Humboldt-Universität zu Berlin
Philosophische Fakultät
Institut für Bibliotheks- und Informationswissenschaft

Modelling and Automated Retrieval of Provenance Relationships

Masterarbeit im Rahmen des Weiterbildenden Masterstudiengangs
Bibliotheks- und Informationswissenschaft im Fernstudium

vorgelegt von
Thomas Schneider

Gutachter:
Prof. Dr. Robert Jäschke
Christian Rüter

Erfurt, den 9. April 2023

Contents

List of Figures

1 Introduction

1.1 Background

In research on the historical holdings of cultural heritage institutions, the origins of cultural objects are of central importance. The origin of an item is also called its *provenance* and comprises the people or corporate bodies that owned this item over time.

In the case of the holdings of university and research libraries, particularly interesting provenances are those related to the change of owners when a book item is passed on or distributed [Hakelberg 2016, p. 2]. Provenances can be reconstructed by marks of ownership such as stamps, bookplates (ex libris), or handwritten signatures: with the help of these features, it is possible to retrace the "history" of a book item or the extent of library holdings that have been scattered in the meantime [Hakelberg 2016, p. 2].

In order to enable provenance research, libraries index the provenances of their historical holdings and make them available to users via their catalogues. A provenance entry refers to a person or coporate body and to a feature that indicates ownership. In order for owners to be identified unequivocally, they are often given via a reference to authority files such as the Integrated Authority File (GND) of the German National Library [Web1]. The indexed provenance data makes it possible, for example, to query and reconstruct the items owned or held by a single person or corporate body, to query the whereabouts of relevant items, or to retrace the distribution of all indexed exemplars derived from a given work.

Nowadays, provenance entries are recorded in electronic catalogues. German university and research libraries typically do not maintain their own individual catalogue; rather they are provided with a central catalogue by the library network in which they participate.[1] These central catalogues are equipped with an underlying database and standardised data formats for internal representation and data export. For example, the networks GBV and SWB maintain and use the common catalogue (database) *K10plus* [Web3], which internally uses the data format PICA [Web4] and allows for exports in the data formats MARC 21, MAB2, and Pica+ [Web5]. Despite the use of a uniform data format, there are several possibilities to record provenance entries. As Hakelberg [2016, Chapter 4] explains, libraries even within the same network often use diverse representations for the same type of provenance entry, and the differences are considerable: for example, some GBV libraries record their provenance entries in data fields on the bibliographic level, while others use the level on the exemplar level. These deviations lead to large differences in the presentation of the holdings in the online catalogue, which hinders the retrieval of relevant items and historical holdings.

Data about persons and corporate bodies are recorded in Germany- or worldwide authority files and further databases, such as the previously mentioned GND, the *WorldCat* [Web6], databases of projects such as ISNI [Web7] and VIAF [Web8] or in Wikidata [Web9]. These data sources usually support standardized data formats for data export, such as MARC 21 and RDF via several interfaces in the case of GND. Data about a person contain, among others, the name and alternative name forms (which can be manifold in the case of, e.g., scholars of previous centuries), places of birth, death, and work, as well as relationships to corporate bodies and other persons (e.g.,

[1] There are six library networks *(Bibliotheksverbünde)* for scientific libraries in Germany [Web2].

coauthors and students). The extent of a dataset of the same person can differ between data sources, which is witnessed, for example, by the entries on the scholar Georg Joachim Rheticus in ISNI, VIAF, WorldCat, GND, and Wikidata.[2] Hence, the state of data on persons and corporate body is heterogeneous as well, and depending on the concrete individual, it may be necessary to consult several data sources and combine the obtained data.

Given this diversity and heterogeneity of the existing data sources, it is currently difficult to impossible to retrieve provenance relationships that are not restricted to bibliographic objects and their owners but which also involve various relationships between works, exemplars, and (multiple) owners. For example, one could be interested in answers to queries of the following form.

Q1 Who read work W, in which manifestation and in which year?

Q2 Which exemplars[3] of work W were passed from one of its owners to a collaborator (or a student)?

Q3 What are the relationships between the recipients of work W (or of manifestation M of W or of exemplar C of W, respectively)?

In these examples, we use variables W, M, C to refer to arbitrary works, manifestations, or exemplars. Therefore, **Q1**–**Q3** are in fact *query patterns*, each of which represents a set of possible queries that can be obtained by assigning concrete objects to the variables. We will elaborate on this thought in Chapter 3.

Query Pattern **Q1** addresses works as well as their manifestations (e.g., editions of the same work in various languages). An answer to a query of this type would allow it to trace the reception of the same work over several eras. For example, Duchess Luise Dorothea of Saxe-Gotha-Altenburg read French editions of English works.[4] Obviously, there is a difference between the action "read" used in this query pattern and the relationship "owned" represented by provenance (entries). We neglect this difference for the moment and will get back to it in Chapter 3.

give more concrete example

Query Patterns **Q2** and **Q3** aim at highlighting the network that spans between the recipients of a work. For example, one of the two exemplars of Nicolaus Copernicus's main work *De revolutionibus orbium coelestium* [Copernicus 1543] that are now held by the Gotha Research Library of the University of Erfurt have been owned, or at least read, by several scholars from the circle around the author, some of which were in the teacher-student relationship. This information can be concluded from the accounts of Gingerich [Gingerich 2002, p. 69] and Salatowsky and Lotze [Salatowsky and Lotze 2015, p. 142], but it can also be obtained by looking up the entry and its owners in the electronic catalogue of the library, following the links to the GND entries of the owners, and inspecting the relationships between the owners in GND; see Chapter 3 for a more detailed derivation.

An important difference between Patterns **Q2** and **Q3** is the following. While **Q2** fixes a relationship between two persons ("collaborator" or "student") and asks for works in whose context this relation occurs, **Q3** does not fix a particular relationship but asks for the entire context of the work or a manifestation or exemplar thereof.

When attempting to answer queries conforming to patterns such as **Q1**–**Q3**, it does not suffice to consult a single data source such as a catalogue, authority file, or knowledge base. Instead, it is

[2] This can be verified by following, at the end of the Wikipedia page on Rheticus in the table "Authority Control" [Web10], the links to ISNI, VIAF, WorldCat und GND, and inspecting the Wikidata data set [Web11].

