

Knowledge Discovery for Provenance Research on Colonial Heritage Objects

Sarah Binta Alam Shoilee¹

¹Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Abstract

Heritage institutions hold rich information on cultural heritage objects involving contextual information about people, places, times, and events. This information is usually kept in institutional silos, where domain researchers often work with data across institutions. Linking entities among different institutions can enrich these data sources and, in turn, aid domain research. The aggregated version of data can be further used to infer insightful knowledge that can excel in one of the time-consuming tasks of the domain, which is provenance research. This research will first focus on entity linking across institutions to construct a Knowledge Graph representing both structured metadata of objects and the collector's biography. This work aims to use this newly formed Knowledge Graph to find interesting patterns to scale-up provenance research and analyse the effect of adding such information to the current data source. Experiments with the different modalities of data and pattern mining techniques will reveal to which extent this data enrichment places a role in finding useful knowledge for the heritage objects' provenance research.

Keywords

Knowledge Graph, Heritage Object, Entity Linking, Pattern Mining, Knowledge Discovery, E-humanities

1. Introduction

Knowledge acquisition and knowledge representation are active research fields within computer science, and countless formalisms have been developed to deal with various types of knowledge (see [1] and [2]). The last 10 years have seen a considerable level of rising popularity of Semantic technologies, including Linked Data in the cultural heritage domain. This has resulted in many heritage institutions' datasets, as well as structured vocabularies and ontologies, becoming published on the Semantic Web (for example, [3], [4]). In addition, recent advances in the modelling of data and object provenance have contributed to a better match to digital humanities needs around source and tool criticism [5].

One of the fundamental problems in the cultural heritage domain is having limited information about objects' biographies. In today's world, it is not enough to represent a heritage object only from the institutional perspective; rather, the representation should be more informative about its past or original purpose. Therefore, it is common among museum professionals to retrieve information about a particular object's biography through provenance research. From

Doctoral Consortium at ISWC 2022 co-located with 21st International Semantic Web Conference (ISWC 2022)


✉ s.b.a.shoilee@vu.nl (S. B. A. Shoilee)

🌐 <https://shoilee.github.io/Profile/> (S. B. A. Shoilee)

🆔 0000-0001-9458-8105 (S. B. A. Shoilee)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

this research, professionals try to establish a connection between objects and their past based on historical patterns, literature studies, or evidence from the past [6].

When hundreds of thousands of objects in museum premises need further provenance information, it is counter-productive for provenance researchers to search through individual data sources and come up with links for individual objects. On the other side, due to the recent collaborative commitments and initiatives, more and more such resources are being available online with an ambition of sharing collection on the open web [7]. The release of linked data from heritage institutions allows the opportunity to bring information from different data sources into a single Knowledge Graph (KG); which in turn can be used to find interesting, new knowledge to guide provenance research.

Concatenating data from multiple sources into a single Knowledge Graph allows understanding nodes association better and may help identify new knowledge through mining patterns from existing data. For example, analysing associations among different entities, perhaps among collectors, their participation in historical events, and object acquisition trends may reveal new information. Moreover, the recent development of graph embedding techniques ([8], [9]) allows us to infer complex inductive knowledge from existing graph data. The availability of more and more open linked data from heritage institutions makes it ideal to use such inference techniques on those data sets to extract new information about heritage objects that may provide a new lens in traditional provenance research.

However, while dealing with colonial heritage objects' datasets, the incompleteness in object's data, ambiguous mention in attribute values and subjective views placed in metadata creation make it a challenging problem for data linking and new knowledge discovery. This work investigates the challenges of entity linking on historical data sources to enrich heritage objects KG. In this graph, knowledge discovery techniques will be adopted to infer usable, new knowledge for provenance research. Experimenting with different pattern mining techniques and accessing acceptability will determine the practical approach for finding useful domain knowledge.

1.1. Colonial Heritage Object Metadata Challenge

Incompleteness, ambiguity and subjective views in metadata should be considered when dealing with colonial heritage objects. In addition, historical events play a part in categorising or listing information around collected objects metadata. This work identifies three issues with the current dataset, which are below.

Incompleteness Object information in the museum database starts from its acquisition date. While traditional provenance research requires considerable human resources, it also emphasises imposing institutional bias in prioritising objects. Which object to choose for further research has always been a bureaucratic choice; therefore, some objects are information-rich, whereas others suffer from a lack of information, resulting in a skewed dataset.

