph. In addition, the language of the label is defined with @en.

As in the previous SAZG example, the mapping has here also been executed with the help of OntoRefine, and the output in Turtle-Syntax renders the result as shown in the snippet below. In the beginning of the datafile, the prefixes and ontologies used are defined:

```
@prefix rico: <https://www.ica.org/standards/RiC/ontology#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix wd: <https://www.wikidata.org/wiki/>.

wd:Q111594077   rdf:type                    rico:person;
                rico:hasOrHadName           "Joseph Maria Weber-Strebel";
                rdfs:label                  "Joseph Maria Weber-Strebel";
                rico:hasBirthDate           "1842"^^xsd:date;
                rico:hasDeathDate           "1933"^^xsd:date;
                rico:hasOrHadOccupationOfType   wd:Q33231 .

wd:Q33231       rdfs:label                  "Photographer"@en.
```

Figure 5: RDF text snippet of mapped RiC-E07 entity of the creator of the example collection object

The agent-entity has been created as IRI and is the subject of the triple-statements with various predicate-object relations. Importing it into a graph DB (with the necessary name-spaces, settings and prefixes; more under 4.4) creates the first very basic knowledge graph as visualized in the Figure below, and makes these pieces of information easier human-readable.

"*Joseph Maria Weber Strebel*" has become an entity, and is a node[65] (the red dot) in the network. That object is linked with the predicates <*type*> and <*hasOrHadOccupationType*> to the other objects/entities <*person*> and <*Photographer*>. These relations, predicates linking multiple nodes, are called arcs[66] or edges. Since the objects "*Person*" and "*Photographer*"

[64] https://www.wikidata.org/wiki/Q33231

[65] ICA (2021): RiC-CM, v0.2, p. 6

[66] ICA (2021): RiC-CM, v0.2, p. 6.

were mapped as IRI, they aren't only object of triples, but could also be a subject in another triple-statement themselves. In other words, they can serve as nodes linking to other nodes in a growing network, and thus expanding the knowledge graph and the information network. The other objects that were mapped to literal values (like `rico:hasBirthDate` or `rico:hasOrHadName`) are shown on the righthand side of the Figure below as attributes of the entity *Joseph Maria Weber Strebel*. Since they aren't mapped to IRI's, they can't function as subjects and nodes themselves, but remain simple "describing" objects linked to the specific subject.

This procedure of cleaning the data, splitting columns where necessary and moving information bits into separate data-fields needs to be executed for all records, all exports, and for all columns.

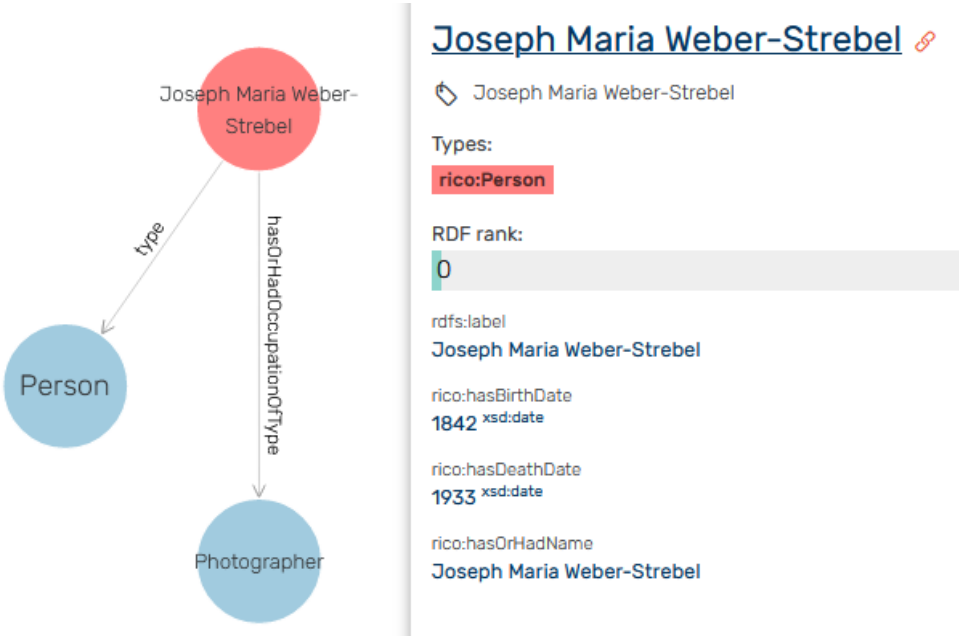


Figure 6: Knowledge graph visualizing information about a person, mapped from a RDB data-field to a triple-store

The original export of MBZ had 31 records/rows with 10 columns. Only focusing on the column "`Pers./Körp.`" a total of 8 different agents were identified (including mister Weber Strebel). The data of these agent-entities was cleaned and mapped all according to the same logic as described above, and loaded into the graph DB. In addition, the entities that were assigned the type `rico:person` or `rico:corporateBody`, were again assigned the superordinate entity-type <`Agent`>; information which doesn't exist as such in the original export, but is defined by RiC-CM. The full knowledge graph depicting all agent-entities is shown on the figure below. With such a small data-set, it is quite clearly visible on first glance, what information the data contains. There are 3 agents of the entity type <`CorporateBody`>, and 5 of entity type <`Person`>, and while <`CorporateBody`> is a sub-class of

<*Group*>, both entity-types ultimately are a sub-class of the entity type <*Agent*>. Of the 5 Agents, only 4 were by occupation photographers, and 1 of them was a painter.

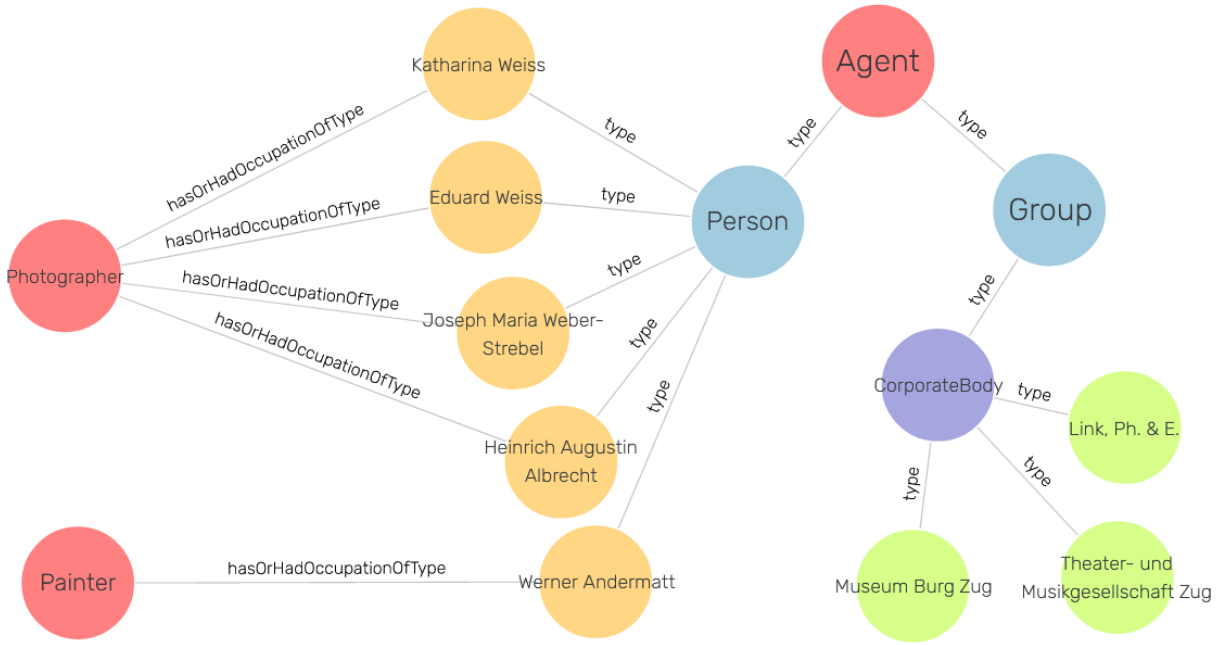


Figure 7: Knowledge graph of Agents mapped to RiC from the MBZ-database export

This short cleaning and mapping task of the single column "Pers./Körp." of the MBZ export is quite useful to illustrate the move from RDB to RDF. It becomes evident, that moving from a previous description (standard) in RDB to an RDF-based standard like RiC, is more than just mere 1-to-1 translation of previous attributes into a new ontology. *The mapping itself must be well prepared, but mapping inadequate data won't improve the data miraculously. Thus, cleaning the data before mapping is of utmost importance, even though it might involve a lot manual and time-consuming labour.*

With the mapping of only one attribute of the original dataset multiple subjects, predicates and objects were created. The full dataset originally holds information about different entities (Agent, Record Set, Instantiation, etc.), as described in 4.2, but only implicitly. The entities aren't defined as such, and thus the relations between them also aren't explicitly defined; an omission that shall be rectified.

When mapping the dataset, I decided to create in this case (in contrast to the SAZG mapping) for each entity a separate csv-table, split the columns in the data-export along the entities and attributes, transfer every entity with its attributes to its separate sheet, and to connect the entities with the respective adequate predicates/objects in each sheet. The splitting of the existing attributes and the mapping is outlined in Appendix 3.

After repeating the data-transformation described above for the full export, it was imported to the graph DB. The original object described at the beginning of this chapter results in the knowledge graph below.

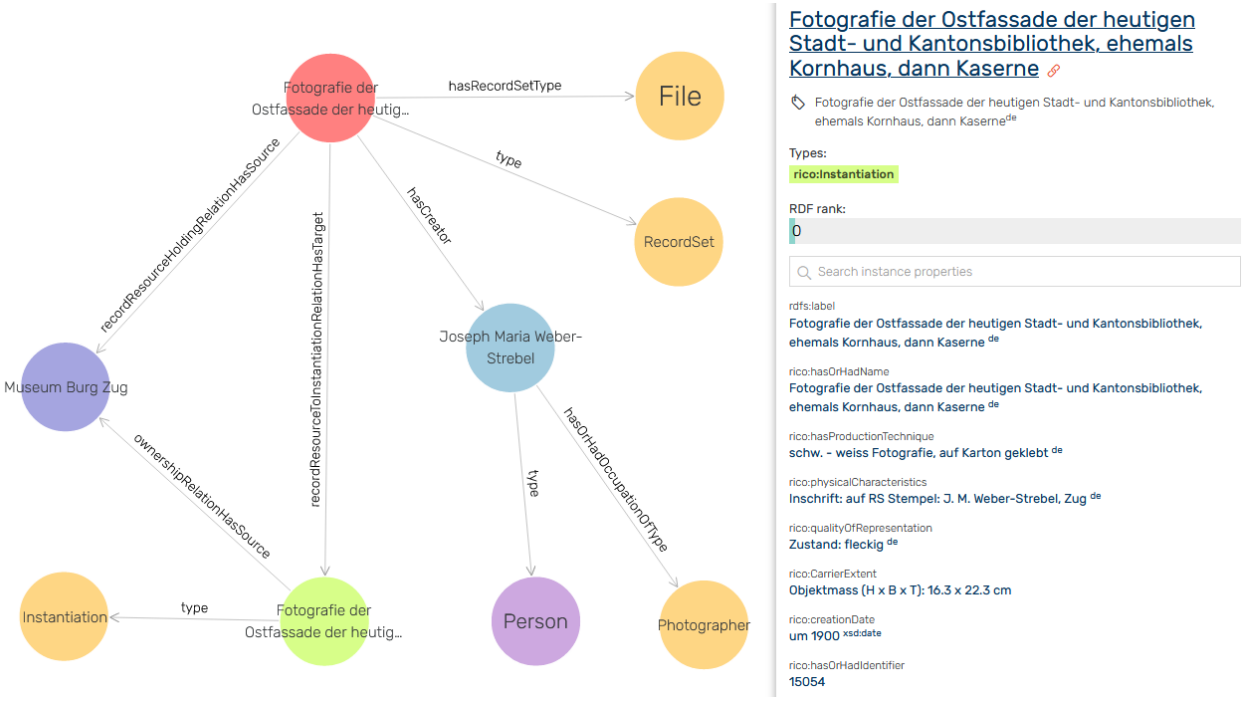


Figure 8: Knowledge graph visualizing an example collection item after mapping from RDB to RDF

Each entity is now depicted as an own node; the photographer Joseph Maria Weber-Strebel and the holder of the item Museum Burg Zug as agents, the intellectual content of the photograph as RecordSet, and the physical positive of the photograph as Instantiation. The data-fields, which were mapped as literals and not as IRI's, are here again depicted for each node on the righthand side. In this example, it is the literals of the Instantiation, describing attributes like physical characteristics or the creation Date.

All the steps so far described were now applied according to Appendix 2 to the full data-export of the holdings of MBZ, and the transformed data was imported into the graph DB. Showing the full dataset with all nodes and edges in one knowledge-graph makes even such a small dataset unreadable. Therefore, I've created for illustrative purposes a knowledge-graph based on 2 manually picked RecordSets with a sub-set of information, as shown below:

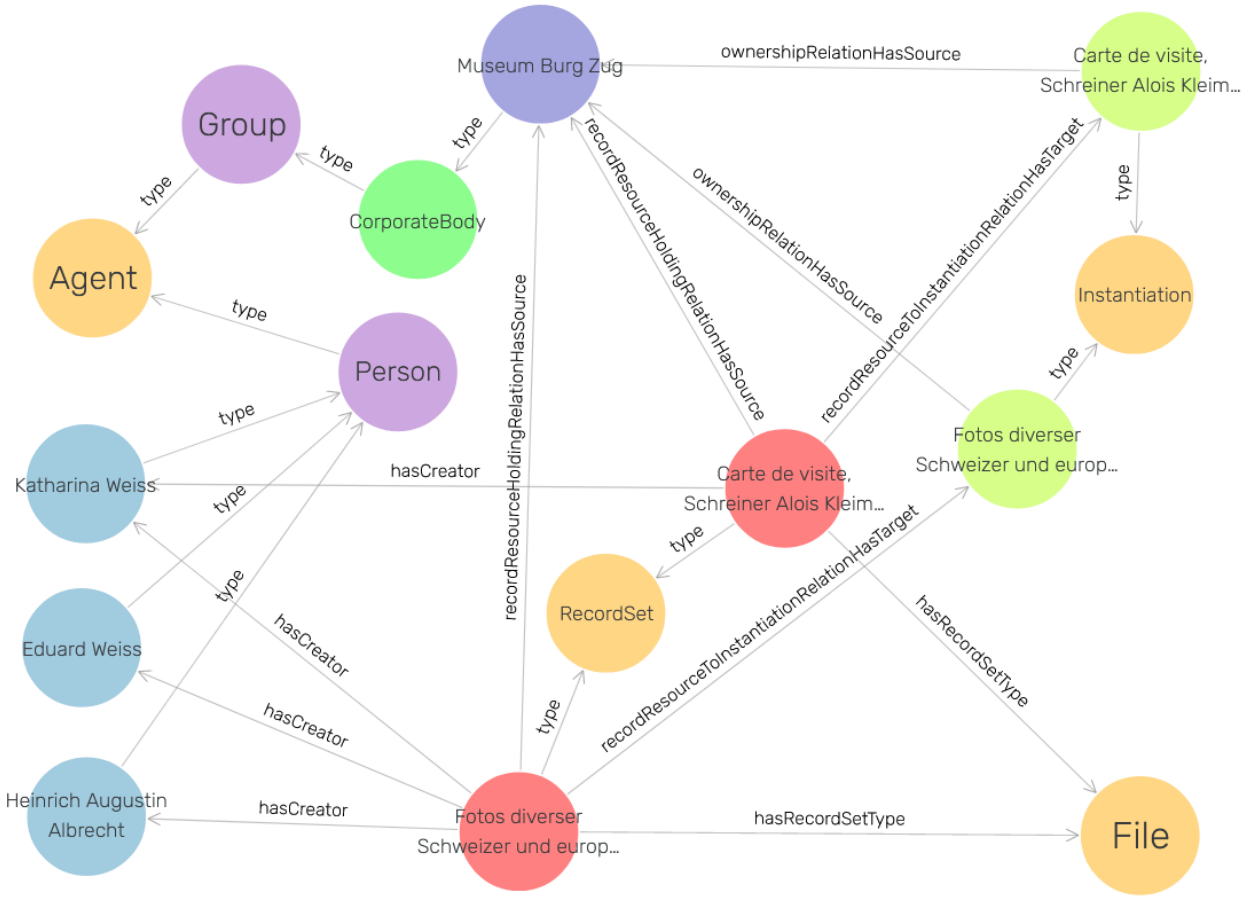


Figure 9: Knowledge graph of an extract of the MBZ-export fully mapped to RiC/RDF

Now with this final knowledge graph of the MBZ mapping, once can make meaningful statements about the objects in the Museum and various relations. What has previously been just data in a RDB now has become what is sometimes referred to as meaningful data. While previously any meaning to the data was implicitly recorded in the RDB (if at all), now the meaning of data is explicitly capture.

To exemplify the meaning of the information in this knowledge graph for the MBZ, I've chosen 5 Record Set Entities (yellow nodes, RiC-E03), that have the creator entity <Katharina Weiss>. One of those Record Sets (<`Fotos diverser Schweizer und europäischer…`>) has not only <Katharina Weiss> as creator entity, but also <Eduard Weiss> as well was <Heinrich Augustin Albrecht>. All 3 of them, as well as <Werner Andermatt> are people, i.e. have the entity type Person (RiC-E08). And out of these, <Katharina Weiss> has the occupation type <Photographer>, and <Werner Andermatt> is a <Painter>. The 5 Record Sets all are owned by the entity <Museum Burg Zug>, which has the entity class <Corporate Body>. <Museum Burg Zug> also owns Instantiations (red nodes, RiC-E06), which are themselves physical instantiations of the previously described Record Set entities.