[3] Conforming to the FRBR model [IFLA SG FRBR 2009], the precise wording should be "examplars of manifestations of expressions of W" but, here and in the following, we omit the intermediate entities for brevity when no misunderstanding is expected.

[4] Dietrich Hakelberg, Research Library Gotha of the University of Erfurt, personal communication

necessary to consult several data sources and combine the information found. This process is highly laborious, given not just the number of data sources but also their diverse data formats. Therefore, automated support is essential. This is one of the reasons for Hakelberg to raise the following question [Hakelberg 2016, p. 46, translated from German]: "How can historical provenance relationships be formulated and represented in a machine-readable way?"

In order to implement suitable tools, it is necessary to analyse available data sources, data models, and data integration techniques, to develop an abstract model of data sources and possible queries, and to devise a method for obtaining answers in this abstract framework.

Address "Raubgut", items acquired without purchase ⇝ add relevant prototype queries?

Further thoughts on prototype queries:

- While the difference between persons and institutions may be negligible from a modelling point of view, it is highly relevant concerning the amount of query answers and their handling.
- Exemplars with traces of ownerships that do not name the owner can be particularly interesting to study in the context of missing holdings. Some libraries record such traces (e.g., HAAB Weimar), but of course not exhaustively so. Could this be a prototype query that can be incorporated into the model, or should it be discussed under possible extensions?

1.2 Aim and Research Question

In this thesis, we pursue the goal of facilitating the automated retrieval of provenance relationships. More precisely, we want to develop a method for answering provenance queries that refer to bibliographic entities, people, and coporate bodies as well as the relationships between those. This method should consult standard data sources such as library catalogues, authority files, and knowledge bases. The method should furthermore be implementable as a software tool that supports the user in formulating their queries, answering them, and exploring the data that supports the query answers. In the long term, we envisage that such a tool will support provenance research by prospectively retrieving potentially interesting constellations.

This goal leads to the following central research question for this thesis.

> *How can provenance relationships be modelled and automatically retrieved?*

This question implies several subordinate questions:

SQ1 *Which data sources are available for answering provenance queries?*

SQ2 *Which techniques and tools are available for the integration of data from heterogeneous sources?*

SQ3 *Based on the structure of the identified data sources, how data sources, queries, and answers be modelled in an abstract framework?*

SQ4 *What is a suitable method for retrieving provenance relationships in that framework?*

1.3 Methods and Outline

In order to answer our research questions, we will proceed as follows.

After a review of related work (Chapter 2), we will revisit in Chapter 3 the exemplary query patterns from Section 1.1, demonstrate a manual attempt at answering them, and discuss the expected quality of query answers and difficulties with this manual process. This discussion will be used as a point of reference for the subsequent considerations.

possibly adapt this paragraph later

In order to answer Questions **SQ1** and **SQ2**, we will review data sources, data models, and data integration techniques in Chapter 4. Based on the results of this analysis, we will develop an abstract model of data sources, queries, and answers in Chapter 5. This mathematical model will subsume the previous examples while being vastly more general: it gives a formal description of how to build admissible queries, without restricting their contents (i.e., the specific names, attributes, concepts, and relationships used) or their complexity. Using this model, we hope to provide a means for library science, which is the origin of the research question, to benefit from established methods from mathematics and computer science

Based on our model and the insights from the preceding analysis, we will develop in Chapter 6 a method for answering queries that can serve as the basis of a software tool as indicated in the previous section. Finally, we draw conclusions in Chapter 7.

2 Related Work

TODO: import from *Exposé* and extend

- state of provenance indexing (with authority data): [Hakelberg 2016]
- data provenance: [Eckert 2012]
- Named Entity Linking and Recognition: [Menzel, Schnaitter et al. 2021; Meiners 2022], see Exposé
- Historical Network Analysis: [Menzel, Bludau et al. 2020], see Exposé
- more from project SoNAR (IDH)?

3 Prototype Queries and Answers

As a first step towards delineating the type of queries that should be covered by our approach, we examine the query patterns introduced in Section 1.1 more closely. Those will serve as a point of reference for the analysis of the available data sources in Chapter 4, and they will be generalised by the abstract framework developed in Chapter 5. That framework will provide a rigorous definition of the queries that can be formulated within it.

3.1 Prototype Queries

The query patterns introduced in Section 1.1 are the following.

Q1 Who read work W, in which manifestation and in which year?

Q2 Which exemplars of work W were passed from one of its owners to a collaborator (or a student)?

Q3 What are the relationships between the recipients of work W (or of manifestation M of W or of exemplar C of W, respectively)?

These query patterns have been identified as important examples by personal communication with provenance researchers,[5] which justifies the choice of considering them as *prototypical*.

ask more experts (V. Petras, M. Hopp, J. Weis)

Obviously, the question arises whether these prototypical query patterns are representative for the range of queries that provenance researchers are interested in asking. Answering that question would require systematic analysis of queries relevant to or useful for researchers. Such an analysis would need to comprise an extensive interview study based on very generic questions of a predominantly open-ended nature, requiring a labour-intensive evaluation. Given that this thesis focusses on the technical prerequisites and realisation, such a study is clearly outside its scope. However, since the general framework that we will develop is informed by the available data sources and designed to cover a wide range of possible queries, it is reasonable to assume that tools developed on its basis will be helpful for provenance researchers. In subsequent work, when our method will hopefully have been implemented in a prototype tool, the extent to which researchers' needs are served can be determined by means of a more focussed user study with more specific questions, which in turn can inform possible extensions of the framework.

3.2 Manual Answer Retrieval

In order to demonstrate how a researcher could proceed (manually) when answering a query, we take Pattern **Q2** and fix work W to be the seminal work *De revolutionibus orbium coelestium* (short: *De revolutionibus*; English translation: *On the Revolutions of the Heavenly Spheres*) [Copernicus 1543] by the astronomer Nicolaus Copernicus (1473–1543). We thus consider the following concrete query.

[5] Dietrich Hakelberg, Research Library Gotha of the University of Erfurt

Q2′ Which exemplars of *De revolutionibus* were owned by some scientist who passed them on to a student?

When answering **Q2′**, an obvious way to proceed is the following: first, our researcher finds exemplars of *De revolutionibus* in online catalogues of libraries and library networks. For each such exemplar, they then inspect the provenance entries that name owners who were people (not corporate bodies). Finally, our researcher will have to find those names in databases such as authority files or Wikidata and, for each entry, explore the specified profession (“scientist”) and relationships to other people (“student”).