Ambiguity As with other historical data, dealing with museum databases of different institutes also comes with the challenge of inconsistent mention of the same entity. For example, someone named "Lars Erikson" in one database may be listed as "L. Erikson" or just "Eriksson" in another database. In addition, the date and location attribute value sometimes does not always have a precise specification. This inherent ambiguity makes it a challenging problem to

tackle such data.

Subjective view Heritage object metadata is a purposeful creation of museum histories, and perspectives [10]. Determining categories in classification is related to the historical, social, and cultural context in which the classification scheme is created and used [11]. Unfortunately, when it comes to linking data from multiple museums, those classification labels do not always mean the same thing [11] which is hard to manage in the automated mining process.

1.2. Use Case: Pressing Matter

This research will be done as part of the project “Pressing Matter” which investigates ownership, value and the question of colonial heritage in museums. This research aims to provide a solution that helps the domain researcher prioritise their investigation by bringing large-scale information closer in a manageable manner.

2. Related Work

This section will explain the research context and the relevant works from the literature. The topics mentioned here are relevant to the research questions mentioned in section 3 and position the proposed work in different topic areas.

2.1. Linked Data and Knowledge Graph in e-humanities

Knowledge organisation systems have a long history in the museum world, where they have been employed in metadata descriptions to arrange objects and increase findability. In addition, scholarly efforts have been made to create authoritative data to describe a particular group of objects, resulting in taxonomies, vocabularies, and thesaurus. The advent of Linked Data technologies offers an opportunity for the institutional data silos to enter the realm of the World Wide Web [12].

In the spirit of the Linked Data vision, several data standards and collective commitments have been made to allow cross-institutional heritage data linking. Europeana Data Model (EDM)¹, CIDOC-CRM² and Union List of Artist Names (ULAN) and Thesaurus of Geographic Names (TGN)³ are just some of the examples [7]. Adhering to the linked-data principle presents the opportunity to go beyond just the institutional databases and excel possibility of research. The diversity and heterogeneity in cultural object metadata encourage the use of Knowledge Graphs that can hold information about places, people, concepts and organisations while bringing context to the cultural heritage objects [13] [14].

While significant effort has been made to develop a suitable data model for cultural heritage data, it remains less explored whether such data can be further used for knowledge discovery [15]. Furthermore, due to different practices in curating metadata descriptions, meta-tags vary significantly across different institutions, even when describing the same object. It creates a

¹<https://pro.europeana.eu>

²<https://cidoc-crm.org>

³<https://www.getty.edu/research/tools/vocabularies/>

significant challenge when linking and analysing different object collections from multiple institutions, which further needs attention from the Linked Data community.

2.2. Provenance Research

According to the Getty Research Institute (GRI), *A complete provenance provides a documented history that can be used to verify ownership, assign the work to a known artist, and establish the legitimacy of the work of art*⁴. This quest for object biography is often based on sources, i.e., digital collection register of museum system(s), public and private archives, online search, literature review, object research, experts' input, etc.; where the significant amount of these sources are already digitised.

In the context of Pressing Matter, provenance research on the colonial object is approached by untangling the complex acquisition history through the collector's biography. This approach is adopted based on Actor-network-theory, where objects are not only singular entities but in relational configuration with other actors (people, place, time, event) that inform their biographies. The project identifies four overlapping collecting modes through which colonial objects have entered museums and will use them as a frame to guide understanding of the colonial object's changing ownership, values, and potentiality.

2.3. Entity Linking

The problem of Entity Linking can be divided into two parts: surface form extraction and named entity disambiguation (NED). Surface form extraction relates to identifying entities from a continuous span of text. The goal of the NED task is to link a named entity to ground truth entities in a knowledge base [16]. A plethora of automated entity disambiguation techniques have been proposed ranging from rule-based approaches to node-embedding-based approaches (current state-of-the-art for entity disambiguation [17] and [18]). Given that the notion of identity can change under different contexts, the task at hand always influences how entity disambiguation algorithms are built.

Entity Linking on Knowledge Graph of historical data comes with a significant challenge, i.e., attributes may have value approximation (e.g., approximate date), lack of naming standard between datasets, attribute that look similar may not mean the same thing, error-prone attribute values etc. Baas et al. mentioned some of these problems in their work and projected on other task-focused Entity Linking approaches in digital humanities [19]. While Baas et al. [19] considered neighbourhood information-based embedding to cluster similar nodes, the other literatures on Entity Linking in the context of digital humanities primarily used deterministic rules based on context to tackle such tasks; therefore not applicable when the task or entity type changes.