With this knowledge graph it becomes clear that the previously inherent meaning of the data can, in fact, be thought rather as a network of information, and not as tabular listing. This knowledge graph also gives an insight, in what direction a larger network of information could go, and how previously closed, self-contained holdings description could be connected to information outside of the description, to better describe the full context in which records are created and exist.

4.3.3 Mapping the BibZug archival holdings

One could expect that the mapping of the archival holdings of BibZug is similar to the one of SAZG, since both share their catalogues and jointly use the same installation of the CMI AIS. The exports for both institutions have been done with the same custom report functionality, and hold the same 10 attributes (see Appendix 1). The export of BibZug contains 2 additional attributes "Urheber" (originator) and "Format" (photography format), which primarily stem from a photography collection, of which the SAZG doesn't have a description equivalent. These attributes need an additional mapping besides the ones that could be taken over from the SAZG mapping.

Unfortunately, not only different attributes require a different mapping, but even the same attributes. In the case of BibZug, this means for example, that some records of the above described "Dokumentation"-fonds (mainly consisting of thematic Files with newspaper articles as Items), might hold information for the attributes "*title*" and "*scopeAndContent*" like this:

Id	*title*	*scopeAndContent*
313181	Adolf Ogi der vergessene Bundesrat (Zuger Zeitung)	Publikation: Zuger Zeitung, Ausgabe: 263 Seite: 6 Untertitel: Der frühere SVP-Magistrat war zur Zeit der Crypto-Affäre Verteidigungsminister. Seine Einschätzung war bei der Aufarbeitung jedoch nicht gefragt. Autor: Lucien Fluri

Table 8: Example item of the "Zuger Dokumentation"- fonds from the BibZug export of the archival holding

In the case of SAZG the attribute values of "*scopeAndContent*" were simply mapped 1-to-1 to literal objects with the predicate *rico:scopeAndContent*. To apply the same mapping for BibZug as well would be a lazy approach, because the attributes "*title*" and "*scopeAndContent*" hold additional information. The attribute "*title*" doesn't only hold information pertaining to the RiC-A28 Name attribute, but also where the article was published (*Zuger Zeitung*), and should thus be split and also mapped with the predicate *rico:hasPublisher*. The attribute "*scopeAndContent*" holds the same information about the publisher ("*Publikation: Zuger Zeitung*") as well, and provides besides that, information about the sub-title

of the article ("*Untertitel*") and the Author *("Autor"),* requiring again splitting and separate mapping.

Another mapping issue, can be illustrated with the items from the postcard-fonds of Bib-Zug. Here again, the attribute *scopeAndContent* holds other information than above.

Id	*title*	*scopeAndContent*
166544	Hotel Restaurant Ochsen Zug	Ort: Zug Verlag: Kalt-Zehnder, Zug Fotograf: Grau, Zug Farbe: s/w Text: nein

Table 9: Example item of the Postcard-collection fonds from the BibZug export of the archival holding

In that attribute in fact things like *rico:place* (*Ort: Zug*), the *rico:hasPublisher* (*Verlag, Kalt-Zehnder, Zug*), the creator/photograph (*Fotograf: Grau, Zug*), and whether the post-card is black-and-white or coloured (Farbe: s/w), and whether there is hand-written text or not (*Text: nein*) on it.

This brief example illustrates an issue, that many institutions face now and will face when transitioning to RiC. If an attribute or an entity is not clearly defined, it can hold all kinds of information. The ISAD(G) attribute *scopeAndContent* is such a field, that is often being used by institutions to "collect" any information that doesn't fit to another attribute. Therefore, it will be vital that all existing attributes undergo thorough inspection and preparation before mapping to RiC. They certainly will hold within a single attribute information, that describes other entities and attributes, and even when identifying different types of information, the same way of splitting an attribute will not necessarily be suitable for all records of the same institution. It can't be emphasized enough, that when transitioning to RiC it is the right time to clean data, and only cleaned data will be able to use the potential of RiC and of Linked Data! This is a big task for institutions, but should primarily be seen as a big opportunity.

How did I go about in this specific case? Similarly to the SAZG mapping, the *ID* from the AIS is used to create a compound IRI for RiC-E03 and for RiC-E06, as shown in the table below, and both entity types are assigned its classes with the predicate *rdf:type*, while the original "ID" remains a literal identifier of the RecordSet with the datatype *localSoftwareID*, as shown below:

BibZug AIS export		subject	predicate	object
		<E03_"id">	*rdf:type*	rico:RecordSet
ID		<E03_"id">	*rico:hasOrHadIdentifier*	"id" ^^localSoftwareID
		<E06_"id"_001>	*rdf:type*	rico:Instantiation
		<E03_"id">	*rico:recordResourceTo InstantiationRelation*	<E06_"id"_001>
Verzeichnungsstufe		<E03_"id">	*rico:hasRecordSetType*	ric-rst:"levelOfDescription"

The relation between every RecordSet RiC-E03 and its respective Instantiation RiC-E06 is stated with *rico:recordResourceToInstantiationRelation*. The existing ISAD(G) "*Verzeichnungsstufe*" is also relabelled to "*levelOfDescription*" within the Record Set Type (*ric-rst*) vocabulary, and mapped with the predicate *rico:hasRecordSetType*.

In this case again an additional column "superordinate" has to be created to be able to preserve the existing archival hierarchy between the various RecordSets RiC-E03 with the predicate *rico:isPartof*.

BibZug AIS export		subject	predicate	object
superordinate		<E03_"id">	*rico:isPartOf*	<superordinate>
Signatur		<E06_"id"_001>	*rico:Identifier*	"signature"^^localHolderRefCode
Titel		<E03_"id">	*rico:hasOrHadName*	"title"@de
		<E03_"id">	*rdfs:label*	"title"@de
		<E03_"id">	*rico:hasPublisher*	<publisher_id>
Titel		<publisher_id>	*rdfs:label*	"publisher"
		<publisher_id>	*rdf:type*	rico:Agent
Titel		<E06_001_"id">	*rico:hasOrHadName*	"Instantiation of" & "title"@de
		<E06_001_"id">	*rdfs:label*	"Instantiation of" & "title"@de

Again, "Signatur" remains a *rico:Identifier* with a datatype *(^^localHolderRefCode)*. For this case, *"Titel"* is split to "*title*" and *"publisher*". The publisher is then assigned an IRI <publisher_id>, while serving as subject after the *rico:hasPublisher* predicate, while either being an entity *rico:Person* or *rico:CorporateBody*. The split literal text is then as *rdfs:label*. "*Title*" is then mapped as name and label of the RecordSet, and also with the prefix *"Instantiation of"* as name and label of the corresponding Instantiation. Both are assigned a suffix @de to denote the German language, as shown above.

For the archival holdings of BibZug, some records (e.g. Photographs) have a creator recorded separately in the field "Urheber", while other records (e.g. newspaper articles in the "Zuger Dokumentation") have an "Autor" recorded in the field "Form und Inhalt"("scopeAndContent"). Now RiC-O would offer for a creator of a photograph the predicate *rico:hasCreator* (RiC-R027). This property is a super-property of a more precise predicate *rico:hasAuthor* (RiC-R079), which would in this case be applicable e.g. to the authors of the newspaper articles. Since in RiC-CM Author is a sub-type of Creator, and for the sake of simplicity, in this case I will extract this information from both attribute fields, and map

both types with *rico:hasCreator*, even though RiC would offer a more detailed distinction. That newly created entity is assigned the type *rico:Person*, and is labelled with the exported attribute value "*creator*".

"Form und Inhalt" ("*scopeAndContent*") holds then besides the creator more information that needs to be extracted from that attribute. This is the right place to introduce a new RiC-entity: Place (RiC-E22). This entity faces a similar problem as the above-described creator/author of a record. It doesn't exist as such in ISAD(G), and thus normally any relation to a place/location is often described in the attributes *"title"* or "*scopeAndContent"* of ISAD(G) catalogues. Therefore, here again as with the creator, references to a place are extracted from the attribute "title" and/or "*scopeAndContent*", and saved in a new column called <*E22_place_name*>. To each distinct place a <E22_place_id> is assigned, and that IRI is assigned its RiC-Entity with *rdf:type*. Finally, the relation of the RecordSet to the Place-entity is modelled with the predicate *rico:isAssociatedWithPlace*.

BibZug AIS export		*subject*	*predicate*	*object*
Form und Inhalt		<E03_"id">	*rico:scopeAndContent*	"scopeAndContent"
		<E03_"id">	*rico:hasCreator*	<creator_ID>
		<creator_ID>	*rdf:type*	rico:Person
Urheber		<creator_ID>	*rico:hasOrHadName* *rdfs:label*	"creator"
Form und Inhalt		<E03_"id">	*rico:isAssociatedWithPlace*	<E22_place_id>
		<E22_place_id>	*rdf:type*	rico:Place
Form und Inhalt		<E22_place_id>	*rico:hasOrHadPlaceName* *rdfs:label*	<E22_place_name>

The remaining exported columns are mapped 1-to-1 as their RiC equivalents to either RecordSet or Instantiation according to the same logic as laid out in the previous SAZG mapping:

BibZug AIS export		*subject*	*predicate*	*object*
Zeitraum		<E03_"id">	*rico:creationDate*	"creationDate"^^xsd:date
Archivalienart		<E03_"id">	*rico:contentType*	<contentType>
Frist		<E03_"id">	*rico:conditionsOfAccess*	<protectionPeriod>
Format		<E06_"id"_001>	*rico:carrierExtent*	"carrierExtent"
Überlieferungsform		<E06_"id"_001>	*rico:hasRepresentationType*	<representationType>
Umfang		<E06_"id"_001>	*rico:instantiationExtent*	"extent"

And ultimately, the initially defined RecordSet and Instantiation are linked to it's holder, BibZug:

	<E03_"id">	*rico:hasOrHadHolder*	<wd:Q856475>
	<E06_"id"_001>	*rico:hasOrHadHolder*	<wd:Q856475>
	<wd:Q856475>	*rdfs:label*	"Bibliothek Zug"@de

The entity *RiC-E22 Place* has only been now introduced, because it normally doesn't exist in a distinct attribute according to ISAD(G), and thus isn't strictly part of a mere mapping, but has to be extracted from other attributes. Nonetheless, this entity is a fitting transition

to the next chapter "Mapping library holdings", because we will encounter the same attribute there again.

The "archival holdings" of BibZug strictly speaking aren't archival holdings, but a documentation following the hierarchical description logic of ISAD(G). Thus, these holding can nonetheless serve as a illustrative example of the integration of the ISAD(G) hierarchy into RiC:

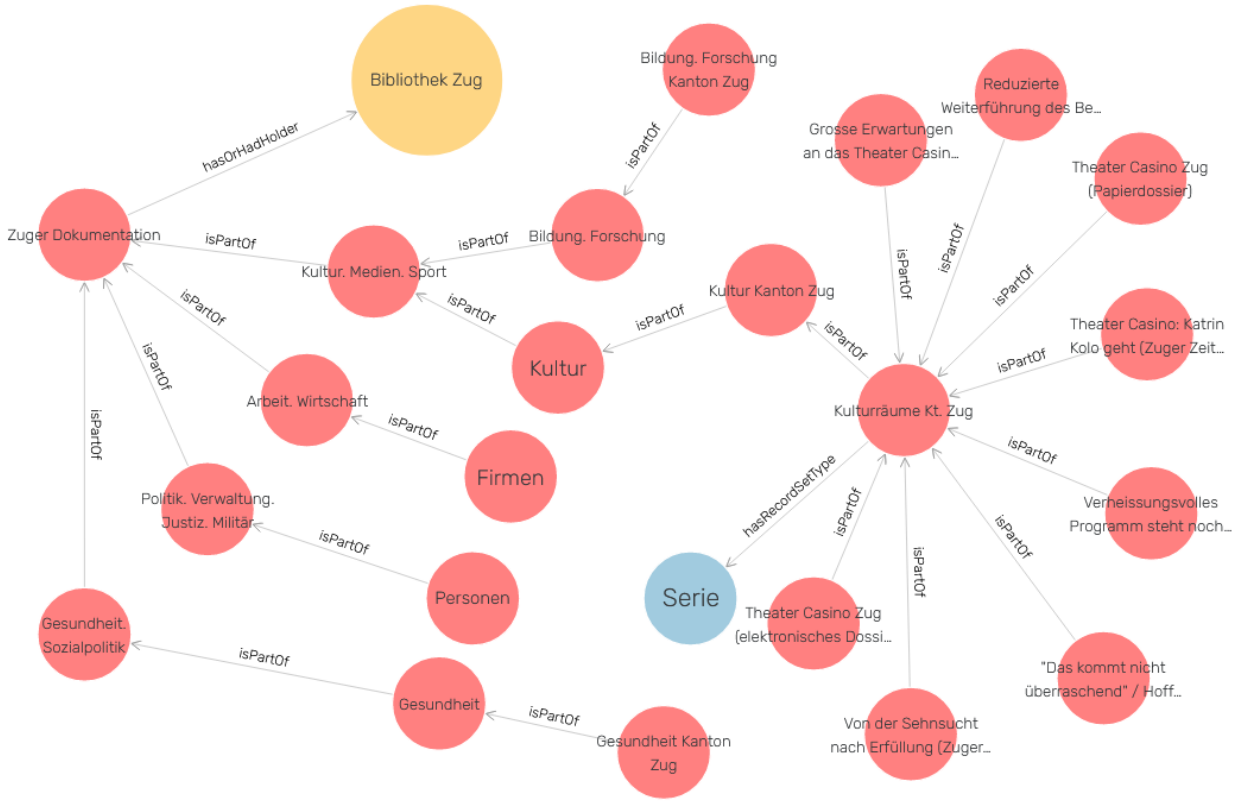


Figure 10: Knowledge graph depicting the archival hierarchy of a sub-set of the "Zuger Dokumentation" of Bibliothek Zug

The knowledge graph above illustrates using some example records of the fonds "Zuger Dokumentation" of BibZug. The levels of arrangements are visible quite well; below the main fonds are some sub-fonds, whit sub-sub-fonds, ultimately ending at the depth of the series "Kulturräume Kt. Zug", which itself holds a couple of files.

4.3.4 Mapping library holdings – a BibZug example

As laid out in the chapter on extracting data, the library catalogue software of BibZug only allows for exports of recorded units at level "*item*" of the RDA-framework, thus effectively exporting data pertaining to physical holdings, i.e. physical Instantiations. The export offers 2 types of identifier attributes, "*Regal*" equivalent to a library signature, and "*Medien-nummer*" equivalent to a unique barcode on the physical holding, and neither are complete

for all rows. For the sake of practicality, "*Mediennummer*" is treated similarly to the datatype *localSoftwareID* as used in the previous exports. The blank attribute values are filled with a unique 12-digit identifier, and then used as a basis for the IRI of the RecordSet-entity and the Instantiation-entity in the same way as in previous exports. The "Regal" attribute is mapped as identifier with the datatype "*localHolderRefCode*", and the RecordSets are linked with their respective Instantiations through *rico:RecordResourceToInstantiationRelation*. In addition, both entities are assigned the title as *rdfs:label* and name (*rico:hasOrHadName*).

BibZug Quria export		*subject*	*predicate*	*object*
Mediennummer		<E03_"id">	*rico:Identifier*	"id"^^<localSofttwareIDe>
		<E03_"id">	*rdfs:type*	rico:RecordSet
		<E06_"id"_001>	*rdfs:type*	rico:Instantiation
Regal		<E06_"id"_001>	*rico:Identifier*	"refCode"^^<localHolderRefCode>
		<E03_"id">	*rico:RecordResourceToInstantiationRelation*	<E06_"id"_001>
Titel (split)		<E03_"id">	*rico:hasOrHadName* *rdfs:label*	"title"
		<E06_"id"_001>	*rico:hasOrHadName* *rdfs:label*	"Instantiation of" & "title"

The library catalogue exports the author of the holding, which is assigned an IRI as well depending on the type of the author the class *rico:Person* or *rico:CorporateBody*. In addition, as mentioned above, the catalogue export states if the holding belongs to BibZug or StAZG, which is also mapped as shown below.

BibZug Quria export		*subject*	*predicate*	*object*
Geistiger Schöpfer		<creatorID>	*rdfs:type*	rico:Person / CorporateBody
		<E03_"id">	*rico:hasCreator*	<creatorID>
		<creatorID>	*rico:hasOrHadName* *rdf:label*	"creator"
Bibliothek		<holder_iri>	*rdfs:type*	rico:CorporateBody
		<E06_"id"_001>	*rico:hasOrHadHolder*	<holder_iri>
		<E03_"id">	*rico:hasOrHadHolder*	<holder_iri>
		<holder_iri>	*rico:hasOrHadName* *rdfs:label*	"library"

The attribute "Publikationsvermerk" is again one of these attributes like previously "scopeAndContent", that holds a multitude of different information. In this case, the attribute holds information regarding the place of the publication (that conforms to the entity RiC-E22 Place), as well as the date of publication and the publishing actor. The attribute is thus split and mapped accordingly:

BibZug Quria export		subject	predicate	object
Publikationsvermerk (split)		<E03_"id">	*rico:hasOrHadPublisher*	<publisherID>
		<publisherID>	*rdfs:type*	rico:Person / Corporate-Body
		<publisherID>	*rdf:label* *rico:hasOrHadName*	"publisher"
		<E03_"id">	*rico:creationDate*	"publishing-Date"^^xsd:date
		<E03_"id">	*rico:isAssociatedWithPlace*	<publishingPlaceID>
		<publishing-PlaceID>	*rdfs:type*	rico:Place
		<publishing-PlaceID>	*rdf:label* *rico:hasOrHadName*	"publishingPlace"

And finally, the last 3 RDA-attributes are mapped to their respective equivalents in I-SAD(G), thus concluding the mapping of the export of the library holding of BibZug.