For example, the online catalogue (OPAC) of the Gotha Research Library of the University of Erfurt (Forschungsbibliothek Gotha) lists two printed exemplars of *De revolutionibus* [Web12]. One of those bears the signature Druck 4° 00466, and its provenance entries name the following previous owners [Web13]:

- Hieronymus Tilesius (1529–1566): autograph and date 1551
- NN: note, date 1553, name scraped out
- Johann Hommel (1518–1562), autograph
- Valentin Thau (1531–1575), note (greek proverb, possibly not denoting ownership)
- Ernest II, Duke of Saxe-Gotha-Altenburg (1745–1804): stamp/seal, initial
- Ducal Library, Gotha (a predecessor organisation of Gotha Research Library): stamp marking a duplicate
- Ernestine Gymnasium, Gotha: stamp
- Landesbibliothek Gotha: stamp

Our researcher can immediately decide that they can ignore the second entry (no name given) and the last three entries (corporate bodies). For the remaining four entries, our researcher follows the links given in the OPAC to the Integrated Authority File (GND) of the German National Library [Web14]. On inspection of these entries, it turns out that Ernest II was a regent and very probably not a scientist, and that the other three people—Tilesius, Hommel, and Thau—had professions such as theologian, mathematician, and astronomer, which qualifies them as scientists. Furthermore, Hommel’s entry contains a reference to Thau with the relationship “has student” (and Thau’s entry contains the inverse reference to Hommel). From this reference, our researcher can conclude that two scientists in the teacher-student relation have both possessed the exemplar.

Unfortunately, the data available does not imply that Hommel passed the exemplar (directly) to Thau; furthermore, the provenance entry for Thau raises doubts as to whether Thau was really an owner of the exemplar. Therefore, the retrieved data can only be regarded as a *candidate answer* which entails the hypothesis that the found exemplar was passed from Hommel to Thau. Our researcher can now engage in further research in order to verify that hypothesis.

TODO:

- Discuss instances of **Q1** or **Q3**?
- Add further prototypical query patterns?
- Explain how **Q1**–**Q3** differ from each other (extend remarks from Section 1.1)?

3.3 The Quality of Query Answers

Our simple example already illustrates that the quality of query answers strongly depends on the quality of the underlying data, regardless of whether answers are obtained manually or automatically. In particular, missing or spurious data will lead to missing or spurious answers. Before we begin to deal with automated query answering, we need to analyse the sources of missing or spurious data and their effects on the answers obtained.

For the sake of a principled discussion, we call the set of answers to a query obtained by some manual or automatic procedure an *answer set*, and we call the (set of) answers that the query has in reality the *true answers* and the *true answer set*. In the above example, the answer set obtained by the described manual process consists of the single answer "Druck 4° 00466", if we assume that our researcher interprets the data retrieved generously (i.e., as an indication that Thau might have been an owner and might have received the book directly from Hommel) and draws additional conclusions (i.e., that every mathematician or astronomer is a scientist). The true answer set is not known and will most probably never be known; it is a term of rather philosophical nature.

Ideally, both answer sets should coincide. Incomplete data may cause the answer set to exclude some true answers, and spurious data may cause it to contain some answers that are not true. We call the respective answers *missing* and *spurious*. On the basis of our example, we can identify four distinct causes for data being incomplete or spurious, as discussed in the following. We use the word *term* as an umbrella term for concepts (such as "scientist") and relations (such as "is student of").

1. In the example, there is no information available on whether any of the persons involved is indeed a *scientist*. However, information on more specific professions (e.g., theologian, mathematician, and astronomer) is available, and the information that, e.g., Thau and Hommel were scientists has to be derived based on general knowledge of the world. The same effect can occur with relations instead of concepts: if **Q2** were to ask for family members instead of collaborators or students, then the data would presumably only support more specific relations such as "has sister" or "has father", and relationships using "has family member" would need to be derived.

   More generally speaking, terms can be missing in the data sources because membership in them is implicit, for example, from membership in more specific terms. If the query uses such terms, then the answer set is always empty unless the person or machine determining the answers makes derivations based on their knowledge of the world.

   find proof in the lit./with experts? Ask DH?

   Clearly, it would not be advisable to attempt adding all implicit knowledge to the data because that would massively inflate the data, as most terms have several superordinate concepts or relations and, furthermore, implicit knowledge is not restricted to taxonomic knowledge.

   Explain this further? A case for ontologies!

2. In the example, there is no information available on whether any of the owners of the exemplar *passed it on* to another owner. Similarly, attempts to answer instances of Query Pattern **Q1** will have to deal with the problem that there is no information available as to who *read* books.

   More generally speaking, terms can be missing in the data sources because they are not recorded at all, as a consequence of general design decisions for the data source. The reasons for such decisions can be manifold: for example, relationships such as who actually *read* a book are very hard to confirm, or terms may not be part of the fixed vocabulary for a data field in a source. If the query uses such terms, then the answer set is always empty, as in

   find example?

Case 1.

3. In the example, it is possible that single owners of the exemplar or single relationships between owners have not been recorded because no evidence has been found yet. More generally, concept memberships or relationships can be missing sporadically, which may lead to missing answers.

4. In the example, if the provenance entry on Hommel is incorrect, then the answer obtained on the grounds of that entry is incorrect. More generally, spurious concept memberships or relationships may lead to spurious answers.

These cases reveal a striking qualitative difference concerning the effects of data incompleteness or spuriousness on the answer set: Cases 3 and 4 have effects on single answers only, i.e., some answers are missing or spurious. However, Cases 1 and 2 generally make *all* answers missing unless further provisions are made. In the above example, such "provisions" might include the conclusion that mathematicians etc. are scientists. and the hypothesis that Thau was an owner and received the book directly from Hommel. Getting back to the general objective of this thesis, it is appropriate to consider answers as *candidates* that necessitate (and inspire) further research. For this purpose, spurious answers (in manageable numbers) are less harmful than missing answers. Consequently, it is important to find ways to avoid missing answers without generating too many spurious answers. In other words, it is desirable to obtain an answer set that is a slight *overapproximation* of the true answer set.