Though the nature of the data in [19] is similar to our current context for Entity Linking, the data setup is entirely different. In the mentioned work, the entity considered for linking has a substantial amount of property value on them, whereas in our case, there is little to no information on the target entity, i.e., collectors. In our primary dataset from museum, it is common to have more information on related objects and events rather than on the collector

⁴<https://www.getty.edu/research/tools/provenance/>

instances. Therefore, it is a matter of investigation if existing approaches of entity linking are still suitable in the given context.

2.4. Knowledge Discovery in Database

The proposed work can best be placed in Knowledge Discovery in Database (KDD). According to Fayyad et al., *KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* [20]. KDD entails a number of phases, including data preparation, pattern search, knowledge evaluation, and refining. Finding intelligible patterns that might be perceived as helpful or interesting knowledge is a priority for KDD. Patterns are statements which describe interesting relationships among a subset of the analysed data, typically resulting from a data mining process (classification, cluster-mining, association rules mining and so on).

Background knowledge extraction using linked data is an example of graph data mining [21] where data patterns are explained with linked data. Another example of graph mining is using background knowledge to traverse through the network to find new information [22]. In digital humanities, both approaches are interesting, given that there is an abundance of both data and experts' knowledge. For this reason, this work will explore a mixed-method approach where data mining techniques and experts' background knowledge will be incorporated together to find new knowledge to aid provenance research.

Heritage object metadata contains complex contextual information about the object, making KG an effective representational means. Nevertheless, it remains an open question how much they support relational learning models in the cultural heritage domain [15], which are known to provide high scalability and accuracy among other domains. Lately, a small body of work ([23], [24]) is emerging on using machine learning or data mining techniques for humanities problem-solving, which had a deep tradition of scepticism towards quantitative and empirical techniques among humanists for long. Nonetheless, a radical departure from previous methods of humanistic inquiry has been proving to bring new perspectives and scalable solutions for domain researchers. Given the lack of related work and possible usefulness in the domain, it is worth investigating contemporary pattern mining techniques in the current problem setting.

3. Problem Statement

This work will investigate if data enrichment of colonial heritage objects has the potential to provide aid in scaling up provenance research. It will take a case-based approach to explore the hypothesis that the collector's biography information might shed more light on the acquisition modes for heritage objects. This will be concluded by answering the three following **research questions (RQ)**.

RQ1 *How to link entities from multiple heritage institutions where ambiguous mention of entity is present?* To extend heritage objects' structured metadata with information on actor, time, place and events, we need to enrich data from multiple sources. However, linking data from historical datasets comes with the challenge of entity disambiguation. This work will investigate if we can use the existing state-of-the-art approach or if there is a need to develop a new algorithm.

RQ2 *Which patterns exist in colonial heritage objects' data in a museum that is actionable and useful for provenance research?* There are several pattern mining techniques available for finding interesting patterns from structured graph data. It is yet to determine which data mining technique to apply in the given task, which modality of data to consider while avoiding institutional bias and most importantly, which patterns are useful for provenance research.

RQ3 *What is the effect in the result of the dominant pattern from the Knowledge Graph when it is populated with collectors' biography information?* This study will examine if collectors' biographies and modes of object acquisition can reveal useful patterns. It will also examine how these patterns align with the current classification scheme. Finally, the results will be used to guide a revised classification scheme based on the context of objects' acquisition modes.

4. Research Methodology and approach

The overall approach of this research is based on entity linking, pattern search, pattern evaluation and assessing usefulness. Having defined the research questions, we will now address the research approaches in more detail:

RQ1 One primary task for data enrichment from different sources is to match entities across different data sets. To address the challenge of Entity Linking where ambiguous and duplicate mention of Entity exists, this work will begin its investigation from the naive string matching technique and then move towards intelligent entity matching based on vector space embedding([17], and [18]) and analyze to what extent they apply to cases where limited to no text around the Entity is available. Moreover, this research will explore whether contextual information from a Knowledge Graph can be used to guide an automated system to find possible matches across data sources.

RQ2 This work will explore data mining techniques and the explainability of the found patterns. It will also examine different data modalities to avoid potential biases in the dataset. In parallel, this research will assess which patterns are interesting for domain experts and useful for their provenance research.