BibZug Quria export		subject	predicate	object
Medientyp		<E03_"id">	*rico:contentType*	<contentType>
Sprache		<E03_"id">	*rico:hasOrHadLanguage*	<language>
Manifestationsart		<E06_"id"_001>	*rico:hasRepresentationType*	<representationType>

4.4 Making the graph database work

The mappings described in the previous chapter output for every executed mapping a separate TTL-file. This file holds the triple-statements created previously, but the data isn't linked or combined yet, and it also doesn't hold any RiC rules. Since the mappings were executed with pre-defined ontologies and IRI-sources, the importing of the data to a graph DB should link the data automatically. As an example, in all datasets we have entities, that have the class (*rdfs:type*) *rico:Person* or *rico:CorporateBody*, whether as the holder of an entity, creator, publisher, or something else. If the mapping was done correctly, every one of these entities should point to the IRI of *rico:Person* or *rico:CorporateBody*, and thus after import automatically be linked through that IRI. The matching rules that aren't contained in the data would in this case be, that the entity RiC-E08 Person is an RiC-E07 Agent, and that the entity Corporate Body belongs to the entity RiC-E09 Group, which is a sub-entity of RiC-E07 Agent itself, and thus ultimately linking all entities Person and Corporate Body together under the entity Agent as well. These missing rules will be addressed later on. Before the data can be linked though, it has to be first loaded into a DB.

4.4.1 Creation of a Repository

The software used to run the RDF-database is GraphDB from the company Ontotext. For this thesis, I use a local installation of GraphDB Free, which doesn't require working with console, and offers a graphic user interface (GUI). After installing the distribution, I can access the locally running DB in-browser on http://localhost:7200/ through a user-friendly GUI.

In order to load the mapping-outputs, a repository has to be created first. GraphDB offers a range of various types of repositories, out of which for this purpose a "GraphDB Free" repository is being chosen, which offers sufficient functionalities like storing data, running and answering queries, and executing data updates[67].

While creating the repository, technical parameters like a Ruleset, consistency checks, indexing entity ID size etc. can be adjusted. In this case, the repository is created with the standard out of the box settings. After creation, various settings can be customized, out of which the following ones were adjusted:

Adding namespaces

- The standard installation of GraphDB has the following namespaces (see 3.3) enabled:
  - gn: / owl: / path: / rdf: / rdfs: / wgs: / xsd:
- Besides these, additional namespaces are manually added:

Prefix	Path
dc11:	https://purl.org/dc/elements/1.1/
dcterms:	https://purl.org/dc/terms/
rico:	https://www.ica.org/standards/RiC/ontology#
ric-rst:	https://www.ica.org/standards/RiC/vocabularies/recordSetTypes# *
ric-dft:	https://www.ica.org/standards/RiC/vocabularies/documentaryFormTypes# *
wd:	https://www.wikidata.org/wiki/

[67] Other types of repositories offered are e.g. "Ontop Virtual SPARQL" repository, that translates directly data stored in an SQL DB to a virtual SPARQL endpoint, or "FedX Virtual SPARQ" repository, that provides federation of multiple SPARQL endpoints to a single endpoint.

*These ressources were at the date of submission of this Master's thesis accessible, but have been in the meantime removed by the ICA

Enabling Autocomplete Index

- As a useful assistance, the "Autocomplete index" was enabled for the repository, This functionality helps with the automatic completion of IRIs in the SPARQL editor and the GUI, and thus facilitates manual correction and expansion of the DB.

With these settings, the repository is successfully created and ready for data-importing.

4.4.2 Loading data

GraphDB allows for import of data from various either external or internal sources. One could access an external SPARQL endpoint, or import RDF data from an external URL, which is both not necessary in this case. For this thesis, uploading RDF-files directly into the DB and importing RDF text snippets is sufficient. With the mappings transferred above, a total of 6 RDF files were created, which then can be loaded simultaneously into the GraphDB via TTL-import. The import then successfully loads a total of 7'439 statements into the repository, and the mapping and transformation has been completed (accessible under https://github.com/specjo/ISADGtoRiC/tree/main/442_full_import)

So far still no additional meaningful information has been added to the data-set (besides replacing literal values with IRIs), but a mere mapping has taken place.

An example is the entity hierarchy. GraphDB allows to inspect class/entity hierarchy and class relations of a given repository. When doing, once can see that entities exist, but they do not carry any hierarchy according to RiC-CM.

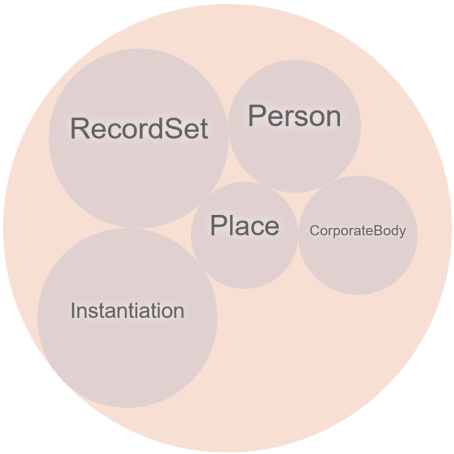


Figure 11: Class hierarchy of mapped entities without implemented RiC-CM hierarchy

There are 3 ways to add additional statements to the repository. One could (1) load additional RDF-files or RDF-snippets to the repository, (2) work with the GraphDB-GUI to manually add triples, or (3) use SPARQL to access the data directly. What Structured Query

Language (SQL) is for RDBs, SPARQL is ifor RDF. It is used to retrieve tabular data and to manipulate it (delete, insert and update triples). And while the knowledge graphs shown so far all have been created through exploration of data through the GUI (i.e. expanding nodes and the corresponding edges by clicking), SPARQL allows to create using queries advanced graphs. Such advanced graphs rely on querying and filtering of data to produce meaningful knowledge graphs, as it wouldn't be possible through simple data-exploration through the GUI.

A simple example of a SPARQL-query can be shown by means of adding the missing RiC-CM hierarchy of classes/entities described above, and would look like this:

```
PREFIX rico: <https://www.ica.org/standards/RiC/ontology#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

INSERT DATA
    {
   rico:Person rdfs:subClassOf rico:Agent.
   rico:CorporateBody rdfs:subClassOf rico:Group.
   rico:Group rdfs:subClassOf rico:Agent.
      }
```

Figure 12: SPARQL query for adding class hierarchies of RiC-E08 / E09 / E11

After running that SPARQL query the subClass-hierarchy is added to the repository, and the inspection of the class hierarchy now shows, that the sub-classes defined by RiC-O have successfully been inserted.

Figure 13: Class hierarchy of mapped entities after implementing RiC-CM hierarchy for Agent-entities

Examining now entities of all classes (limited for 50 entities / class) as defined in the data model in 3.2, we see in the knowledge graph for the full data-set, that the entities and the hierarchy have successfully been loaded into the DB, and the heavy lifting of transformation is completed:

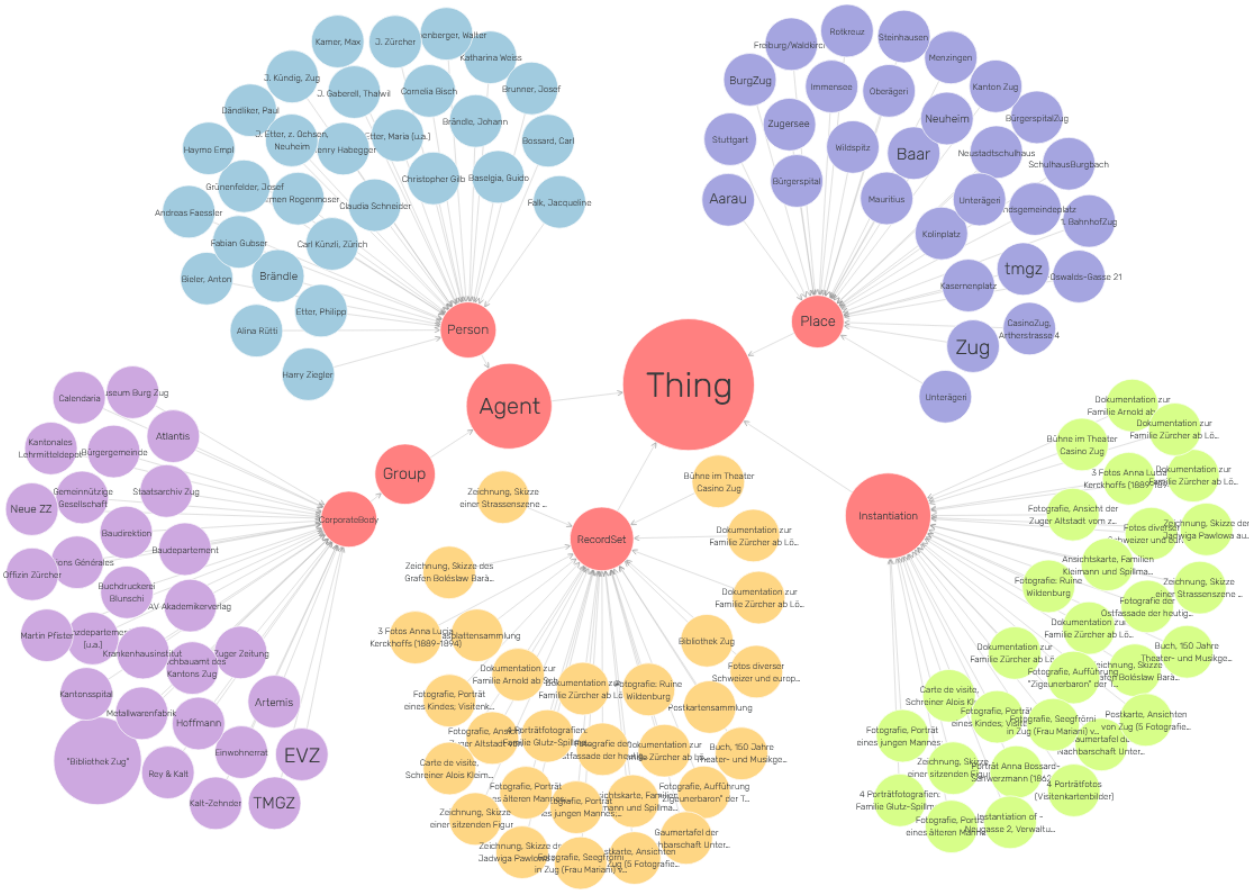


Figure 14: Knowledge graph of entities belonging to a class according to the Data model RiC "light"

With the example query above I've added only one part of the sub-classes of Agent, while the full RiC-Ontology isn't loaded in the dataset yet. Adding the full RiC-O to the data-set is quite an easy task, since GraphDB allows the import of RDF-files from a link, and the RiC-O is accessible under https://www.ica.org/standards/RiC/RiC-O_v0-2.rdf. Loading the ontology adds besides class hierarchies as shown above also all classes, domains, ranges, properties, inverse relations, respective inferred statements and such. An example might be the relation of a Record-Entity to an Agent, that is the holder of that Entity. With the executed mapping I've only added implicitly the "one-way" relation of `<Record-entity> rico:hasOrHadHolder <Agent-entity>`. Importing the full ontology then adds as inferred statement the logical inverse `<Agent-entity> rico:isOrWasHolderOf <Record-entity>`, expanding the "one-way" triple to a "two-way" statement.

4.4.3 Cleaning data

Clean, consistent and correct data is at the core of every useful catalogue or database. As almost every DB has erroneous data of some kind, I dare to claim that every GLAM-institution could improve the content of their own DBs, even if it was only for typos. Data cleaning is in data science a broad issue, that could easily fill an own thesis by itself. Thus,

I will here only briefly outline why and how data cleaning is important for GLAM-institutions. The sources of erroneous data are manyfold, and various methods exist to clean it. Errors might stem from faulty analogous description and metadata, faulty migration from old to new DBs, wrong data-formats, structural mistakes, human errors of various kinds, and many more that certainly every practitioner has encountered here and there while working with catalogues.

Data cleaning then is for most people quite uninteresting or dull, but is nonetheless essential for quality data and quality catalogues. What constitutes quality data has multiple factors like usability, reliability, consistency, completeness, and many others[68]. Since there are many sorts of errors, and many goals and ways of fixing them, at the beginning of a data cleaning process is the inspection and the detection of incorrect and inconsistent data. After identification of the problem, appropriate data cleansing techniques have to be defined and applied. They can take the form of conversion of datatypes, removal of irrelevant data or duplicate entries, normalization, standardization, correction of typos, adding missing values etc.[69].

There are various tools to support such data cleaning processes, of which one is the before-mentioned OntoRefine. The preferred way would be, that data would be cleaned before being loaded into a new database. Since this isn't the focus of this thesis, I will only describe one short illustrative data cleaning measure. In the case of GraphDB, one could either reload a data-set or a data-snippet and overwrite the existing data, use the GUI to manually fix single triple-statements, or work with SPARQL queries. Since it's difficult to oversee changes with the re-loading of data, and manual correction through the GUI is quite self-explanatory, I will use the SPARL query option.

After importing data from 4 various holders of Records or Items, I might want to double-check which entities do now exist in the combined dataset, that are linked to a Record or Instantiation thereof. I can use a simple SPARQL query, to check for any entity/subject *(?s)* that has a holder (`rico:hasOrHadHolder`), and define that holder in the so-called "query variable" `?HolderIRI`. The question mark (?) prefixes a query variable, that is created and used only in the context of this query.

[68] Cai, L. and Zhu, Y. (2015): The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, p.4

[69] Elgabry, Omar (2019): The Ultimate Guide to Data Cleaning

```
PREFIX rico: <https://www.ica.org/standards/RiC/ontology#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?HolderIRI ?HolderName WHERE
{
    ?s rico:hasOrHadHolder ?HolderIRI.
    ?HolderIRI rdfs:label ?HolderName.
}

limit 100
```

Figure 15: SPARQL query for retrieval of RiC-O entities, that are a holder

I might then not only be interested in the IRI defined above only, but also the label of each holder, which is queried with the second triple-statement and stored under the variable *?HolderName*. With the initial command SELECT DISTINCT, I query only single occurrence of each holder, and not every occurrence on all entities held. The query above then outputs the following result:

	Holder	HolderLabel
1	wd:Q27490186	"Museum Burg Zug"
2	wd:Q27490192	"Stadtarchiv Zug"
3	wd:Q856475	"Bibliothek Zug"
4	wd:Q856475	"Stadtbibliothek"
5	wd:Q856475	"Bibliothek"
6	wd:Q14848231	"Staatsarchiv Zug"

Apparently, the labels of the entities MBZ, SAZG and StAZG are distinct and good as they are, but BibZug is labelled with 3 different labels. One might only want one single distinct label in a knowledge graph instead of many (e.g. "Bibliothek Zug", the current official name), and would therefore delete 2 other labels with removal of the respective triple-statements with the query below:

```
PREFIX wd: <https://www.wikidata.org/wiki/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

DELETE WHERE
{
    wd:Q856475 rdfs:label "Bibliothek".
    wd:Q856475 rdfs:label "Stadtbibliothek".
}
```

Figure 16: SPARQL query to delete 2 existing triple statements

Running the previously described query for retrieval of holders and their labels again, one now receives the following results:

	Holder	HolderLabel
1	wd:Q27490186	"Museum Burg Zug"
2	wd:Q27490192	"Stadtarchiv Zug"
3	wd:Q856475	"Bibliothek Zug"
4	wd:Q14848231	"Staatsarchiv Zug"

This very brief example shows the simplicity and powerfulness of RDF alike. With the removal of a single triple-statement, all Records that had the holder *wd:Q856475* and the "wrong" label, were updated right away. Comparing with RDBs, where one would have to update the entry of a holder in every single attribute value field of every respective record in the whole DB, this illustrates one aspect of the flexibility of graph DBs quite well. I will not do a thorough examination and cleaning of data for the exports, since it is a time-consuming task that isn't the topic of this thesis, and a task a data scientist is better equipped to execute. While working with the data in the next chapters, minor errors might occur, which I will nonetheless correct manually.

5 Step 2: RiC use cases

In this chapter I will based on the research topics defined at the beginning, explore simultaneously how existing data could be enriched, and how the enriched data could offer added value. Enriching of the data could have happened during mapping or import, as was done for example with creating IRI's for entities out of literal values. Other possibilities are to use different data from source systems, matching with pre-defined vocabularies or IRI sources during import, machine-aided cleaning and normalization of data or automatic detection from literals such as the infamous "scope and content" field, and many others. For this thesis, I will add meaningful information to the existing data post-import through SPARQL, to illustrate various use-cases, in what direction mass-enriching could go, and to give directly insight into what is happening with data instead of just showing a GUI. The following sub-chapters are written with the research questions 3. and 4. in mind, and offer possible answers.

5.1 Enriching: RiC-E08 Person

The first set of research topics defined at the beginning were the following 4 Persons:

Category	Research topic	RiC class / entity
Person	“Keiser, Dagobert” jun. (architect, 1879 - 1959)	Person (E08)
Person	“Keiser, Dagobert” sen. (architect, 1847 – 1906)	Person (E08)
Person	“Weiss, Katharina” (photographer, 1834 - 1911)	Person (E08)
Person	“Etter, Philipp” (federal chancellor, 1891 – 1977)	Person (E08)

Let's take Keiser Dagobert senior[70] and junior[71] as first example. The latter is the son of the former, and since both were renowned architects in Zug, their identical name can lead to confusion, and improving data is surely useful. When querying for any "Keiser, Dagobert", I only get a single person entity as a result (junior), that was initially imported as creator of a RecordSet, and holds a IRI based on an entry in the online HLS[72]. Since I defined in 3.3 that Wikidata IRIs have priority, I want to change that, and also add additional attributes to that Person-entity, to make it clearer distinguishable from his father.