3.4 From Manual to Automated Query Answering

The manual process that we have described in Section 3.2 is cumbersome, laborious, and prone to errors and omissions for several reasons: The search in library catalogues for exemplars of works and their provenances requires expert skills. Catalogues with potential matches need to be selected manually, and each catalogue needs to be queried individually, using its own search functionality and syntax. The traversal through all retrieved exemplars and the pursuit of each potential relevant provenance entry per exemplar multiplies the amount of manual work necessary. Finally, it is not clear what an effective and efficient way to "explore" relationships would be: while it is easy to find direct relationships such as "is student of" in the view for a person's entry in databases such as GND or Wikidata, there are relationships that cannot be discovered easily by hand, e.g., "P_1 and P_2 are students of the same scholar".

give examples of commonalities (OPAC) and differences (discovery vs. OPAC)?

These considerations suggest that query answering will strongly benefit from automated support, which can help reduce the amount of manual work, integrate heterogeneous data sources, incorporate background knowledge (e.g., every mathematician is a scientist), and discover relationships between entities that are not necessarily direct. We envisage a software tool that enables researchers to formulate queries and which computes answers consulting data sources selected by the user. The vocabulary used for formulating queries should be based on the vocabulary present in the data sources, but it should also include terms such as "scientist" or "read", which are not recorded in the data, as discussed in Section 3.3. The tool thus needs to implement ways to map those terms with the available vocabulary, as well as techniques for integrating data from heterogeneous sources. It is our general vision that the software tool will serve as an instrument for prospectively finding interesting candidates that inspire further research. In the remainder of this thesis, we want to lay the foundations and develop a precise method that can serve as a basis for a future implementation of this tool.

4 Analysis of Available Data Sources and Techniques

…

4.1 Data Sources

- library catalogues: OPAC, network catalogues (K10plus etc.), DNB, **KVK** (“federated search”: shallower than in single catalogues, but (sufficient) information on all copies!); further catalogues?
- authority files: GND, what else?
- KBs: Wikidata, what else?
- focus the following considerations on a narrow choice of these data sources; the general approach to be developed should be largely independent on that concrete choice
- in-depth analysis of data sources (e.g., overlap and differences) is worthwhile but must be deferred to future work

4.2 Data Models

- FRBR et al.
- **OPAC or K10plus data model as ER diagram? Binary vs. n-ary relations? $\rightsquigarrow$ check literature (possibly also papers on the FRBR model)!**

…

4.3 Data Integration and Further Techniques

- upper-level ontologies?
- ontologies on research
- FRBRoo?
- RDF, SPARQL, N-Quads?

…

4.4 Implications on Modelling

…

- **constants plus unary and binary relationships suffice (?)**

5 Modelling Provenance Relationships

In this chapter, we develop a generic approach to modelling provenance relationships. More precisely, we need to model three central notions: queries that a user may want to ask, data sources that are to be consulted in order to answer a query, and answers given to a query. In order to obtain a generic approach, we aim at providing rigorous definitions for these central concepts, and we seek intensional rather than extensional definitions. In particular, those definitions should not depend on concrete example queries or data sources such as the ones discussed in Chapters 3 and 4; neither should they depend on concrete objects, concepts, or relationships (such as “Copernicus”, “exemplar”, or “student”). Instead, we will develop an abstract model that formalises the notions of a query, data source, and answer. This model can then be instantiated with a multitude of concrete queries and data sources, and it specifies how future implementations need to proceed in order to provide answers.

As a basis for our abstract model, we choose standard concepts and techniques from graph theory. The concept of a graph is widely used in computer science and discrete mathematics; see standard textbooks [e.g., Diestel 2012]. Graphs and graph techniques are widely applied in various areas such as computer science, linguistics, physics and chemistry, social sciences, and biology [Web15]. In particular, they are used to represent large knowledge bases [e.g., Ehrlinger and Wöß 2016] and underlie the Resource Description Framework (RDF) [Web16], which is a standard of the World Wide Web Consortium (W3C) [Web17]. Furthermore, the basic definition of a graph is conceptually simple, and graph theory provides a plethora of well-understood concepts and algorithms. By utilising graph theory, our application scenario can benefit from these concepts, algorithms [Diestel 2012; Even 2012], and implementations [Web18; Web19].

Delineate from classical database theory (text snippets commented out)? This argument should be a logical consequence of the requirement analysis in the previous chapter(s)!

In the following sections, we introduce our model by defining the underlying notion of a graph and the abstract notions of data sources, queries, and answers (Sections 5.1 and 5.2). We also discuss decision procedures related to query answering in Section 5.3. Up to that point, the model remains very basic, which is a deliberate choice in order to make it conceptually comprehensible. In particular, although the definitions are elementary and self-contained from a mathematical point of view, we provide additional explanations and illustrations for the sake of readers with little or no background in mathematics. In order to obtain a more flexible and comprehensive model, we discuss our modelling decisions in Section 5.4 and possible extensions in Section 5.5.

5.1 Labelled Directed Graphs

The basic components of a graph are nodes and edges. Edges link nodes, and they can be directed or undirected. Graphs are easy to visualise; nodes are typically represented as circles or rectangles, and edges as arrows (directed) or lines (undirected). Figure 5.1 gives an abstract example of a directed graph.

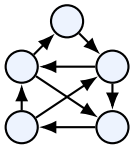

Figure 5.1: A directed graph

For our purposes, nodes represent objects or literals. Objects

include, e.g., works, expressions, manifestations, items, persons, or corporate bodies; literals include, e.g., publication years, birth years, or identifiers. Edges represent relationships between those objects or literals, and those relationships are typically directed: e.g., has_owner points *from* an item *to* a person or corporate body, whereas is_owner_of points into the opposite direction. Therefore we use *directed* graphs. Symmetric relationships, such as collaborates_with, can be represented via two edges in both directions.

Furthermore, we want to assign a unique name to each node of a graph and one or several labels to each node and each edge: The name of a node specifies the *object* that is represented by that node. The labels of a node specify the *concepts* of which that node is an *instance*. For example, a node representing the physicist Albert Einstein may be labelled, among others, with the concepts Person, Scientist, and Physicist. The labels of an edge specify the *relations* of which the pair of nodes represented by that edge is an instance. For example, if a person p_1 has a student p_2 and, in later life, collaborates with p_2, then this can be represented via an edge from p_1 to p_2 with the label {has_student, collaborates_with} (and/or an edge from b to p with the label {is_student_of, collaborates_with}). These considerations lead us to a straightforward extension of the notion of a directed graph: a *labelled directed graph*.