There will be two parallel research works to answer the RQ2. One will explore the suitability of pattern mining techniques considering explainability in domain use. The other will explore which data modality to consider, i.e., KG only with objects' structured metadata, KG with extracted information from a text description, KG including literal node values etc. As an input, these two works will consider the entire object collection from the museum while experimenting with the different modalities of these collections and will produce actionable patterns as an output. The pattern validity will be measured based on user satisfaction.

RQ3 This research question will explore how much the newly found dominant patterns from the resultant Knowledge Graph from RQ1 differ from the found patterns from RQ2. The previous research question tries to find the dominant pattern based on the data sources only from one museum. The current research question will explore the effect of the data enrichment in pattern-finding when the data graph is enriched with collectors' biography. Thus, it will experiment with two different versions of graph data. Emphasis will be given to the clustering methods to determine whether the objects can be grouped based on acquisition modes.

5. Evaluation Plan

In the following section, we will discuss how to evaluate the results.

RQ1 The Entity Linking algorithms will be evaluated based on the experts' assessment. A small statistically significant portion of the automated established link by each algorithm will be randomly chosen for further assessment by a group of domain experts. The result will be communicated as precision and recall of the system based on their findings overlapping with experts.

RQ2 User studies will be designed to understand the acceptability of found pattern. Designing user experiments is also a part of the second research under research question RQ2. To find effective evaluation technique, the result of the pattern mining algorithms will be presented using a different method of visualization and explainability. The design of evaluation techniques and experiments with the different modalities of data will be conducted in parallel so that we can report the usefulness of found patterns and can also use the found patterns to understand the effectiveness of chosen evaluation.

RQ3 For the comparison purpose, RQ3 will also use the established evaluation method for the research question RQ2. The user evaluation will report the found pattern's usefulness to the provenance researchers. In addition, the comparison study with research output from the previous study will examine the effectiveness of using collectors' biography with object metadata.

6. Preliminary Work

Given that this work is in the first year of the PhD program, this section will project on the dataset considered for the first year's research work. This work considers object meta-data from the National Museum of World Culture (NMvW) as the primary data source. Each object is represented as a node and further connected with related information nodes, i.e., geographic location, title, material descriptions, collector's name etc. The NMvW has published two versions of linked data. In the first version, artefacts are placed in the centre of the model and directly connected to the data⁵. The second version uses an even-centric approach (i.e., CIDOC-CRM) adopted by the LinkedArt⁶ community for describing artefacts using related events.

For the second dataset for Entity Linking, this work considers data from Museum Bronbeek⁷ which contains biography information from military personnel and their involvement in different historic events (e.g., wars, expedition). From the unstructured retrieved data from the database (i.e., text description), Named Entity Recognition and Extraction tools will be used to enrich the primary data source. However, how many collectors we can spot from automated linking and which version of NMvW data is more appropriate for entity disambiguation is still a matter of investigation.

A small fraction of object collectors in the chosen data set happen to have their wiki-data identifier, which can be used as a gold standard for Entity Linking performance metric. From