[70] Morosoli, Renato (2007): Dagobert Keiser (Version vom 18.05.2007)

[71] Morosoli, Renato (2014): Dagobert Keiser (Version vom 26.11.2014)

[72] IRI taken from HLS <https://hls-dhs-dss.ch/de/articles/042168/2014-11-26/>

Working with data will be done through SPARQL queries, and the queries described below are accessible under:

https://github.com/specjo/ISADGtoRiC/tree/main/51_enriching_Person.

In order to do that, working with SPARQL replacing/updating an IRI happens through inserting new data and deleting the old instead of "overwriting" the existing data. In this case, I would therefore delete the triple where its IRI Is the object of the *rico:hasCreator* predicate, and also delete all triples, where the IRI is subject of a statement (see below). In order to "update" the IRI, I would then insert new triples, that are almost the same triple-statements as the deleted ones, except for the changed IRI that I want to replace, and a small adjustment in the label (added Suffix (jun.)):

```
PREFIX rico:  <https://www.ica.org/standards/RiC/ontology#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wd:    <https://www.wikidata.org/wiki/>

DELETE DATA
    {
<http://example.com/base/E03_105130916099> rico:hasCreator <https://hls-dhs-dss.ch/de/articles/042168/2014-11-26/>.
<https://hls-dhs-dss.ch/de/articles/042168/2014-11-26/> a rico:Person;
                                                        rdfs:label "Keiser, Dagobert";
                                                        rico:hasOrHadName "Keiser, Dagobert" .  }

  INSERT DATA
    {
<http://example.com/base/E03_105130916099> rico:hasCreator wd:Q78065037.

wd:Q78065037        a  rico:Person;
                    rdfs:label "Keiser, Dagobert (jun.)";
                    rico:hasOrHadName "Keiser, Dagobert". }
```

Figure 17: SPARQL query replacing (deleting and inserting) the IRI of a Person-entity

The IRI for the Person-entity is now replaced, but I want to add additional meaningful information to that IRI. In this case, I know various things about Keiser D. junior, either from the article in HLS, the entry in Wikidata, or even information that an archive holds in-house. The statements from Wikidata could e.g. be fully queried into the local DB, but I'd rather add various datapoints manually for illustrative purposes. What might be now useful to know about the Person-entity, is a *birthDate* and *deathDate,* and a triple stating the main area of activity of the Person (*rico:isAssociatedWithPlace wd:Q68144; #City of Zug*). I can also add buildings that the Person created with the predicate *rico:isCreatorOf* , and as object IRIs addresses of that particular building in the Linked Data Portal of the Swiss government[73], define the occupation with the predicate *rico:hasOrHadOccupationOfType*.

[73] For example, https://geo.ld.admin.ch/location/address/100293767

The occupation-type *wd:Q42973* as well as the buildings that were added as object with the IRI from the *geo.ld.admin.ch* portal now exist in the DB only as IRI's, which is why they additional were given a human-readable label (with *rdfs:label*), and the buildings were defined as entity of *rdf:type rico:Place*. As the WD IRIs would allow to load further triples from WD into the local database, so would the IRIs of geo.ld.admin.ch also allow to load any existing statements regarding e.g. postal code, building category or GPS coordinates into the local database as well. All these information exist, and a GLAM-institution would be able to tap into existing Linked Data repositories in order to enrich its own local databases quite easily, and especially without having to enrich its own catalogue manually by itself.

```
    INSERT DATA
    {
wd:Q78065037        rico:birthDate "1879/10/12"^^xsd:date;
                    rico:deathDate "1959/05/22"^^xsd:date;
                    rico:isAssociatedWithPlace wd:Q68144; #City of Zug
                    rico:isCreatorOf <https://geo.ld.admin.ch/location/address/100293767>; #Theater Casino Zug
                    rico:isCreatorOf <https://geo.ld.admin.ch/location/address/100306366>; #Kantonals Verwaltungsgebäude
                    rico:isCreatorOf <https://geo.ld.admin.ch/location/address/100307000>; #Neustadtschulhaus Zug
                    rico:isCreatorOf <https://geo.ld.admin.ch/location/address/101775735>; #Sanatorium Adelheid
                    rico:hasOrHadOccupationOfType wd:Q42973;
                    rico:isChildOf wd:Q78065032. #Keiser, Dagobert (sen.)

wd:Q42973 rdfs:label "Architect".
<https://geo.ld.admin.ch/location/address/100293767>  a rico:Place;
                                                      rdfs:label "Theater Casino Zug".
<https://geo.ld.admin.ch/location/address/100306366>  a rico:Place;
                                                      rdfs:label "Kantonales Verwaltungsgebäude".
<https://geo.ld.admin.ch/location/address/101775735> a rico:Place;
                                                      rdfs:label "Sanatorium Adelheid".   }
```

Figure 18: Enriching the person-entity wd:Q78065037 with meaningful information

Now the last newly added triple is `wd:Q78065037 rico:isChilfOf wd:Q78065032`, where Keiser Dagobert junior as subject is defined as child of Keiser Dagobert senior (respectively his IRI). One might add similar information as for Keiser sen., as was added for jun., e.g. birthDates and buildings that were built by Keiser senior[74]. The final adding of additional information through a SPARQL query is shown below:

[74] Strictly speaking isn't the creator of the building "Hotel Ochsen Zug", but only renovated it. Since rico: doesn't offer a predicate to describe a renovating role of an architect, I used *rico:isCreatorOf* for the sake of argument

```
INSERT DATA
      {
wd:Q78065032        a  rico:Person;
                    rdfs:label "Keiser, Dagobert (sen.)";
                    rico:hasOrHadName "Keiser, Dagobert";
                    rico:birthDate "1847/03/22"^^xsd:date;
                    rico:deathDate "1906/02/09"^^xsd:date;
                    rico:isAssociatedWithPlace wd:Q68144; #City of Zug
                    rico:isCreatorOf <https://geo.ld.admin.ch/location/address/100306178>; #Hotel Ochsen Zug
                    rico:isCreatorOf <https://geo.ld.admin.ch/location/address/100002474>; #Turnhalle Burgbach
                    rico:isCreatorOf <https://geo.ld.admin.ch/location/address/100306344>;#Zeughaus Zug
                    rico:hasOrHadOccupationOfType wd:Q42973;
                    rico:hasOrHadOccupationOfType wd:Q11518629.

wd:Q11518629 rdfs:label "Timber merchant".
<https://geo.ld.admin.ch/location/address/100306178>  a rico:Place;
                                                      rdfs:label "Hotel Ochsen Zug".
<https://geo.ld.admin.ch/location/address/100002474>  a rico:Place;
                                                      rdfs:label "Turnhalle Burgbach".
<https://geo.ld.admin.ch/location/address/100306344>  a rico:Place;
                                                      rdfs:label "Zeughaus Zug". }
```

Figure 20: Enriching the person-entity wd:Q78065032 with meaningful information

With the SPARQL queries described below, only some additional information was added to the entities Person. Now depending on the information available in existing repositories, or to be added manually, as well as depending on the needs of a certain institution, various additional information and linked data might be of use. For now, for both jun. and sen. Keiser's, we can create the following knowledge graph:

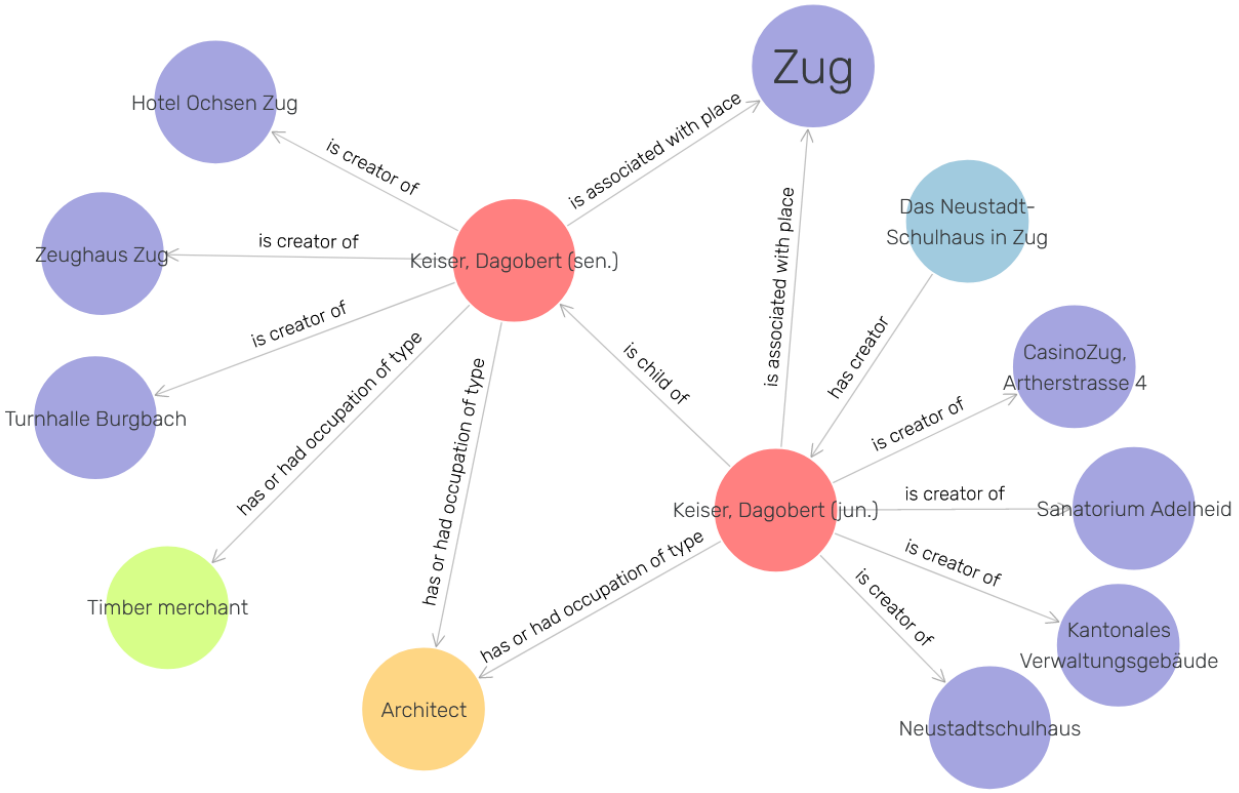


Figure 19: Knowledge graph depicting Keiser Dagobert sen. and jun. with its related holdings and additional information

An archive of a legislative body now might be interested to use for its Person-entities predicates such as *rico:hasOrHadPosition, rico:isOrWasMemberOf* or *rico:hasOrHadLeader*. For a library other predicates might be more interesting to use, such as *rico:isCreatorOf*. Depending on the context, each institution would have to define its own data-model with the datapoints that should be at the focus of the (archival) description, or could offer any other additional value.

Now for the other two Person-entities of interest, Katharina Weiss and Philipp Etter, similar data as described above based on their HLS and WD entries, was loaded via SPARQL query into the database. The respective queries are accessible as .ttl: https://github.com/specjo/ISADGtoRiC/tree/main/51_enriching.

The Person-entities are now enriched with data about the Persons themselves. While analysing the data-exports and the mapping, it became evident that in the institutions various holdings exist, which weren't only created/published/written by these Persons, but have them as subject. Having a DB based on Linked Data and RiC, one might not only want to link a certain Record with its creator, but also link to something that describes what the content of that record is about, e.g. a person, a place or any other topic.

In order to enrich the data-set with such information as well, I proceeded as follows. The relevant holdings, that have one of the 4 research persons as topic, were identified through a SPARQL query. With the following example, I've queried all existing Record Sets (line 9) and outputted their name (line 10), and then filtered their names to contain the word "Weiss" (line 11).

```
6   SELECT * WHERE
7
8       {
9   ?r rdf:type rico:RecordSet.
10  ?r rico:hasOrHadName ?rName.
11     FILTER contains(?rName,"Weiss")
12  ?r rico:hasCreator ?c.
13  ?c rico:hasOrHadName ?cName.
14      }
```

Figure 21: SPARQL query to extract all Record Sets with a certain string ("Weiss") in the Name-attribute

In addition, I've queried the IRI of the entity that is the creator of that Record Set (line 12), and the corresponding name (line 13). The result of the query then looks as follows:

	r	rName	c	cName
1	http://example.com/base/E03_185847_0356	"Erstes Bahnhofsgebäude von Zug, Bild von Katharina Weiss"@de	wd:Q98931764	"Katharina Weiss"
2	http://example.com/base/E03_400187400099	"Zugs erste Fotografin : Jungfer Katharina Weiss (1834-1911)"	http://example.org/base/Steiner,%20Herman	"Steiner, Hermann"
3	http://example.com/base/E03_14942370452	"Zugs erste Photographistin : Katharina Weiss, 1834-1911"	http://example.org/base/Staub,%20Ignaz	"Staub, Ignaz"

The RecordSet in Row 1 then is according to the query a Photograph created by "Katharina Weiss" of the first train station of the city of Zug, and only the results in Rows 2 and 3 were created by somebody else and have as subject (according to the name of the Record Set) the Person "Katharina Weiss".

```
INSERT DATA
    {
 wd:Q98931764    rico:isOrWasSubjectOf <http://example.com/base/E03_400187400099>;
                 rico:isOrWasSubjectOf <http://example.com/base/E03_14942370452>.
    }
```

FIgure 22: SPARQL query linking the person "Katharina Weiss" to RecordSets, of which she is subject of

As a result, the IRIs of these two RecordSets were connected through a SPARQL query with the predicate *rico:hasOrHadSubject* to the person Katharina Weiss:

The same procedure of extracting relevant RecordSets[75] and linking them to Persons was done as well as for both "Keiser, Dagobert" sen. and jun. as well as for "Etter, Philipp" [76].

SPARQL now doesn't only allow for simple SELECT queries to extract data, or simple DELETE/INSERT queries to update data. The knowledge graphs shown so far all were created through manual exploration of the data (i.e. double-clicking nodes expands the node to connected ones, and from there one can further develop the knowledge graph by "opening" the following nodes).

[75] Access: https://github.com/specjo/ISADGtoRiC/blob/main/51_enriching/51_SPARQL_select_RecordSets_Persons.ttl

[76] Access: https://github.com/specjo/ISADGtoRiC/blob/main/51_enriching/51_SPARQL_insert_RecordSets_Persons.ttl

SPARQL also allows to create advanced knowledge graphs, through a targeted query and the CONSTRUCT command. As shown in the example below, the commands in the

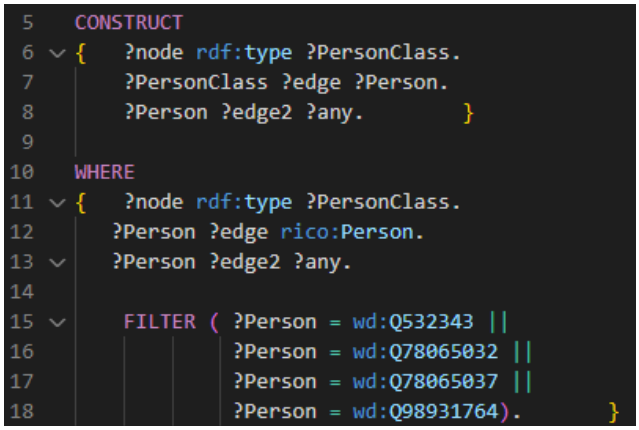


Figure 23: Example of a CONSTRUCT WHERE SPARQL query , creating an advanced knowledge graph

angled brackets {} after WHERE describe the retrieval of data from the DB, and the commands in the angled brackets after CONSTRUCT describe, how the knowledge graph expands from a starting ?node (in this case ?node is `Katharina Weiss <https://www.wikidata.org/wiki/Q98931764>`).