Refer to the literature for labelled graphs? Refer to description logic terminology regarding the terms “object”, “concept”, “relation”, *instance*, etc.? Distinguish clearly between “relation” and “relationship”.

In order to visualise a labelled directed graph, node names are written into the respective node, and node and edge labels are written next to the node or edge. Multiple labels of the same node or edge are delimited with commas. An example is given in Figure 5.2. The shown graph represents a part of the data described in Section 3.2 concerning an exemplar of Copernicus’ *De revolutionibus* at the Gotha Research Library *(FB Gotha)*. It contains a node for the work (labelled with the FRBR entity Work), a node for the exemplar (labelled with the FRBR entity Item), and nodes for the author and two of the owners (labelled with their professions according to their GND entries). For the sake of this example, the owners are additionally labelled with the profession Scientist, which is implicit in the real data. The ten edges represent the FRBR relationships between work and item, the creator relationship between work and author, the owner relationship between exemplar and two owners, student and collaboration relationships between the owners, and the converses of these relationships. For the sake of simplicity, the graph deviates from the FRBR model [IFLA SG FRBR 2009] by omitting the FRBR entities “Expression” and “Manifestation” that should occur between the nodes labelled “Work” and “Item”.

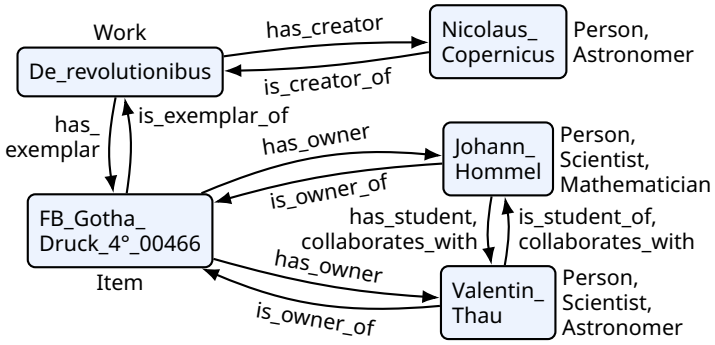

Figure 5.2: A labelled directed graph that represents data concerning an exemplar of Copernicus’ *De revolutionibus* and some of its owners

As we will see in the following, labelled directed graphs can be used in our setting to represent (combinations of) data sources as well as queries. They allow us to draw on standard notions from

graph theory and query answering in order to define admissible query answers and to devise methods for obtaining those.

The above explanations can be cast into a rigorous mathematical definition, which uses sets to represent nodes, a binary relation over the set of nodes to represent edges, and functions over the nodes and edges to represent names and labels. In order for the range of those functions to be well-defined, the definition of a labelled directed graph is relative to a namespace which contains all the names of objects, concepts and relations that are relevant. The contents of this namespace is arbitrary and may consist, for example, of all the names found in the relevant data sources. The following definition introduces the notions of a namespace and a labelled directed graph.

Definition 1. Let $\mathsf{N} = (\mathsf{N_O}, \mathsf{N_C}, \mathsf{N_R})$ be a *namespace* consisting of a set $\mathsf{N_O}$ of *object names*, a set $\mathsf{N_C}$ of *concept names*, and a set $\mathsf{N_R}$ of *relation names*. A *labelled directed graph over* N is a triple $G = (V, E, \mathcal{N}, \mathcal{L})$ where

- V is a set, whose members are called *nodes*;[a]
- $E \subseteq V \times V$ is a set of pairs of nodes, whose members are called *edges*;
- $\mathcal{N} : V \to \mathsf{N_O}$ is an injective function that assigns to each node a unique object (called the node's *name*);
- $\mathcal{L} : V \cup E \to \mathsf{N_V} \cup 2^{\mathsf{N_R}}$ is a function that assigns to each node a set of concept names (called the node's *labels*) and to each edge a non-empty set of relation names (called the edge's *labels*); we call $\mathcal{L}$ a *labelling function*.

[a] In classical graph theory, nodes are called *vertices*; thus the set of nodes of a graph is denoted by V. We adopt the denotation V for conformity and the more modern term "node" for brevity.

Definition 1 formalises the following commitments regarding names and labels. (1) Every node has a unique name, and no two nodes have the same name (the latter being ensured by injectivity). (2) A node can have an arbitrary number of labels, including no label (in case the node belongs to no concept). (3) An edge can have an arbitrary number of labels, but that number must not be zero – the effect of an edge having no labels can be achieved by simply omitting that edge.

In order to illustrate the components of Definition 1, we refer to the graph depicted in Figure 5.2: V consists of five nodes, and E of ten edges (each single arrow constitutes an edge since the direction matters). There are, among others, the following node names and labels:

- The node at the top left has the name De_revolutionibus and the single label Work.
- The node at the bottom right has the two labels Person and Astronomer.
- The edge from the node named Johann_Hommel to the node named Valentin_Thau has the two labels has_student and collaborates_with, and the edge pointing into the converse direction has the labels is_student_of and collaborates_with.

5.2 Modelling Data Sources, Queries, and Answers

We can now use our notion of a labelled directed graph to model data sources and queries, and to obtain a rigorous definition of a query answer.

5.2.1 Data Sources

Data sources correspond exactly to our notion of a graph.

Definition 2. A *data source over the namespace* $\mathsf{N} = (\mathsf{N_O}, \mathsf{N_C}, \mathsf{N_R})$ is a labelled directed graph over N.

In our model, we assume that there is always a *single* data source against which a query is posed and evaluated. When the model is applied to real-word queries and data sources, the abstract notion of a data source is instantiated by the union of all available concrete data sources (such as catalogues, authority files, knowledge bases), including mappings between them if applicable.

5.2.2 Queries

In order to model queries based on the same notion of a graph, we need to distinguish two special groups of nodes that act as placeholders (1) for the object(s) after which the query asks and (2) for further objects that are mentioned in the query without being named explicitly. For example, consider Query **Q2′** from Section 3.2:

Q2′ Which exemplars of *De revolutionibus* were owned by some scientist who passed them on to a student?

To model this query, we do not only need a node representing the work *De revolutionibus*, but also a node representing an exemplar that satisfies the conditions stated in the query and whose name is asked for (Group 1), and two nodes representing the owner and their student (Group 2). Since these three individuals are not known, we need to use *variables* for naming them. Query **Q2′** can then be modelled by the graph shown in Figure 5.3.

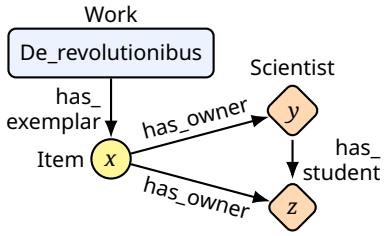

Figure 5.3: A graph representing example query **Q2′**

The nodes of this graph fall into three groups:

(1) The node named De_revolutionibus represents that work;

(2) Node x falls into Group 1 as explained above;

(3) Nodes y and z fall into Group 2 as explained above.

Node names x, y, z are the variables mentioned above, and we call x the *answer variable* and y, z the *anonymous variables* of this query. From now on, we fix two sets $\mathsf{VAR_{ANS}}$ and $\mathsf{VAR_{ANON}}$ of *answer variables* and *anonymous variables*, respectively, and we assume that they both contain a countably infinite number of elements—which simply ensures that there is an unlimited supply of variables. We furthermore require that these two sets are disjoint with each other and with any set $\mathsf{N_O}$ of object names. Thus, according to Definition 2, graphs representing data sources cannot use any variables as node names.

In order to allow queries to use variables, the following definition of a query is now immediate.

Definition 3. A *query over the namespace* (N_O, N_C, N_R) is a labelled directed graph over $(N_O \uplus VAR_{ANS} \uplus VAR_{ANON}, N_C, N_R)$.

The operator "$\uplus$" used in this definition stands for *disjoint union*, i.e., "$A \uplus B$" stands for the union of the *disjoint* sets A, B. The wording of the definition also ensures that we do not have to mention variables explicitly when specifying the namespace of a query.

5.2.3 Query Answers

Based on the representation of both data sources and queries as graphs, we can now define the notion of an answer to a query with respect to a data source. For this purpose, it is important to realise that, typically, a query is a small graph and a data source is a large graph, and that finding answers simply means finding small parts of the large graph that have the same structure as the small graph. For the example query given in Figure 5.3, this means that every subgraph of the data source consisting of four nodes with the same edges and labels should be an answer. Regarding the previous example, the subgraphs given in Figure 5.4 (a, b) constitute answers, while the subgraphs in Figure 5.4 (c, d) do not: Subgraph (a) is identical to the query graph with the only exception that it contains proper objects instead of variables in the node names; Subgraph (b) is the same graph but extended with additional information; Subgraph (c) lacks the edge labelled has_student between the two owners; Subgraph (d) resembles the structure of the query but does not contain the required node named De_revolutionibus.

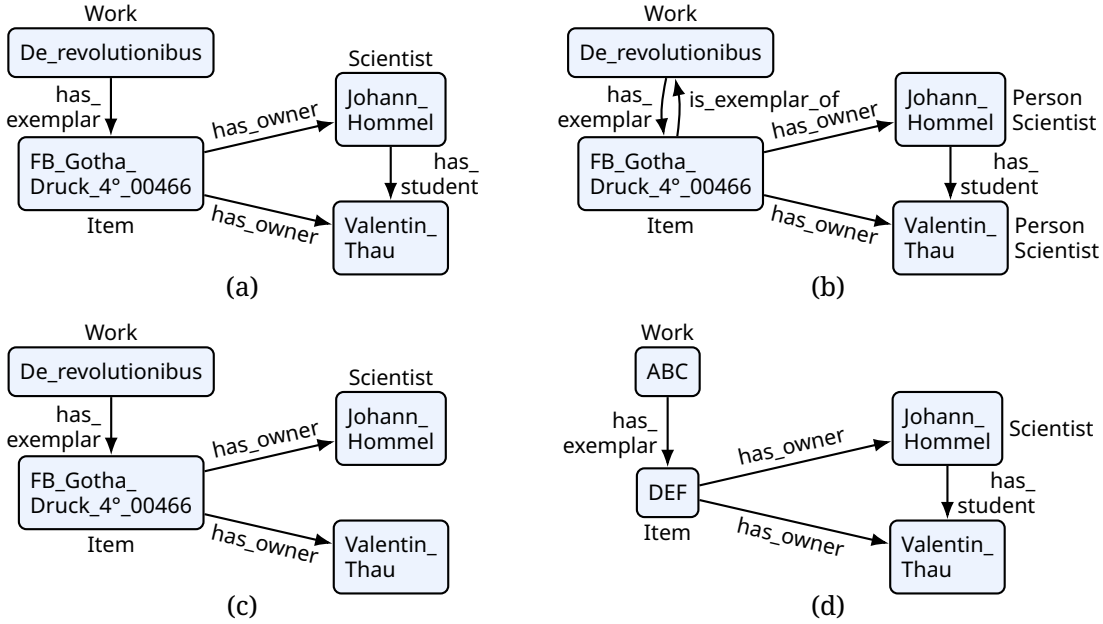

Figure 5.4: Positive (a, b) and negative (c, d) examples for query answers

In order to identify subgraphs of a given graph that have the same structure as another given graph, we use the notion of a homomorphism. A homomorphism is a function that maps some object to another while preserving the structure of the former. We therefore need to define a variant of homomorphisms that maps queries to data sources. This variant is given in the following.

Definition 4. Let $\mathsf{N} = (\mathsf{N_O}, \mathsf{N_C}, \mathsf{N_R})$ be a namespace, $G = (V, E, \mathcal{N}, \mathcal{L})$ a query over N, and $G' = (V', E', \mathcal{N}', \mathcal{L}')$ a data source over N. A *homomorphism from G to G'* is a map $h : V \to V'$ that satisfies the following properties.

H1 $\mathcal{N}(v) = \mathcal{N}'(h(v))$ for every node $v \in V$ with $\mathcal{N}(v) \in \mathsf{N_O}$.

H2 $\mathcal{L}(v) \subseteq \mathcal{L}'(h(v))$ for every node $v \in V$.

H3 $\mathcal{L}(v_1, v_2) \subseteq \mathcal{L}'(h(v_1), h(v_2))$ for every edge $(v_1, v_2) \in E$.

If h is a homomorphism from G to G', we write $h : G \to G'$. If there is some homomorphism from G to G', we write $G \lesssim G'$.

Property **H1** requires that a homomorphism maps each node in G that is named with an object to that node in G' which is named with the same object. Nodes named with variables in G can be mapped to arbitrary nodes in G'. Properties **H2** and **H3** require that homomorphisms preserve node and edge labels; more precisely, the image of a node (or edge) under h must have *at least* the same labels (and may have additional labels).