In Line 11, the variable *`?PersonClass`* is defined as the *`rdf:type`* of the starting ?node (is in this case *`wd:Q98931764`*, the IRI of Katharina Weiss), effectively loading into the *`?PersonClass`* variable the IRI *`rico:Person`*. Line 6 defines, that from the starting node an edge is drawn to that *`?PersonClass`* node. Further, in Line 12 any entity that has the class `rico:Person` is loaded into the variable ?Person, and Lines 15 to 18 filter from all the Persons in the DB only the ones that I am interested in. Line 7 then further expands the knowledge graphs and draws from the Node "`rico:Person`" an edge to all filtered ?Persons. Finally Line 13 loads and Line 8 draws all entities, that are subject of the *`?Persons`* that we filtered, completing the knowledge graph[77]:

[77] Access: https://github.com/specjo/ISADGtoRiC/blob/main/51_enriching/Persons-Records.PNG

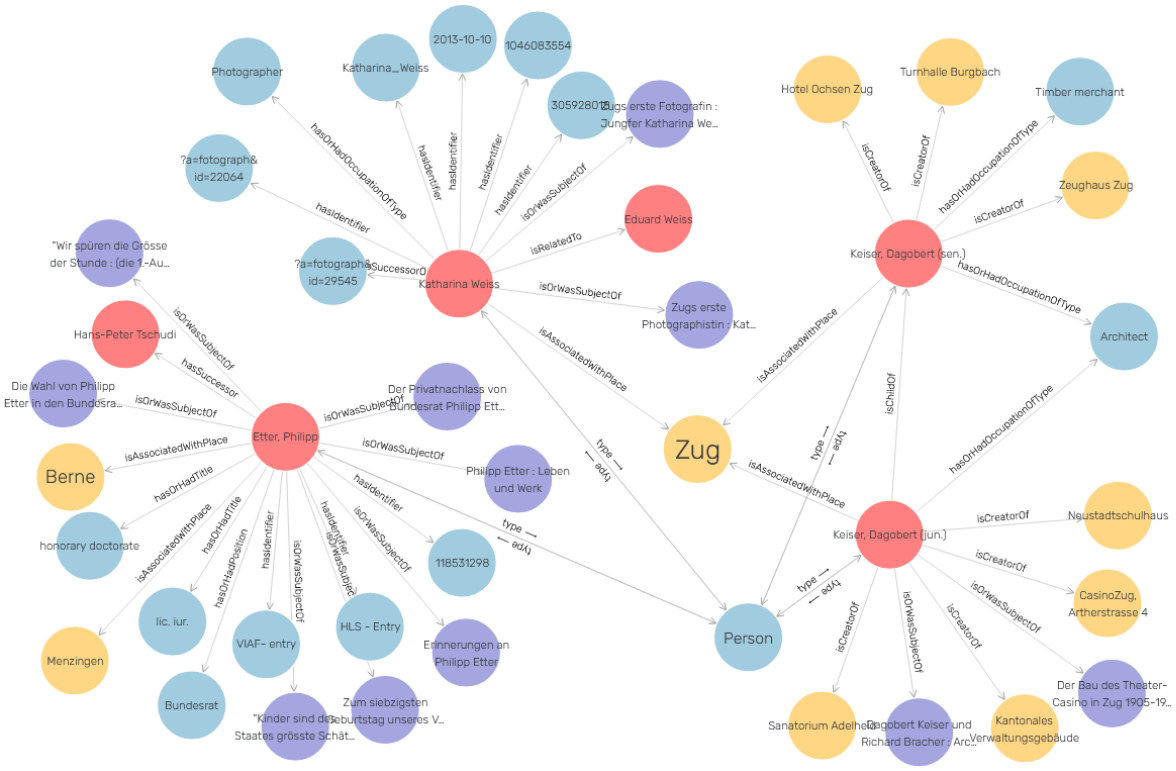


Figure 24: Knowledge graph depicting all nodes related to the 4 Persons defined as research topics

The finalized graph after enriching Person-entities with contextual data now shows various directions in which linked data and data enriching efforts could go. Examples are the linking to all Records created by a Person, and to all Records where that same person is subject of, linking to a relative Person or a predecessor/successor within a position, or the linking to a geographical place with relevance to that Person.

5.2 Enriching: RiC-E22 Place

The initially defined exemplary research topics that classify as RiC-E22 Place were the following ones:

Category	Research topic	RiC class / entity
Building	St. Oswalds-Gasse 21, 6300 Zug	Place (E22)
Building	Schulhaus Neustadt	Place (E22)
Building	Eishalle Herti / Stadion Herti	Place (E22)

Let's look at the first example, "St. Oswalds-Gasse 21, 6300 Zug". This address is quite an interesting example, because that particular building is originally a granary, was later used as barracks, and is nowadays as a library the building where the `rico:CorporateBody` "Bibliothek Zug" is located.

There are various directions of data enriching. An institution holding a photographic collection might want to link the corresponding motives of the holdings to that particular building. Another entity, e.g. an office for the preservation of historic monuments, might be more interested to use the RiC/RDF to depict the history of that particular building and its "interaction with its environment". For now, that building has been mapped as a Place-entity with the IRI coming from the entry in the Swiss Linked Data Service portal[78], it was assigned a name and label, and some Records were linked to the entity during the mapping:

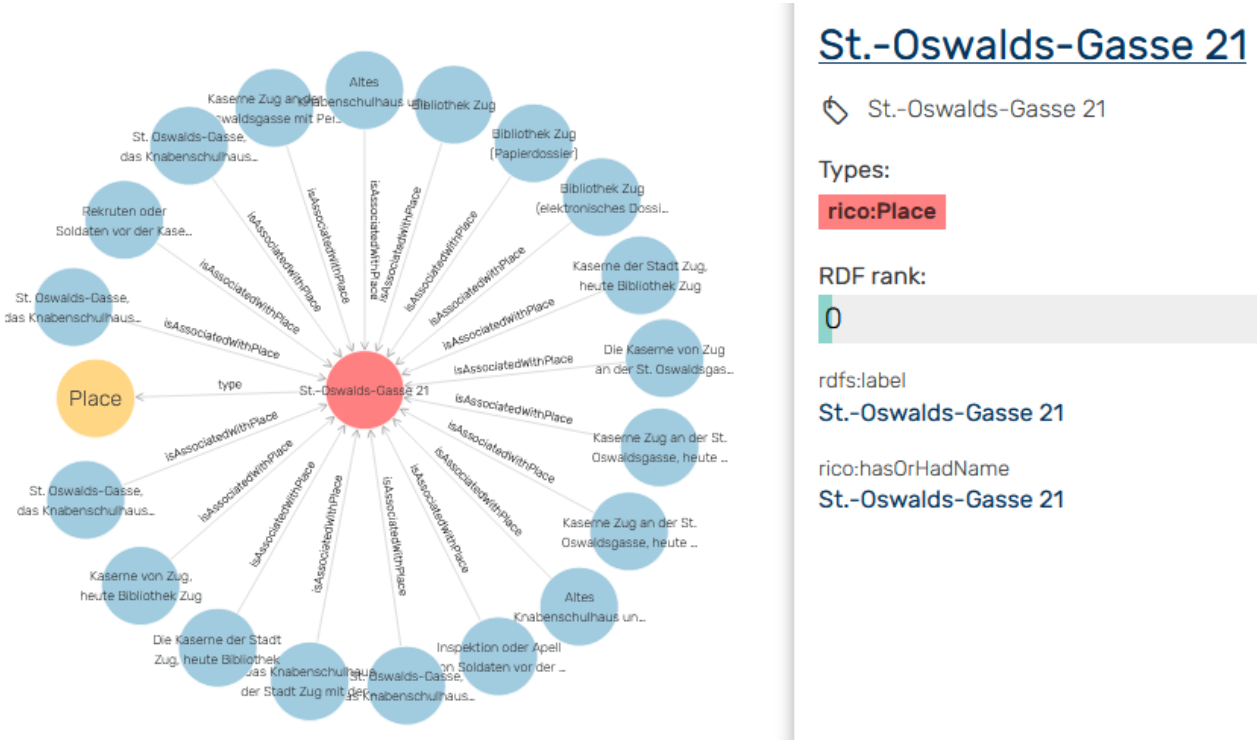


Figure 25: Knowledge graph depicting RecordSets linked to the Place "St.-Oswalds-Gasse 21"

To start with, the "missing" RecordSets that pertain to that particular building, but aren't linked to it yet, would have to be identified and linked. The underlying assumption is, that if a Record mentions that building or an entity located in the building in its name, that Record is also associated with the building. For that purpose, a SPARQL Query was created, assigning any entity that has the type `rico:RecordSet` to the variable *?rs* (Line 6), and

[78] https://geo.ld.admin.ch/location/address/101239133

outputting its name in `?l (Line 7)`. In addition, any of these RecordSets defined as ?rs that are already associated with the Place in question, are excluded from the result with the command FILTER NOT EXISTS and the corresponding triple (Line 8).

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX rico: <https://www.ica.org/standards/RiC/ontology#>
3  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4
5  select * where
6  {   ?rs rdf:type rico:RecordSet;
7          rico:hasOrHadName ?l.
8      FILTER NOT EXISTS {?rs rico:isAssociatedWithPlace <https://geo.ld.admin.ch/location/address/101239133>}.
9      FILTER CONTAINS(?l, "Bibliothek")
10 }
```

Figure 26: SPARQL query for RecordSets associated with the Place IRI of "St-Oswalds-Gasse 21", that aren't linked to it yet

Finally, the variable `?l` containing the names of the Records is filtered with CONTAINS to include the string "Bibliothek" (while in later iterations the filters "Kornhaus" [for the granary] and "Kaserne" [for the barracks] shall be applied as well). An excerpt of the SPARQL query result is then shown below:

	rs	l
3	http://example.com/base/B_E03_81844	St.-Oswalds-Gasse 21, Stadt- und Kantonsbibliothek Zug (Assek.-Nr. 3012a), GS 1226: Umbau Kaserne zur Bibliothek Projekt Hochbau
4	http://example.com/base/E03_146686	Bibliotheken. Bildungshäuser Kt. Zug
5	http://example.com/base/E03_100862737099	Vom Kornhaus über die Kaserne zur neuen Biblio-thek : geglückter Umbau in der Stadt Zug
6	http://example.com/base/E03_105023025099	Katalog der Stadt-Bibliothek von Zug : Sommer 1858
7	http://example.com/base/E03_105038373099	Unsere Bibliothek
8	http://example.com/base/E03_105114727099	Katalog der Stadt-Bibliothek von Zug vom Jahre 1876 : II. Nachtrag

These results illustrate quite well possible pitfalls of data-cleaning or data-enriching, and various queried Records now have various entity-relations. The RecordSets in Rows 3 and 5 in fact have the building at the given IRI/address as their main subject, since they are holdings describing the renovation of that specific building. The holdings in Rows 6 and 8 describe historic catalogues of the library, at a time when BibZug was located in another building. These records thus don't have that Place as subject of its content, but since they are held by BibZug, they are physically located at that address. In addition, since they are the historic catalogues of BibZug, they relate to the BibZug Agent-entity

with a creator-relation[79]. And ultimately, the RecordSet at Row 4 fulfils multiple of the above-described relations; the RecordSet is physically located at the Place-entity, it has as creator the BibZug Agent-entity, and the content of that holding has as subject (among other educational institutions in the canton of Zug) the Agent-entity.

The same SPARQL query as described above with "`FILTER contains`" for "Bibliothek" was run filtering for "Kaserne"[80], and the results were analysed and connected to either the Place- or the CorporateBody-entity according to the same way as described above[81]. The final SQL query filtering for "Kornhaus" didn't deliver any results, because the query excludes any RecordSet associated with the Place-entity, which were apparently previously linked with the Place-entity.

The relevant RecordSets then have been as described above have been linked to its respective entities. From customer perspective, somebody might now be researching about that particular building at St.-Oswalds-Gasse 21, and would want to know where to find resource, and what kind of resources they are. An adequate advanced knowledge graph would depict these information based on a CONSTRUCT WHERE SPARQL Query, where the starting ?node of the knowledge graph is the IRI of the Place-entity/building in question. Lines 16 / 17 query into various variables any element, where the ?node is either subject or object of a triple statement (i.e. …) and construct with the Lines 6 / 7 these triples in a knowledge graph.

```
4   CONSTRUCT
5   {
6       ?node   ?edge   ?s1.
7       ?q      ?edge2  ?node.
8       ?s1     rico:hasOrHadHolder ?h1.
9       ?q      rico:hasOrHadHolder ?h2.
10      ?s1     rico:contentType ?c1.
11      ?q      rico:contentType ?c2.
12  }
13
14  WHERE
15  {
16      ?node   ?edge   ?s1.
17      ?q      ?edge2  ?node.
18      ?s1         rico:hasOrHadHolder ?h1.
19      ?q          rico:hasOrHadHolder ?h2.
20      OPTIONAL{?s1    rico:contentType ?c1}.
21      OPTIONAL{?q     rico:contentType ?c2}.
```

FIgure 27: SPARQL querying from the Research Topic to RecordSets, and its holders and content types

[79] In this case, BibZug wasn't catalogued as creator of these catalogues in the original export, which is a catalogue error

[80] Access: https://github.com/specjo/ISADGtoRiC/blob/main/52_enriching_Place/52_SPARQL_query_Kaserne.ttl

[81] Access: https://github.com/specjo/ISADGtoRiC/blob/main/52_enriching_Place/52_SPARQL_insert_Kaserne.ttl

From the entities linked to/from our original *?node*, Lines 18 / 19 query any holder ?h1 or ?h2, and draw these connections in Lines 8 / 9. And ultimately Lines 20 / 21 query, where applicable, the content type of the RecordSets, and draw them with Lines 10 / 11. The final advanced knowledge graph[82] then depicts these interconnections quite illustratively:

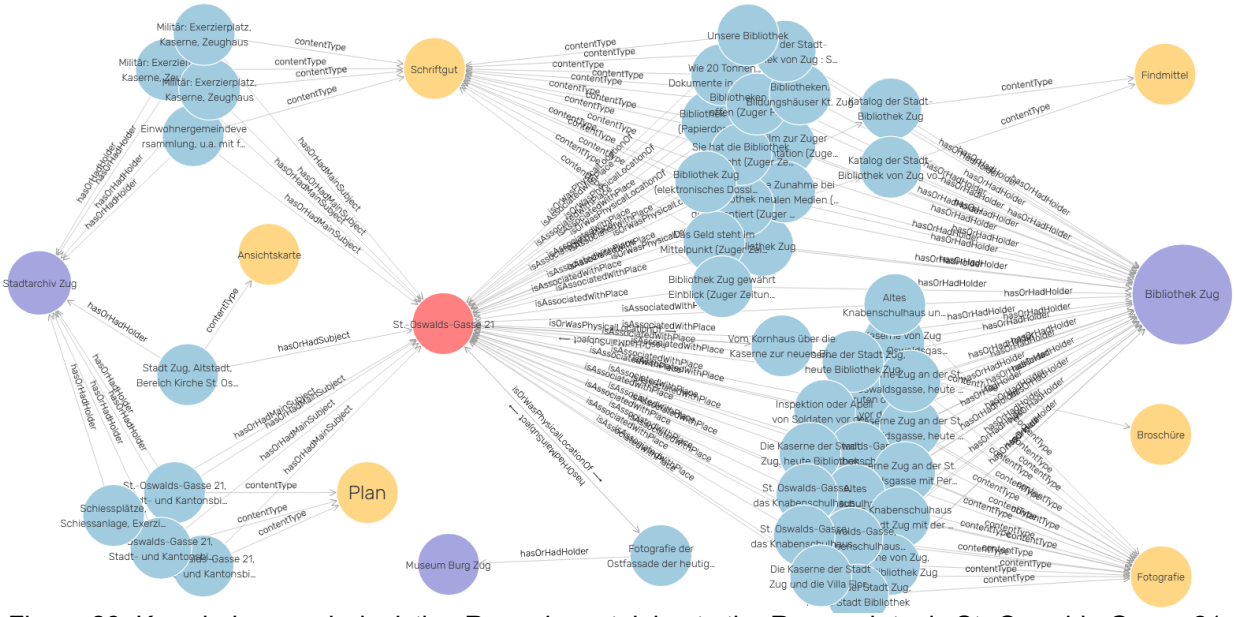


Figure 28: Knowledge graph depicting Records pertaining to the Research topic St.-Oswalds-Gasse 21

The red node is our starting node and the blue ones are the RecordSets that are associated with the original node, or have it as a subject. These RecordSets have one of the purple nodes as a holder, where BibZug holds the most relevant Records, and MBZ holds only a single one. And finally, the yellow nodes depict where a customer can find what kinds of content in the RecordSets, from Photographs [Fotografie] over construction plans [Plan] to classic written materials [Schriftgut] and others.

The originally taken assumption, that if the term "Bibliothek" is present in the name of a RecordSet, that entity must also relate to the Place (i.e. building) the "Bibliothek" is in, has been proven wrong. This brief example illustrates quite well, that mapping and data enriching can has its tricky sides. Such a simple assumption that on first glance might be flawless can quickly lead to a wrong mapping and/or enriching of data. The example also illustrates quite well, that in order to map and/or enrich data, the existing description/data has to be well understood. In the case of GLAM institutions, such work might be done on the technical side by a computer scientist. But in order to make also sense of the content of the data, cooperation between e.g. a local historian and a computer scientist alike might be most fruitful.

[82] Access: https://github.com/specjo/ISADGtoRiC/blob/main/52_enriching_Place/SOG21_RS_H_type.PNG

5.3 Enriching: RiC-E11 Corporate Body

An interesting use case for the entity RiC-E11 CorporateBody is the initially as research topic defined "Crypto AG". To give a quick overview, "Crypto AG" was a company active in cryptographic services, and based in the city of Baar in the canton of Zug. It was founded in 1952[83] and its entry in the Swiss Trade Registry was officially deleted in 2019[84]. The company was then split in 2 successors, an internationally oriented "Crypto International AG" and a domestically oriented "Crypto Schweiz AG"[85] . "Crypto Schweiz AG" has its seat in the city of Steinhausen, in the canton of Zug, was founded in 2018, and was renamed to "CyOne Security AG" in the same year[86]. "Crypto International AG" was founded as well in 2018, but has its seat in the municipality of Hünenberg, canton of Zug[87]. Some of such datapoints regarding founding dates, locations and board members were manually extracted from the publicly available Swiss Trade Registry, and linked to existing or newly added IRIs. For a broader and more in-depth application, all data of the trade registry could also be harvested and implemented automatically from the Swiss governmental open data platform[88]. Such an application might be interesting for institutions like the "Swiss Economic Archives"[89], or any other institution holding records regarding various corporate bodies.

Since the original "Crypto AG" was subject to some controversy for its business practices and involvements, plenty further information was researched and published in various sources. Thus, I've also added to the SPARQL query publicly available information about board mandates of various politicians[90] and some of their ties[91]. The full query adding additional triple-statements is accessible under:

https://github.com/specjo/ISADGtoRiC/blob/main/53_enriching_CorporateBody/53_Insert_Crypto.rq

[83] Hagelin, Boris (1979): Die Geschichte der Hagelin-Cryptos, p. 2

[84] Swiss Trade Registry ZEFIX (2022): UID CHE-100.069.067

[85] Brandenberg, Livio (2018): 14 Entlassungen bei Zuger IT-Sicherheitsfirma Crypto AG

[86] Swiss Trade Registry ZEFIX (2022): UID CHE-260.677.443

[87] Swiss Trade Registry ZEFIX (2022): UID CHE-421.008.995

[88] https://opendata.swiss/de/dataset/zefix-zentraler-firmenindex

[89] "The national competence centre for sources on the Swiss economy and economic history". https://wirtschaftsarchiv.ub.unibas.ch/en/

[90] Zentralplus (2020): Auch CVP-Vertreter waren mit der Crypto AG verbandelt

[91] Schorno, Urs-Ueli (2020): Zwei Zuger Politiker wussten über die Crypto-Affäre Bescheid

For a person researching the topic of "Crypto AG", the following knowledge graph might then be interesting. The red nodes show the three mentioned corporate bodies and its

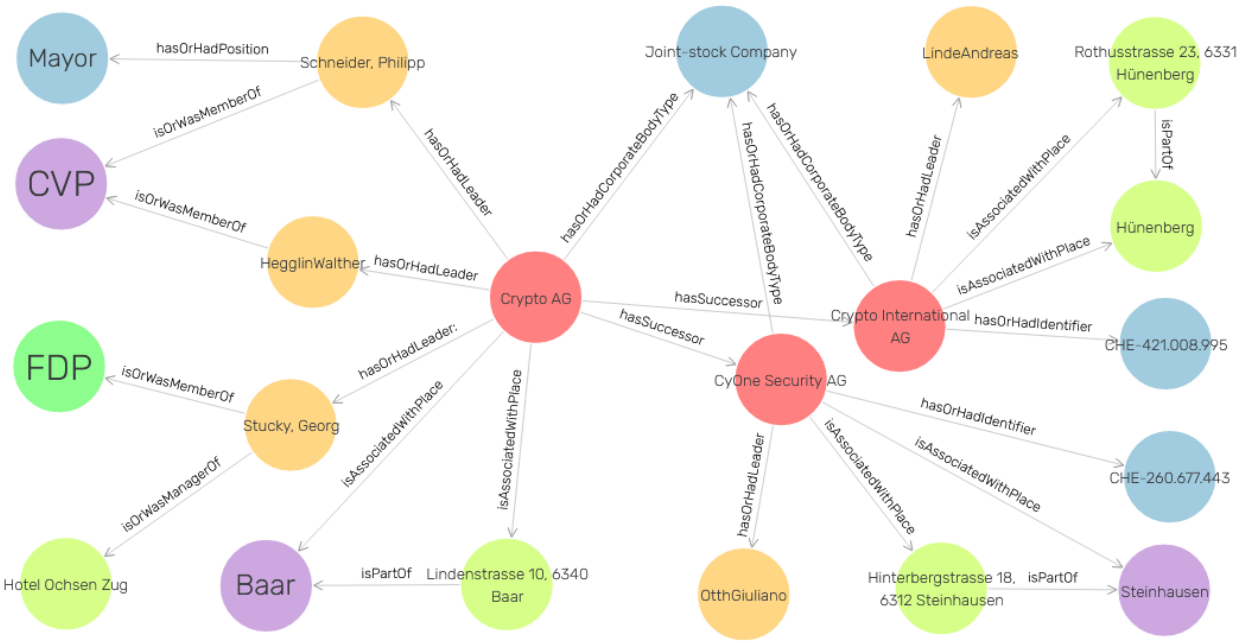


Figure 29: Knowledge graph of the CorporateBody "Crypto AG" and its successors

connection, as well as some Person-entities in yellow nodes, that are members of the board of directors. The knowledge graph then also depicts the spatial relations of the CorporateBody-entities, as well as the party-affiliations of some of the board-members (<CVP> and <FDP>) and a position held by one of them (<Mayor>).

Naturally, a researcher now might also be interested in archival and library holdings pertaining to the topic. Since linking of Records wasn't done during the mapping, these links have added at this point, and a similar procedure as in the previous example of "Bibliothek Zug" seem adequate.

Here again, to extract any relevant records, a small SPARQL-query was executed, asking for any triple-statement, where the subject is a RecordSet (Line 7), and filtering for where the predicate (whether a name, label, or anything else) contains the word "Crypto" (Line 8).