Figure 5.5 shows a homomorphism h (dashed lines) from the query depicted in Figure 5.3 to the graph from Figure 5.2.

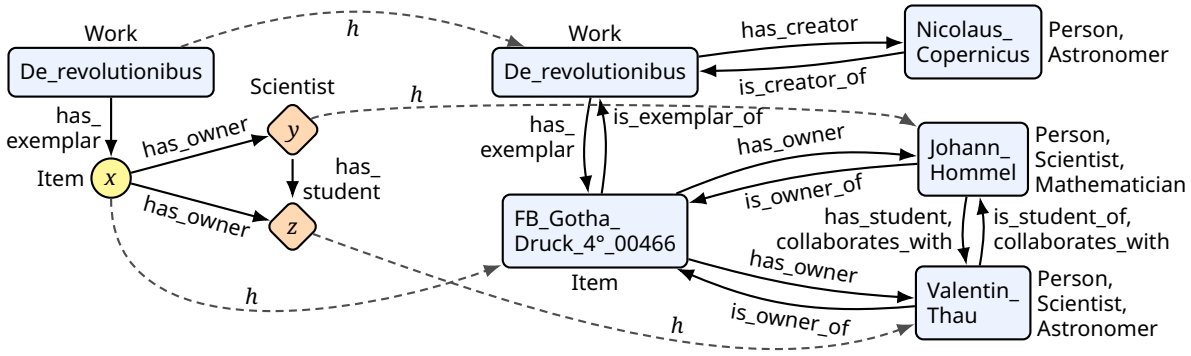

Figure 5.5: An example homomorphism

5.3 Decision Procedures

Formulate decision problem, discuss (data) complexity & algorithms. (Reduction to FO/SQL?)

5.4 Discussion of the Modelling Decisions

Relate this “machinery” to the prototype queries. In particular:

- Comment on Boolean queries if necessary.
- Discuss specific requirements for modelling **Q1** and **Q3**:
  - **Q1** seems to require answer variables representing sets (the owners) and an appropriate extension of the definition of a homomorphism;

  - the same holds for **Q3**; additionally the answer should include the relationships between the images of the answer variables (the relationships between the owners), i.e., some sort of spanned subgraph

⇝ extensions needed; description or sketch of these and other extensions in the next section

5.5 Possible Extensions

TODO: Discuss further extensions:

- relations of arbitrary arity
- data provenance
- concrete values (“year of publication”)
- attributes on relationships:
  - sketch idea: e.g., provide year for relationship has_owner – example: “passed on to” requires descending year numbers *and* no successor with intermediate year number
  - solution: quads instead of triples (in RDF speak); add attributes to the graph model? (Markus Krötzsch’s work?)
  - explain difficulties: more complex formal machinery (def. of graphs, queries, and matches); missing data, e.g.:
    * From which year *to which year* did person X own item Y?
    * Was person Z a student of person X’s *at the point in time when X passed the item on to Z*?
  - discuss usefulness: false positives due to incomplete data as discussed in Section 3.3 ⇝ manual inspection is necessary anyway; attributes may still help hide false answers

6 Automated Retrieval of Provenance Relationships

- develop method for answering queries in the model just defined

7 Conclusion

- get back to initial research question and its subordinate questions
- Acknowledgements? See comments.

References

Bibliography

Copernicus, Nicolaus (1543). *Nicolai Copernici Torinensis De revolutionibus orbium coelestium, Libri VI*. Latin. Ed. by Andreas Osiander. Nürnberg: Petreius, Johann, [6], 196 sheets. URL: https://opac.uni-erfurt.de/LNG=EN/DB=1/PPNSET?PPN=567506266 (cit. on pp. 2, 6).

Diestel, Reinhard (2012). *Graph Theory, 4th Edition*. Vol. 173. Graduate texts in mathematics. Springer. ISBN: 978-3-642-14278-9 (cit. on p. 11).

Eckert, Kai (2012). "Metadata Provenance in Europeana and the Semantic Web". German. Master's thesis. Humboldt-Universität zu Berlin. DOI: http://dx.doi.org/10.18452/14172 (cit. on p. 5).

Ehrlinger, Lisa and Wolfram Wöß (2016). "Towards a Definition of Knowledge Graphs". In: *Joint Proceedings of SEMANTiCS 2016 and SuCCESS 2016, Posters and Demos Track*. Ed. by Michael Martin, Martí Cuquet and Erwin Folmer. Vol. 1695. CEUR Workshop Proceedings. CEUR-WS.org. URL: https://ceur-ws.org/Vol-1695/paper4.pdf (cit. on p. 11).

Even, Shimon (2012). *Graph algorithms*. Ed. by Guy Even and Richard M. Karp. 2nd. Literaturangaben. Cambridge [u.a.]: Cambridge Univ. Press. XII, 189. ISBN: 9780521517188 (cit. on p. 11).

Gingerich, Owen (2002). *An annotated census of Copernicus' "De revolutionibus" (Nuremberg, 1543 and Basel, 1566)*. Studia Copernicana. Leiden: Brill. 402 pp. ISBN: 9004114661 (cit. on p. 2).

Hakelberg, Dietrich (2016). "Herkunft finden und vernetzen: Stand und Perspektiven der Provenienzerschließung mit Normdaten". German. Master's thesis. Humboldt-Universität zu Berlin (cit. on pp. 1, 3, 5).

IFLA Study Group on the Functional Requirements for Bibliographic Records (Feb. 2009). *Functional Requirements for Bibliographic Records. Final report*. Ed. by International Federation of Library Associations and Institutions (IFLA). München. URL: https://repository.ifla.org/handle/123456789/811 (cit. on pp. 2, 12).

Meiners, Ole (2022). "Evaluation von Named Entity Recognition-Modellen für historische deutschsprachige Texte am Beispiel frühneuzeitlicher Ego-Dokumente". German. Bachelor's thesis. Humboldt-Universität zu Berlin. 62 pp. (cit. on p. 5).

Menzel, Sina, Mark-Jan Bludau et al. (2020). "Graph Technologies for the Analysis of Historical Social Networks Using Heterogeneous Data Sources". In: *Graph Technologies in the Humanities 2020 (GRAPH 2020)*. Vol. 3110. CEUR Workshop Proceedings, pp. 124–149 (cit. on p. 5).

Menzel, Sina, Hannes Schnaitter et al. (2021). "Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten". In: *Qualität in der Inhaltserschließung*. Ed. by Michael Franke-Maier et al. Berlin, Boston: De Gruyter Saur, pp. 229–258. ISBN: 9783110691597. DOI: doi:10.1515/9783110691597-012. URL: https://doi.org/10.1515/9783110691597-012 (cit. on p. 5).

Salatowsky, Sascha and Karl-Heinz Lotze, eds. (2015). *Himmelsspektakel. Ausstellung der Universitäts- und Forschungsbibliothek Erfurt/Gotha*. German. Veröffentlichung der Forschungsbibliothek Gotha 52. Gotha: University und Research Library Erfurt/Gotha. 231 pp. (cit. on p. 2).

Web Resources

[Web1] German National Library. *The Integrated Authority File (GND)*. URL: https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd.html (visited on 30/03/2023) (cit. on p. 1).

[Web2] Wikipedia. *Bibliotheksverbund: Deutschland*. German. URL: https://de.wikipedia.org/wiki/Bibliotheksverbund#Deutschland (visited on 30/03/2023) (cit. on p. 1).

[Web3] Bibliotheksservice-Zentrum Baden-Württemberg and Gemeinsamer Bibliotheksverbund. *K10plus – Kooperationsprojekt BSZ und GBV*. German. URL: https://www.bszgbv.de/services/k10plus/ (visited on 30/03/2023) (cit. on p. 1).

[Web4] Gemeinsamer Bibliotheksverbund. *PICA-Format*. German. URL: https://format.gbv.de/pica (visited on 30/03/2023) (cit. on p. 1).

[Web5] Bibliotheksservice-Zentrum Baden-Württemberg and Gemeinsamer Bibliotheksverbund. *Exportformate*. German. URL: https://wiki.k10plus.de/display/K10PLUS/Exportformate (visited on 30/03/2023) (cit. on p. 1).

[Web6] OCLC, Inc. *WorldCat®*. URL: https://www.worldcat.org/ (visited on 31/03/2023) (cit. on p. 1).

[Web7] ISNI International Agency. *ISNI*. URL: https://isni.org/ (visited on 31/03/2023) (cit. on p. 1).

[Web8] OCLC, Inc. *VIAF*. URL: https://viaf.org/ (visited on 31/03/2023) (cit. on p. 1).

[Web9] Wikimedia Foundation. *Wikidata*. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page (visited on 31/03/2023) (cit. on p. 1).

[Web10] Wikipedia. *Georg Joachim Rheticus*. URL: https://en.wikipedia.org/wiki/Georg_Joachim_Rheticus#External_links (visited on 31/03/2023) (cit. on p. 2).

[Web11] Wikidata. *Georg Joachim Rheticus*. Data Set Q93588. URL: https://www.wikidata.org/wiki/Q93588 (visited on 31/03/2023) (cit. on p. 2).

[Web12] University of Erfurt. *OPAC search result for "De Revolutionibus"*. URL: https://opac.uni-erfurt.de/DB=1/CMD?ACT=SRCHA&IKT=1016&SRT=YOP&TRM=tit+de+revolutionibus+and+per+kopernikus+and+jah+15**+and+bbg+a* (visited on 31/03/2023) (cit. on p. 7).

[Web13] — *OPAC entry for one exemplar of "De Revolutionibus"*. URL: https://opac.uni-erfurt.de/LNG=EN/DB=1/XMLPRS=N/PPN?PPN=567506266 (visited on 31/03/2023) (cit. on p. 7).

[Web14] German National Library. *DNB Catalogue*. URL: https://katalog.dnb.de/EN/home.html?v=plist (visited on 31/03/2023) (cit. on p. 7).

[Web15] Wikipedia. *Graph Theory: Applications*. URL: https://en.wikipedia.org/wiki/Graph_theory#Applications (visited on 06/04/2023) (cit. on p. 11).

[Web16] W3C – World Wide Web Consortium. *Resource Description Framework (RDF)*. URL: https://www.w3.org/RDF/ (visited on 08/04/2023) (cit. on p. 11).

[Web17] — *W3C Homepage*. URL: https://www.w3.org/ (visited on 08/04/2023) (cit. on p. 11).

[Web18] Python Software Foundation. *Python Graph Libraries*. URL: https://wiki.python.org/moin/PythonGraphLibraries (visited on 15/04/2023) (cit. on p. 11).

[Web19] Naveh, Barak and Contributors. *JGraphT: a Java library of graph theory data structures and algorithms*. URL: https://jgrapht.org/ (visited on 08/04/2023) (cit. on p. 11).

Humboldt-Universität zu Berlin
Philosophische Fakultät
Institut für Bibliotheks- und Informationswissenschaft

Name: Schneider Vorname: Thomas

Matr.Nr.: 624025

Eidesstattliche Erklärung zur

- ○ **Hausarbeit ***
- ○ **Bachelorarbeit ***
- ◉ **Masterarbeit ***
- ○ **Abschlussarbeit im Bibliotheksreferendariat ***

***** Die eingereichte PDF-Datei ist mit den Printexemplaren identisch.

Ich erkläre ausdrücklich, dass es sich bei der von mir eingereichten schriftlichen Arbeit mit dem Titel

Modelling and Automated Retrieval of Provenance Relationships

um eine von mir erstmalig, selbstständig und ohne fremde Hilfe verfasste Arbeit handelt.

Ich erkläre ausdrücklich, dass ich *sämtliche* in der oben genannten Arbeit verwendeten fremden Quellen, auch aus dem Internet (einschließlich Tabellen, Grafiken u. Ä.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos sowohl bei wörtlich übernommenen Aussagen bzw. unverändert übernommenen Tabellen, Grafiken u. Ä. (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen bzw. von mir abgewandelten Tabellen, Grafiken u. Ä. anderer Autorinnen und Autoren (Paraphrasen) die Quelle angegeben habe.

Mir ist bewusst, dass Verstöße gegen die Grundsätze der Selbstständigkeit als Täuschung betrachtet und entsprechend der Prüfungsordnung und/oder der Fächerübergreifenden Satzung zur Regelung von Zulassung, Studium und Prüfung (ZSP-HU) geahndet werden.

Datum …….. Unterschrift ……