```
prefix rico: <https://www.ica.org/standards/RiC/ontology#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

select ?s ?o ?p where
{
    ?s ?p ?o.
    ?s ?p2 rico:RecordSet.
    FILTER CONTAINS (?o, "Crypto")
}

ORDER BY ?o
```

Figure 30: SPARQL query to extract any predicate containing the word "Crypto"

The result of that Query outputted 14 RecordSets[92]. The RecordSets were assigned with a SPARQL query after examination, a predicate (with `rico:hasOrHadSubject` or with `rico:hasOrHadMainSubject`) pointing to the respective CorporateBody-subject[93].

In this case, one might be more interested in all RecordSets related to any of these 3 corporations, in any Agent related to the creation or publication of these Records, and in where these RecordSets are held. A SPARQL Query querying and constructing for these parameters[94] renders then the following knowledge graph:

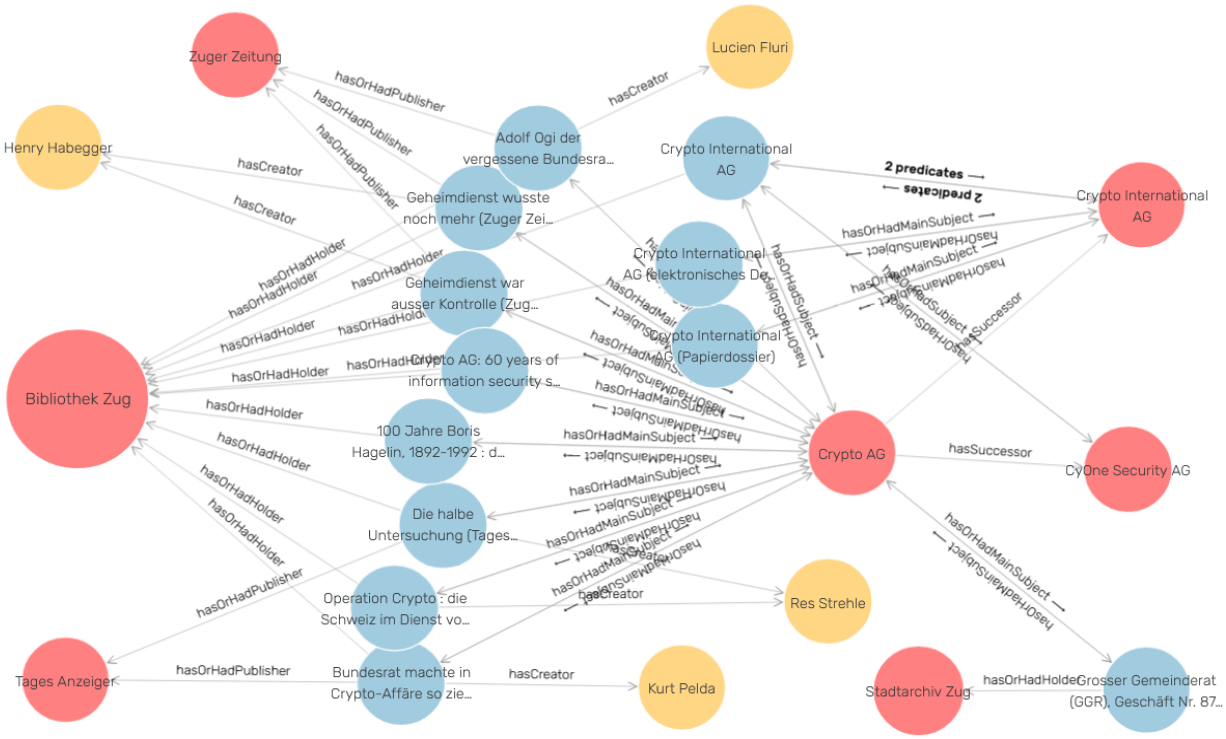


Figure 31: Knowledge graph focusing on Record relevant to Crypto AG and its successor companies

[92] https://github.com/specjo/ISADGtoRiC/blob/main/53_enriching_CorporateBody/53_SELECT_WHERE_Crypto_result.csv

[93] https://github.com/specjo/ISADGtoRiC/blob/main/53_enriching_CorporateBody/53_Insert_Records-Crypto.rq

[94] https://github.com/specjo/ISADGtoRiC/blob/main/53_enriching_CorporateBody/53_CONSTRUCT_WHERE_Crypto.rq

A researcher can here then see where relevant RecordSets are to be found, which authors were active on that topic and potentially published other relevant pieces, and in which Newspaper (Zuger Zeitung, Tages-Anzeiger) such articles were mainly published.

5.4 Other applications

The previous three chapters focus on certain applications of the core entities RiC-E07 Agent (respectively their sub-entities RiC-E08 Person and RiC-E11 Corporate Body) and RiC-E02 Record Resource as well as the supporting entity RiC-E22 Place. The final export of the repository can be accessed under:

https://github.com/specjo/ISADGtoRiC/blob/main/54_further_use/Zug_repository.ttls .

The other core entity RiC-E06 Instantiation and RiC-E15 Activity, as well as other supporting entities[95] weren't explored at all. RiC-CM alone with its total of 22 Entities, 41 Attributes, 79 Relations and 5 Relation Attributes offers a broad buffet of description possibilities. Since RiC is extensible, any other ontologies or vocabularies can used to expand an institutions description, and thus in theory it should be possible to meet any description use there is. As such, RiC and the underlying technology is not only interesting for archives, but offers also interesting elements for all kinds of GLAM-institutions.

A gradual transition to RiC might be imaginable as well. In the case of archives, that could mean that data cleaning and transformation to IRIs could happen step-by-step. While RecordSets and Instantiations probably should be assigned IRIs from the very beginning, the transformation of other entities like Person or Place could take place gradually, similar to the examples described in the previous chapters. Implementation of IRIs for Persons, CorporateBodies, Places and the like might thus also depend on the local availability of useful IRIs. As shown previously with the example for Place, I was only able to extract an IRI for a postal address, but not for an exact building. In such a case it might make sense to only transfer Literals of buildings to IRIs, once a meaningful IRIs for the needed purpose is available by an external source. Depending on the use-case, one could resort to existing Authority Files, or would have to create IRIs for its own internal use.

It is naturally to be expected that RiC won't be homogenously implemented, as was already the case with ISAD(G). Nonetheless, the standard provides quite the well-defined structure for description, but also allows for enough individual creative freedom to meet an institution individual requirement. The underlying RDF allows then with enough

[95] RiC-CM v0.2, p.17

flexibility, that adjustments or corrections to description could be undertaken quite easily ex post. How RiC will ultimately be implemented, is very much dependent on the individual institutions themselves, and on "lighthouse"-institutions, that will lead the transformation by good example and grant support to smaller ones.

6 Recommendations for Institutions

This chapter concludes the practical part of this thesis, and answers the initial research question number 5. What are the learnings of the previously described processes, and what recommendations for institutions can be made? Going through the full process of exporting data from current description, mapping and transforming it to RiC, loading it into a new database, enriching the data, and querying it for different purposes, various insights were gained and lessons learned. The topic is broad, and so are the learnings. The recommendations below are neither exhaustive nor final, but rather summarize various experiences made and thoughts developed during the writing of this thesis. Summarizing the lessons learned and recommendations under (sometimes overlapping) categories helps to keep an overview of various directions.

Extracting, mapping, transformation

Extracting data

While the export of existing descriptions in the case of MBZ was quite straightforward, getting data out of the systems in the case of the AIS of StAZG, SAZG and BibZug, as well as out of the library catalogue of BibZug, has proven quite cumbersome. The StAZG even has its own archive IT-specialists, and even being able to access the existing RDB through SQL wouldn't allow them to extract data in a time-efficient and easy way. The creation of description is a time-consuming process, which takes up a lot of time for institutions, and that data thus poses a valuable asset. That asset should be accessible and extractable easily to the creators of the description. While data regarding a single unit of description was relatively easy to extract, it is worrying that exporting the structural hierarchy as a main feature of ISAD(G) (as well as of the RDA catalogue) was in both cases not possible. Thus, future software should not only facilitate the input of data, but also the extraction thereof in order to make it accessible and reusable for further use, and reduce dependency on external software vendors. The National Archives of the UK are not exaggerating when they say, that the catalogue is " a crucial business asset"[96]. Data is money, and the ownership of and accessibility to that data should fully lie within the hands of each institution.

[96] The National Archives (2022): Project Omega

Data modelling

RiC is a broad standard, offering many different elements. Most likely very few institutions will use the full range of the standard, but only parts of it. Thus, the development of a suitable data-model for each institution should be of high priority, in order to make most use of RiC and other expanded ontologies. Such a data-model should be exhaustive, but not extensive, and finding the right balance between both will be an important task to solve. The data model then should not only define what elements of RiC should be used, but can be broader and/or extended by something one might call an "internal implementation guide". For an institution whose holdings have a strong logical link to geographical Places, the integration of the RDF vocabulary *wgs*:

(http://www.w3.org/2003/01/geo/wgs84_pos#) to represent latitude, longitude and altitude information in the WGS84 geodetic space might make for example. For other institutions other vocabularies and ontologies might be more useful to use.

Such a guide should also specify how and what kind of IRIs and which controlled vocabularies (e.g. for `rico:CarrierType`, `rico:classification`) are to be used. The first rule of Linked Data is "*Use URIs as names for things*", and the second rule is *"Use HTTP URIs so that people can look up those names."* [97]. The second rule might deter archives, since using publicly available URIs and dealing with potentially confidential information at the same time might not seem to go to well together. But linked data doesn't equal linked open data. This shouldn't deter archives from applying the first rule though, because URIs can also strictly limited to internal usage withing a given institution. What is beyond question, is that for RiC URIs trump Literals in any case, and Literals should be avoided wherever possible. Especially "catch-all" fields, as currently in ISAD(G) "scope and content" is often used for, should be avoided like the plague.

Of importance might also be to define, if and how a certain archive wants to map the administrative history of the body the archive is in charge of, or whether and how to describe and link relations between Persons and other entities withing that administrative body, within a geographic space or similar. The applications and implications are broad, and have to be prepared carefully.

[97] Berners-Lee, Tim (2006): Linked Data

Data cleaning

The importance of clean data has been already elaborated in chapter 4.4.3, but can't be stressed enough. It is a completely normal phenomena, that while working with data also faulty data of various kinds is generated. Thus, correct, consistent and clean data is not a one-time job, but should be a high priority that is taken regularly care of. Transferring from one DB to another, imports in new systems and similar "waypoint" are a good opportunity to do a deep and thorough cleaning and revisiting of the data an institution has. If data isn't taken care of regularly, DBs tend to accumulate errors quickly, and thus correct catalogues should be a high priority for every institution. Only correct data and description is useful and usable data.

Given the age of "big data" and extensive, over centuries grown catalogues, cleaning should preferably not happen manually because of lack of feasibility, but rather with the aid of adequate tools to automate various cleaning processes.

Next generation technologies and tools

Internal vs. external use

Most archival information systems have an inherent flaw. The archival tectonic and description, as archivists use it in their internal clients, are most-often mirrored as online accessible "light" catalogues. What archivists input in the system, is more or less what clients receive as output. Such an implementation requires the external client to use the catalogue in a similar way as the internal archivist, and to have a similar understanding of description and tectonics; a pre-requisite that is rarely the case in reality. These differences might become even bigger with the implementation of RiC, and the requirements for the use of catalogues through internal archival users and external clients will diverge even more. Thus, it most likely makes sense for institutions in the future not to use one single software for internal description and external presentation, but to use separate software (or two separate integrated components of a similar software) for input versus output of description. An example of how such a presentation software might look is SPARNATURAL[98], which allows for an intuitive access of clients, while differing very much from how an internal software for archival specialists would have to look. Any new tools and/or catalogues (whether a single one or multiple ones are to be used) must take these differences into consideration, and only then can offer the most beneficial solutions that the standard offers for both stakeholder groups.

[98] http://sparnatural.eu/index.html

User friendly GUI

RiC is quite a sophisticated standard, and one can't expect archival experts to know all the traits of it by heart. Thus, the development of new AIS must take into consideration facilitating the use for archivists as much as possible. How the specifications predefined through RiC might be implemented in a software would preferably be developed in co-development by software engineers and archival experts in cooperation. So far, archival software has rarely been developed through institutions themselves, but rather through specialized software development companies. Some kind of public private partnerships for the development of such software is imaginable. My personal belief is, that archival and similar institutions must commit and engage in the open-source development and provisioning of such software. A national or supra-national archival institution might be e.g. leading in such a development, where smaller archives could profit from that development and implement local versions of such software. That would need additional financial and personnel investments by archives, but would transfer the responsibility over and ownership of precious data from proprietary software and dependence of external vendors back into the hands of the institutions themselves. Ownership, in-house knowledge and more robust data-protection are only some of different possible benefits.

New tools with new potential

As should have become clear, RiC is not only new metadata fields, but much more. The new standard and its underlying technology have a huge potential. But one can only tap fully into that potential, by using the adequate tools. Adjusted or new AIS as well as the use of and connection to external resources will unleash the full potential. This doesn't only need a development of mind-set, but also a development of the (software) tools used. Inferring logical statements, graphical exploration of graph databases and external links, geo-referencing on multiple dimensions of geospatial information are only three keywords, that will become possible only with correct data and the adequate software.

Personnel

Archivists and other collection and records managers have been facing in the last decades with technical and other developments new challenges and changing requirements to their profession in general, and to specific individual skills in particular[99]. The introduction of RiC certainly will underline this development again. I don't believe that archival experts shall become computer scientists, but generally speaking, the trend of requiring a broader technological understanding of archivists will probably continue in the future even more. Therefore, I believe that institutions should invest in their personnel in two general directions:

- **Archival specialists**: Introducing RiC will certainly require further in-depth training of the existing personnel that one would summarize as archival specialists. Such training could focus on two main topics, and should probably emphasize both simultaneously and weigh equally.
  - RiC Standard: One focus will have to be on understanding the standard itself. Which entities and which attributes exist, what are the meanings and intended uses of these element. How does the new standard relate to ISAD(G); what stays same or similar and what is new. How will it be implemented in your own local institution? What is the difference between the intellectual Record Resource and the materialized instantiation, and what are their specific traits?
  - The whole standard is well documented in its existing components RiC-IAD, RiC-O and RiC-CM, and will be supplemented by an additional "Application Guidelines" (RiC-AG). Nonetheless, one can't expect archivists to read through hundreds of pages, and right-away know how to use the new standard. The standard is as already mentioned not only about additional attributes, but requires also an evolving understanding of provenance, of description, and of the possibilities that RiC offers. After 3 decades of ISAD(G), most archivists have that standard figuratively burned in in their DNA. Thorough in-depth training of current and of future generations of archivists will be crucial for the success of the new standard.

[99] Hickerson, H. T. (2001). Ten Challenges for the Archival Profession. P. 6 – 16

  - Technology: Understanding the standard without understanding the technological underpinnings seems to me impossible. One doesn't have to re-educate archivists into computer scientists, but understanding how linked data works, what RDF and what ontologies are, and what possibilities exist will definitely help to link the theoretical basis of RiC-CM to the real-world application, and use it professionally in one's everyday work.
- **IT specialists**: Most medium-sized and archival institutions already nowadays employ IT specialists in charge of various aspects of technology; be it software, hardware, integrated systems and similar. I dare to say, that in the future proportionally even more personnel with main focus on various IT-fields will have to be hired, whether they are mainly IT-personnel, or bring a mixture of archival and IT skills to the table. IT specialists must play an important role on the technical side in implementing quiet a complex model in such a way, that it integrates as seamlessly as possible in an archives daily business and facilitates the transition for all personnel.

7 Conclusion and outlook

There are multiple facets to answering the initial research questions 1a and 1b, how I-SAD(G) and non-ISAD(G) description can be transformed to RiC. RiC-CM is documented in quite detail, which shouldn't make a mapping from ISAD(G), nor from other description methods too complicated. Nonetheless, there are as shown in this thesis other factors coming into play, that make the transition more challenging. Two main challenges are the source systems and the current description. Existing DBs in source systems don't necessarily make it easy to export complete datasets, as was shown below, and is an issue to be solved by IT-technicians. Current description on the other hand is an issue more related towards archival specialists. Only because a current attribute field is labelled a certain way, doesn't mean that the data in that field exactly conforms to that description. Inconsistent or faulty description, catch-all attributes and the like don't allow for direct transfer to RiC, but require a thorough data-cleaning.

If these mapping issues are solved adequately and a consistent data model is used for the mapping to RiC, then the research question 2, how various descriptions can be combined, is already solved as well. The underlying RDF-standard and logic of linked data makes it fairly easy, to combine description from various source, if mapped consistently. As shown in the examples above, the various data-exports were mapped individually, and loaded individually into the database. Since a consistent data-model was used, no further alignment was needed, and all description was ready to be used as a combined catalogue.

In summary, the process of transitioning to the new RiC standard definitely isn't a trivial one. Institutions will face on the way various challenges, and the new standard certainly won't be implemented over night. To successfully walk this path, the archival community, its institutions, and the archival experts will require the proper mindset and determination. A revisited understanding of records, their description, the contexts and connections to the physical and digital world outside the walls of the institutions, as well as embracing the benefits of adequate use of IT tools is required and will facilitate that transition.

This thesis has only explored some of the challenges of data mapping and transfer, and of general use of RiC, but surely many more questions will arise. The appropriate procedure of transitioning has to be planned by institutions well ahead. It is not only about mapping and translating of existing description, not only about procurement of new software, or not only about the training of existing or hiring of new personnel. All of the points mentioned above and many others are interlinked, and only when taking care of all the components simultaneously and adequately. There will not be a one-size-fits-all solution

to these issues. In this thesis some use-cases and benefits were shown, but there are (one dare say endless) more possibilities and use-cases. For some issues general or standard solutions might be applicable, but for many individual challenges institutions will have to find individual solutions. On basis of the experiences made during the writing of this thesis and with focus on the future, various recommendations were given at the end of this thesis to answer research question 5. I thus firmly believe that the success of RiC depends on multiple factors. In the recommendations some factors playing a role like user-friendliness, personnel training, usability, adequate tools and general mindset towards further development were mentioned. In addition, the archival community as whole and especially leading institution with flagship projects implementing RiC are challenged to lead the standard to success together.

With the publication of RiC 1.0 and the gradual implementation, certainly a new era for archives and archival description is awaiting. The coming years will be challenging and exciting at the same time. I am curious about future projects and what kind of solutions will be developed from different sides. RiC and Linked Data is a great opportunity for archives and other glam institutions alike, and will hopefully benefit all stakeholders, from archivists over scholars and researchers to the general society

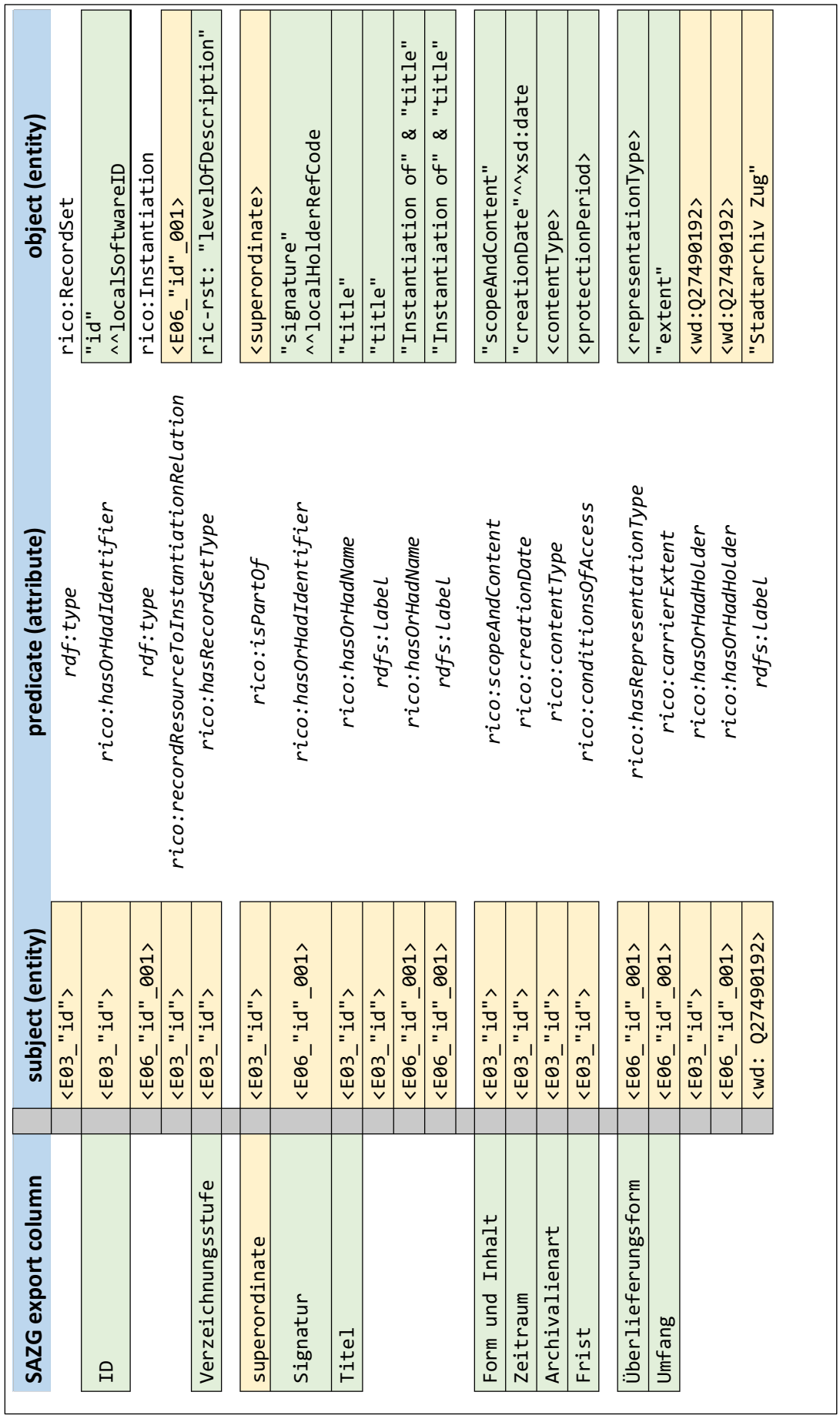
8 Appendices

8.1 Appendix 1: Overview of database export attributes/columns

This table shows in the header row which institution the export was taken from and the relating chapter. It illustrates then in the respective columns all attributes, that were extracted from the original DB. The labelling of the attributes is original as it was extracted in German language.

SAZG [CMI]	MBZ	BibZug Quria	BibZug CMI
ID	Inv.-Nr.	Mediennummer	ID
Verzeichnungsstufe	Objekt/Titel	Manifestationsart	Verzeichnungsstufe
Titel	Pers./Körp.	Titel	Titel
Signatur	Datierung	Regal	Signatur
Form und Inhalt	Masse	Geistiger Schöpfer	Form und Inhalt
Überlieferungsform	Inschr./Sign.	Bibliothek	Überlieferungsform
Zeitraum	Mat./Technik	Medientyp	Zeitraum
Umfang	Beschreibung	Sprache	Umfang
Archivalienart	Zustand	Publikationsvermerk	Archivalienart
Frist	Reprorecht		Frist
			Urheber
			Format

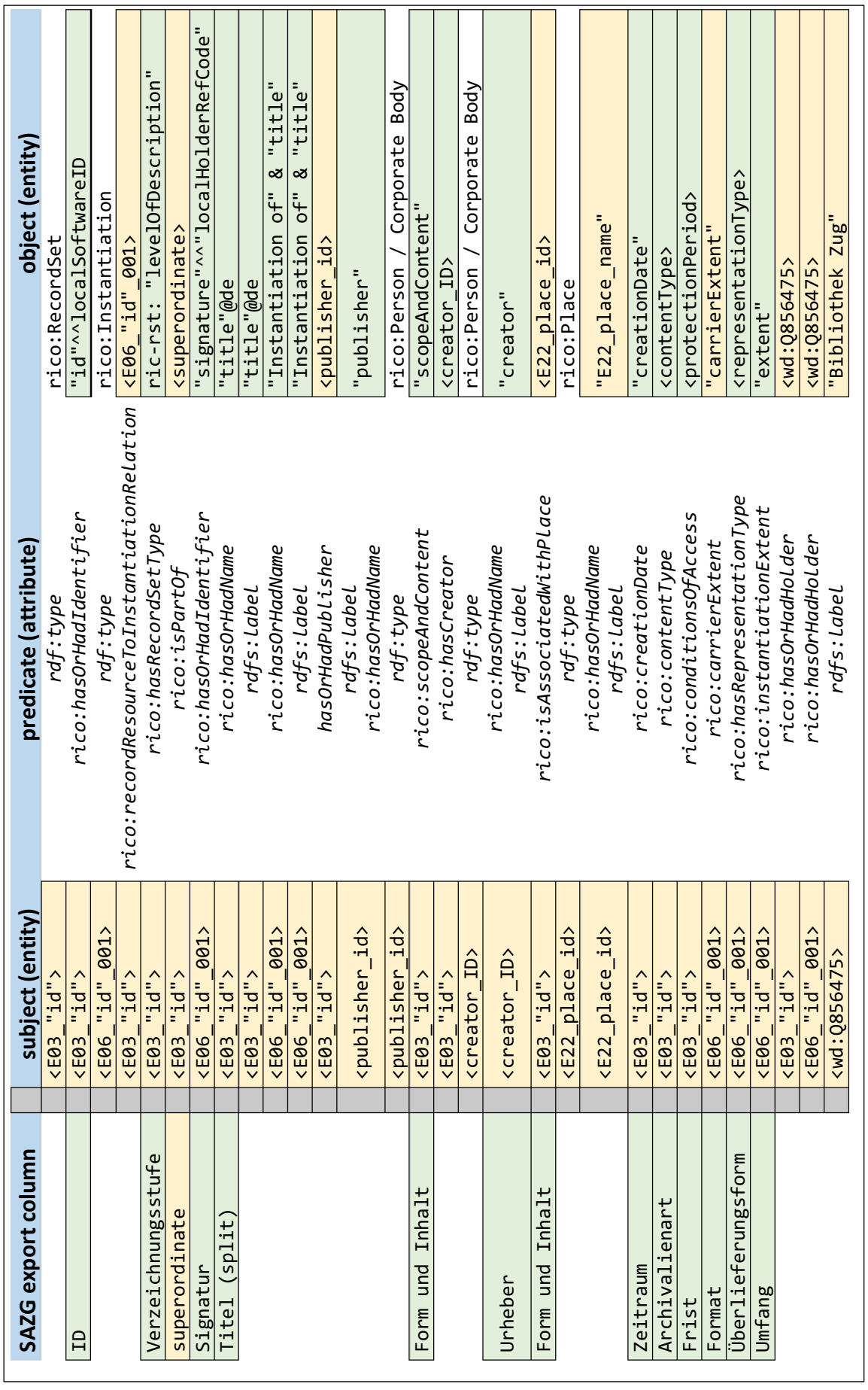
8.2 Appendix 2: Target renaming / mapping for SAZG



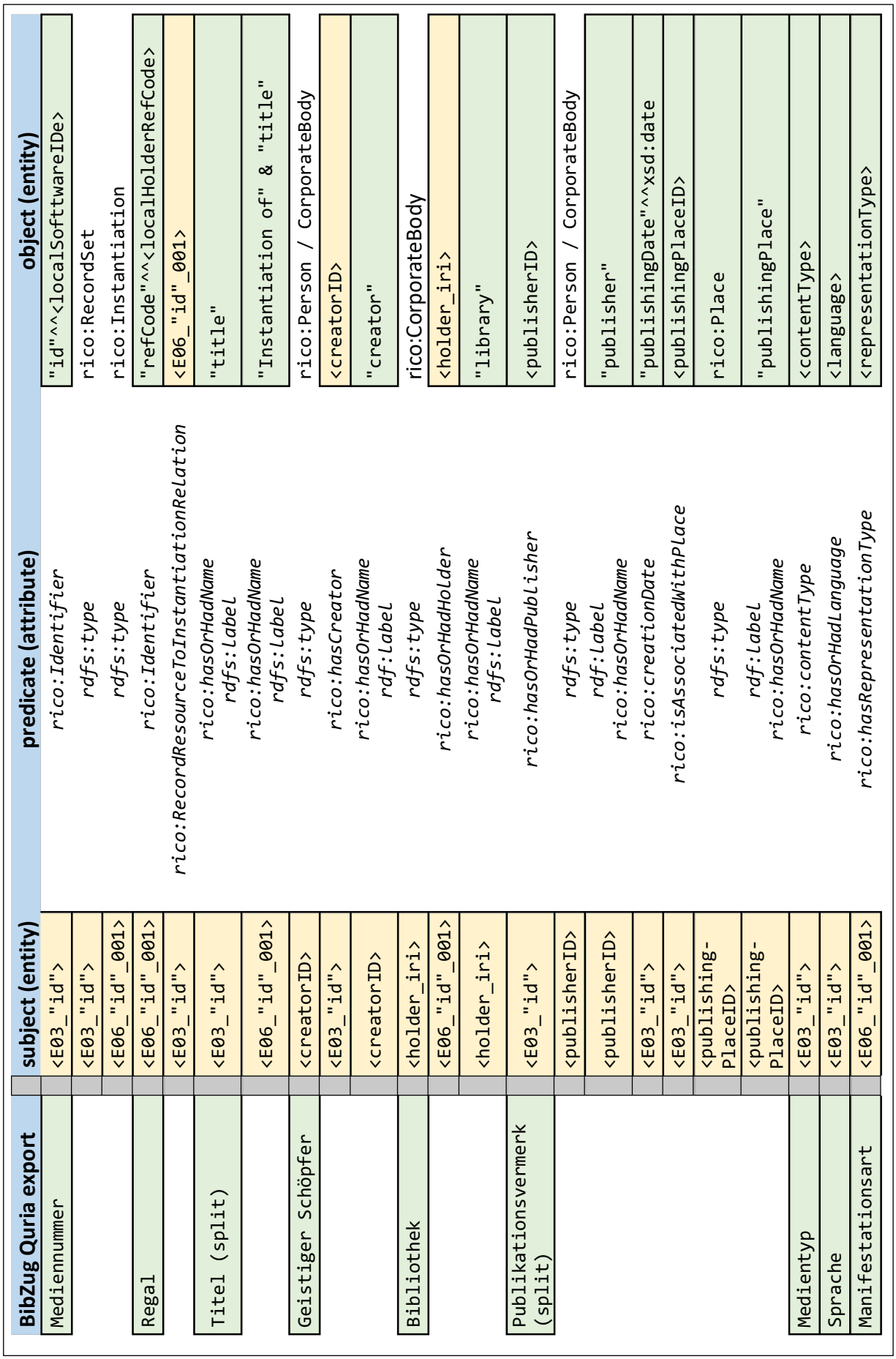
8.3 Appendix 3: Target renaming / mapping for MBZ

MBZ export column	subject (entity)	predicate (relation)	object (entity)
	<E06_"invnr"_001>	*rdf:type*	rico:Instantiation
Inv.-Nr.	<E06_"invnr"_001>	*rico:hasOrHadIdentifier*	"invnr"
Objekt/Titel	<E06_"invnr"_001>	*rico:hasOrHadName*	"title"
Objekt/Titel	<E06_"invnr"_001>	*rdfs:label*	"title"
Datierung	<E06_"invnr"_001>	*rico:creationDate*	"creationDate"^^xsd:date
Masse	<E06_"invnr"_001>	*rico:carrierExtent*	"carrierExtent"
Inschr./Sign.	<E06_"invnr"_001>	*rico:physicalCharacteristics*	"inscription"
Mat./Technik	<E06_"invnr"_001>	*rico:hasProductionTechnique*	"productionTechnique"
Zustand	<E06_"invnr"_001>	*rico:qualityOfRepresentation*	"quality"
Reprorecht	<E06_"invnr"_001>	*rico:hasOrHadHolder*	<E07_agent>
	<E03_"invnr">	*rdf:type*	rico:RecordSet
	<E03_"invnr">	*rico:hasRecordSetType*	ric-rst:File
	<E03_"invnr">	*rico:RecordResourceToInstantiationRelation*	<E06_001_invnr>
Objekt/Titel	<E03_"invnr">	*rico:hasOrHadName*	"title"
Objekt/Titel	<E03_"invnr">	*rdfs:label*	"title"
Beschreibung	<E03_"invnr">	*rico:hasScopeAndContent*	"scopeAndContent"
	<E03_"invnr">	*rico:hasOrHadHolder*	<E07_agent>
	<E03_"invnr">	*rico:hasCreator*	<E07_agent>
Pers./Körp. (split)	<E07_agent>	*rdf:type*	rico:<agent_type>
	<E07_agent>	*rico:hasOrHadName*	"name"
	<E07_agent>	*rdfs:label*	"name"
	<E07_agent>	*rico:hasBirthDate*	"birthDate" ^^xsd:date
	<E07_agent>	*rico:hasDeathDate*	"deathDate"^^xsd:date
	<E07_agent>	*rico:hasOrHadCorporateBodyType*	<corporateBodyType>
	<corporateBodyType>	*rdfs:label*	corporateBodyType_label
	<E07_agent>	*rico:hasOrHadOccupationOfType*	<occupation_id> (Wikidata)
	<occupation_id>	*rdfs:label*	"occupation_title"

8.4 Appendix 4: Target renaming / mapping for BibZug CMI



8.5 Appendix 5: Target renaming / mapping for BibZug Quria



9 List of Sources

Ashraf, Sharjeel (2020): *ETL is Not Dead. It is Still Crucial for Business Success* https://dataintegrationinfo.com/etl-is-not-dead/ [Accessed: 13.06.2022]

Axiell (2022): *Quria*; https://www.axiell.com/solutions/product/quria/ [Accessed: 30.05.2022]

Berners-Lee, Tim (2006): Linked Data; https://www.w3.org/DesignIssues/LinkedData [Accessed 20.07.2022]

Brandenberg, Livio (2018): 14 Entlassungen bei Zuger IT-Sicherheitsfirma Crypto AG; in: Luzerner Zeitung, 24.01.2018;

https://www.luzernerzeitung.ch/zentralschweiz/zug/steinhausen-14-entlassungen-bei-zuger-it-sicherheitsfirma-crypto-ag-ld.102942 [Accessed: 07.07.2022]

Brüning, Rainer; Heegewald, Werner und Brübach, Nils (2006): *Internationale Grundsätze für die archivische Verzeichnung*. Durchgesehener Nachdruck der 2. überarbeitete Auflage; in: Veröffentlichungen der Archivschule Marburg Institut für Archivwissenschaft (23)

Bundesarchiv (2022): Bundesverfassung von 1848;

https://www.recherche.bar.admin.ch/recherche/#/de/archiv/einheit/5998435 [Accessed: 13.06.2022]

Cai, L. and Zhu, Y. (2015): *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*; in: Data Science Journal (14), p. 2 – 10; DOI: http://doi.org/10.5334/dsj-2015-002

Cambridge Dictionary (2022): *Garbage In, Garbage Out;* https://dictionary.cambridge.org/dictionary/english/garbage-in-garbage-out [Accessed 07.06.2022]

DB-Engine (2022): *DB-Engines Ranking of Graph DBMS*; Version of July 2022; https://db-engines.com/en/ranking/graph+dbms [Accessed 02.07.2022]

Elgabry, Omar (2019): *The Ultimate Guide to Data Cleaning,* in: Towards Data Science; https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4 [Accessed 10.07.2022]

Expert Group on Archival Description (2021): Call for comments on ICA RiC-CM 0.2; https://www.ica.org/en/call-for-comments-on-ica-ric-cm-02 (Accessed 04.07.2022)

Gutierrez Claudio, Hurtadob Carlos A., Mendelzon Alberto O., Pérezcal Jorge (2011): *Foundations of Semantic Web Databases*; in: Journal of Computer and System Sciences, 77(3), p. 520-541; https://doi.org/10.1016/j.jcss.2010.04.009

Hagelin, Boris (1979): Die Geschichte der Hagelin-Cryptos

Hickerson, H. T. (2001): Ten Challenges for the Archival Profession; in: The American Archivist, 64(1), p. 6–16

International Council on Archives, Brien Brothman (1992): *ISAD(G): General International Standard Archival Description*; in: Archivaria 34: International Congress on Archives (ICA) (Summer 1992); p. 17 - 32

International Council on Archives (2000): *ISAD(G): General International Standard Archival Description*; Second Edition; Adopted by the Committee on Descriptive Standards; Stockholm, Sweden, 19-22 September 1999

International Council on Archives (2021): *Records in Contexts – Conceptual Model (RiC-CM)*, Consultation Draft v0.2, July 2021; https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf [Accessed 13.06.2022]

International Council on Archives (2021): *Records in Contexts – Introduction to Archival Description (RiC-IAD) v 0.2;* https://www.ica.org/en/records-in-contexts-introduction-to-archival-description [Accessed 11.06.2022]

International Council on Archives (2021): *Records in Contexts – Ontology (RiC-O) v 0.2;* https://www.ica.org/standards/RiC/ontology [Accessed 24.07.2022]

International Council on Archives (2021): *RiC-O projects and tools;* https://ica-egad.github.io/RiC-O/projects-and-tools.html [Accessed 03.07.2022]

International Federation of Library Associations and Institutions (2009): *Functional Requirements for Bibliographic Records,* Final Report, September 1997, As amended and corrected through February 2009

Kasenchack, B., Lehnert A, Loh G (2021): *Use Case: Ontologies and RDF-star for Knowledge Management*; in: Eighteenth Extended Semantic Web Conference - Industry Track; https://openreview.net/forum?id=Prc8onC-jfq [Accessed 02.07.2022]

Keil, Jan Martin (2016): *IRI, URI, URL, URN and their differences* https://fusion.cs.uni-jena.de/fusion/2016/11/18/iri-uri-url-urn-and-their-differences/ [Accessed 24.06.2022]

Malhotra, Ashok. (2012): *Mapping Relational Data to RDF;* https://www.w3.org/blog/2012/08/mapping-relational-data-to-rdf/ [Accessed 13.06.2022]

Meindertma, Joep (2018): *A brief introduction to linked data*

https://ontola.io/blog/what-is-linked-data/ [Accessed 24.06.2022]

Meindertma, Joep (2019): *What's the best RDF serialization format?*

https://ontola.io/blog/rdf-serialization-formats/ [Accessed 24.06.2022]

Merzaghi, Michele; Ohnesorge, Krystyna (2020): *Neue Wege in der archivischen Erschliessung – Scopriamoli assieme*; in: arbido 2020/4: Virtualisierung; https://arbido.ch/de/ausgaben-artikel/2020/dematerialisation/neue-wege-in-der-archivischen-erschliessung-scopriamoli-assieme [Accessed: 02.03.2022]

Messner, Philippe (2017): *Records in Contexts – vom Baum zum Netz*; in: arbido 2017 / 3: Metadaten – Datenqualität; https://arbido.ch/en/ausgaben-artikel/2017-1/metadaten-datenqualit%C3%A4t/records-in-contexts-vom-baum-zum-netz [Accessed: 01.03.2022]

Morosoli, Renato (2007): *Dagobert Keiser (Version vom 18.05.2007)*; in: Historisches Lexikon der Schweiz Online; https://hls-dhs-dss.ch/de/articles/027372/2007-05-18/ [Accessed: 12.07.2022]

Morosoli, Renato (2014): *Dagobert Keiser (Version vom 26.11.2014)*; in: Historisches Lexikon der Schweiz Online; https://hls-dhs-dss.ch/de/articles/042168/2014-11-26/ [Accessed 12.07.2022]

National Archives, the (2022): Project Omega.https://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-plans/our-digital-cataloguing-practices/project-omega/ [Accessed 07.07.2022]

ontotext (2022): *Loading Data using OntoRefine;*https://graphdb.ontotext.com/documentation/free/loading-data-using-ontorefine.html *[Accessed 16.06.2022]*

Schorno, Urs-Ueli (2020): Zwei Zuger Politiker wussten über die Crypto-Affäre Bescheid; in: zentralplus, 24.06.2020; https://www.zentralplus.ch/gesellschaft/drei-politiker-wussten-ueber-die-crypto-affaere-bescheid-1826469/ [Accessed 08.07.2022]

Swiss Trade Registry ZEFIX (2022): UID CHE-100.069.067;https://www.uid.admin.ch/Detail.aspx?uid_id=CHE-100.069.067 [Accessed 06.07.2022]

Swiss Trade Registry ZEFIX (2022): UID CHE-260.677.443;https://www.zefix.ch/de/search/entity/list/firm/1335789 [Accessed 06.07.2022]

Swiss Trade Registry ZEFIX (2022): UID CHE-421.008.995;https://www.zefix.ch/de/search/entity/list/firm/1333057 [Accessed 06.07.2022]

Tropper, C. und Zeloth, T. (2010): *Kärntner Landesarchiv - Richtlinien für die Ordnung und Verzeichnung von Archivgut, auf der Grundlage von ISAD(G) und hausinternen Traditionen*

Verein Schweizerische Archivarinnen und Archivare (2009): *Schweizerische Richtlinie für die Umsetzung von ISAD(G)*, 1. Ausgabe, September 2009

Weber, Lisa B. (1989): *Archival Description Standards: Concepts, Principles, and Methodologies*; in: American Archivist (52), Fall 1989. p. 504 – 513

Wehrle M. (StAZG) (2022a): *eMail reply to Josip Spec on 29 April 2022*

Wehrle M. (StAZG) (2022b): *eMail reply to Josip Spec on 09 June 2022*

World Wide Web Consortium (1999): *Resource Description Framework (RDF) Schema Specification;* https://www.w3.org/TR/1999/PR-rdf-schema-19990303/ [Accessed 02.07.2022]

World Wide Web Consortium (2014)*: RDF 1.1 Primer;* https://www.w3.org/TR/rdf11-primer/ [Accessed 02.07.2022]

World Wide Web Consortium (2015): *Use cases*; last modified on 24 July 2015, at 12:31. https://www.w3.org/community/rsp/wiki/Use_cases [Accessed 02.07.2022]

World Wide Web Consortium (2022): *Resource Description Framework (RDF)* https://www.w3.org/RDF/ [Accessed 02.07.2022]

Zentralplus (2020): Auch CVP-Vertreter waren mit der Crypto AG verbandelt; 21.02.2020; https://www.zentralplus.ch/news/auch-cvp-vertreter-waren-mit-der-crypto-ag-verbandelt-1734521/ [Accessed 08.07.2022]

Bisher erschienene Schriften

Ergebnisse von Forschungsprojekten erscheinen jeweils in Form von Arbeitsberichten in Reihen. Sonstige Publikationen erscheinen in Form von alleinstehenden Schriften.

Derzeit gibt es in den Churer Schriften zur Informationswissenschaft folgend
Reihe Berufsmarktforschung

Weitere Publikationen

Churer Schriften zur Informationswissenschaft – Schrift 144
Herausgegeben von Wolfgang Semar
Jasmin Suter
TikTok User sind einfacher manipulierbar
Einfluss von Videoplattformen auf das Verhalten in der Pre-Purchase Phase am Beispiel TikTok
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 145
Herausgegeben von Wolfgang Semar
Lea Bächli
Die Veränderungen der Angebote öffentlicher Bibliotheken in der Deutschschweiz durch die COVID-19-Pandemie
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 146
Herausgegeben von Wolfgang Semar
Jeffrey Santana de Jesus
Mithilfe von Digital Nudging mehr Privatsphäre in sozialen Netzwerken?
Digital Nudging in sozialen Netzwerken
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 147
Herausgegeben von Wolfgang Semar
Regina Eicher
Die Entwicklung inhaltlicher Sprachbegriffe für eine verbesserte Erschliessung von Kinder- und Jugendzeichnungen
Eine qualitative Inhaltsanalyse von 12 ausgewählten Märchen
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 148
Herausgegeben von Wolfgang Semar
Andrej Kilian
«Die Bibliotheksthematik hat sich in den letzten Jahren stark relativiert»
Interne Bibliotheken in der Deutschschweiz und in Liechtenstein – Versuch eines Einblicks
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 149
Herausgegeben von Wolfgang Semar
Sandra Freiburghaus
Untersuchung von Anzeige- und Reservationssystemen zur Lernplatzorganisation in Bibliotheken
Unter Betrachtung der Bedürfnisse und Erfahrungen der Institution
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 150
Herausgegeben von Wolfgang Semar
Nicole Fässler
User Adoption bei der Einführung einer Kollaborations- und Kommunikationssoft-ware im Modern Workpl
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 151
Herausgegeben von Wolfgang Semar
Marina Inglin
Re- und Upskilling-Empfehlung
Kriterien für die automatische Auswahl von Re- und Upskilling-Angeboten
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 152
Herausgegeben von Wolfgang Semar
Lisa Heller
Zur Genese eines nationalen Bibliothekspro-jekts: Swiss Library Service Platform (SLSP)
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 153
Herausgegeben von Wolfgang Semar
Antonin Friberg
Die Effektivität von Social Norms Nudging in der Customer Journey
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 154
Herausgegeben von Wolfgang Semar
Curdin Marxer
«Drug Repurposing»
Wie können unstrukturierte Textdaten für die Ermitt-lung neuer «Drug Repurposing» Kandidaten nutzbar gemacht werden und wie können sie Datenbanken ergänzen?
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 155
Herausgegeben von Wolfgang Semar
Samir Limani
Sicht der administrativen Mitarbeitenden von Bündner Spitälern und Kliniken auf den Digitalisierungsstand ihres Unternehmens
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 156
Herausgegeben von Wolfgang Semar
Marina Lea Schürmann
Deep Learning für Part-of-Speech-Tagging
Vergleich eines auf Transformers basierenden POS-Taggers mit bestehenden Modellen
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 157
Herausgegeben von Wolfgang Semar
Yannick Mireille Kaufmann
Einsatz von Unternehmenswikis als Wissens-management-Tool in einer Netzwerkorganisation
Evaluationsstudie zu «wikimia», eine Wissensdaten-bank in der schweizerischen Berufs-, Studien- und Laufbahnberatung Masterthesis 2022
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 158
Herausgegeben von Wolfgang Semar
Franco Malacrida
Standortfindung von Schweizer Start-ups
Welche Standortfaktoren sind für Schweizer Start-ups am wichtigsten?
Chur, 2023
ISSN 1660-945X

Über die Informationswissenschaft Fachhochschule Graubünden

Die Informationswissenschaft ist in der Schweiz noch ein relativ junger Lehr- und Forschungs
ternational weist diese Disziplin aber vor allem im anglo-amerikanischen Bereich eine jahrzehnt
dition auf. Die klassischen Bezeichnungen dort sind Information Science, Library Science oder
Studies. Die Grundfrage-stellung der Informationswissenschaft liegt in der Betrachtung der Ro
Um-gangs mit Information in allen ihren Ausprägungen und Medien sowohl in Wirtschaft und G
Die Informationswissenschaft wird in Chur integriert betrachtet.

Diese Sicht umfasst nicht nur die Teildisziplinen Bibliothekswissenschaft, Archivwissenschaft
mentationswissenschaft. Auch neue Entwicklungen im Bereich Medien-wirtschaft, Information
sensmanagement und Big Data werden gezielt aufge-griffen und im Lehr- und Forschungspro
rücksichtigt.

Der Studiengang Informationswissenschaft wird seit 1998 als Vollzeitstudiengang in Chur ang
seit 2002 als Teilzeit-Studiengang in Zürich. Seit 2010 rundet der Master of Science in Busines
ration das Lehrangebot ab.

Das Forschungsfeld Informationswissenschaft vereinigt Cluster von Forschungs-, Entwick
Dienstleistungspotenzialen in unterschiedlichen Kompetenzzentren:

- Bibliothek und Digitalisierung von analogem Kulturgut
- Bildungsinformatik
- Data Analytics
- Digital Business and Usability Engineering
- Information Lifecycle Management
- Knowledge and User Research
- Practical Data Science
- Process Data, Visualization, and Machine Learning
- Scientific Computing

Diese Kompetenzzentren werden im Swiss Institute for Information Science (SII) zusammenge

Impressum

Impressum

FHGR – Fachhochschule
Graubünden
Information Science
Pulvermühlestrasse 57
CH-7000 Chur

www.informationscience.ch
www.fhgr.ch
ISSN 1660-945X

Institutsleitung

Prof. Dr. Ingo Barkow
Telefon: +41 81 286 24 61
Email: ingo.barkow@fhgr.ch

Sekretariat

Telefon: +41 81 286 24 24
Fax: +41 81 286 24 00
Email: clarita.decurtins@fhgr.ch