⁵<https://collectie.wereldculture.nl/thesaurus/#/query/>

⁶<https://linked.art>

⁷<https://www.defensie.nl/onderwerpen/bronbeek/over-bronbeek>

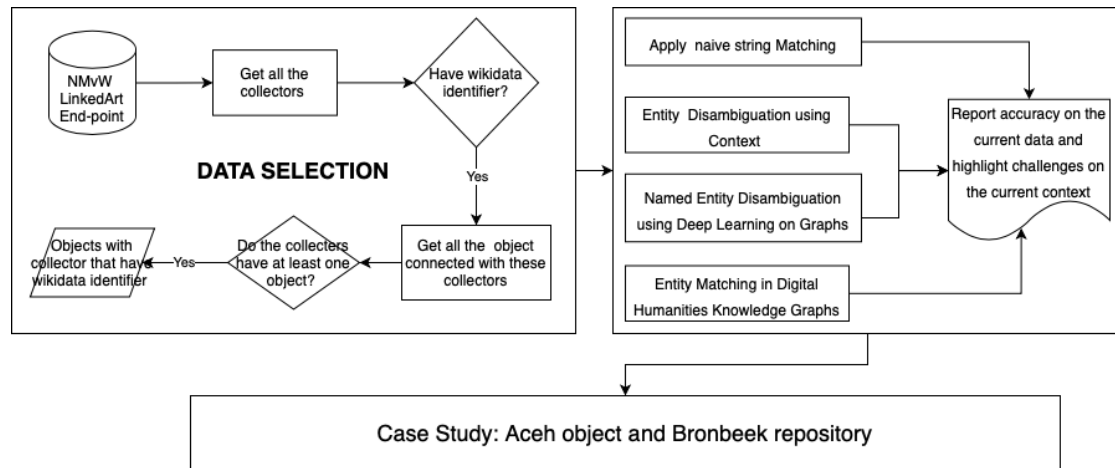


Figure 1: The experiment pipeline for Entity Linking where “Entity Disambiguation using context”, “Named Entity Disambiguation using Deep Learning on Graphs” and “Entity Matching in Digital Humanities Knowledge Graphs” refer to work [17], [18], and [19] respectively.

an initial study, it has been observed that the number of collectors with a wiki-data identifier is very small; therefore, to conclude on the success rate of entity disambiguation algorithms in the given context, we still need the human evaluation of the output.

As an initial case study and to reporting purpose in a practical use-case, this work will investigate a small set of NMvW material objects (consists of 4242 objects) from Aceh, Indonesia, collected between 1873 and 1942; which has been researched under the PPROCE⁸ project. From this research, dedicated links have been established between military personnel who has been stationed in Aceh war camps and objects’ collection. This subset will be used for reporting the performance of adopted entity disambiguation algorithms. The complete experiment pipeline to answer the question, *whether the existing state-of-the-art approaches is suitable for the given context or if there is a need for developing a new algorithm*, is given in Figure-1.

7. Conclusion and Limitation

This work plan shows how we intend to answer the question: *To what extent does Knowledge Graph constructed from heritage object’s metadata and further enriched with collector’s biography information has the potential to scale-up objects’ provenance research for museum experts*. While exploring the three focused research questions, this work aims to answer this overarching question.

The first year’s work will be focused on constructing a dataset by linking two data sources. The remaining years will focus on making this new and existing dataset usable for provenance research. This work will contribute to the field of digital humanities by providing tools or techniques to excel in provenance research. This paper can be seen as a proposal for a hy-

⁸<https://www.niod.nl/en/projects/pilotproject-provenance-research-objects-colonial-era-pproce>

pothesis generation tool for the domain experts rather than a concluding outcome for domain understanding.

This research takes a bottom-up approach instead of the traditional process mining approach for object provenance research. Given many objects with missing provenance information, we choose such a data-driven approach where research hypotheses can be generated based on data patterns. Though it is believed that both top-down and bottom-up approaches can benefit our findings in the given context, this work will explore the bottom-up approach primarily to manage the volume of work.

Acknowledgments

This work is conducted under the Pressing Matter project (<https://pressingmatter.nl/>) which is financed by the Dutch National Science Agenda (NWA) and coordinated from the Vrije Universiteit Amsterdam. I would like to thank my supervisors, Dr Victor De Boer, Prof. Dr Jacco van Ossenbruggen and Prof. Dr. Susan Legene for their guidance and feedback.

References

- [1] F. Van Harmelen, V. Lifschitz, B. Porter, Handbook of knowledge representation, Elsevier, 2008.
- [2] A. T. Schreiber, G. Schreiber, H. Akkermans, A. Anjewierden, N. Shadbolt, R. de Hoog, W. Van de Velde, B. Wielinga, R. Nigel, et al., Knowledge engineering and management: the CommonKADS methodology, MIT press, 2000.
- [3] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, H. Van de Sompel, The europeana data model (edm), in: World Library and Information Congress: 76th IFLA general conference and assembly, volume 10, IFLA, 2010, p. 15.
- [4] C. Dijkshoorn, L. Jongma, L. Aroyo, J. Van Ossenbruggen, G. Schreiber, W. Ter Weele, J. Wielemaker, The rijksmuseum collection as linked data, Semantic Web 9 (2018) 221–230.
- [5] N. Ockeloen, A. Fokkens, S. Ter Braake, P. Vossen, V. De Boer, G. Schreiber, S. Legène, Biographynet: Managing provenance at multiple levels and from different perspectives., in: LISC@ ISWC, Citeseer, 2013, pp. 59–71.
- [6] A. Tompkins, Provenance Research Today, Lund Humphries, 2021.
- [7] V. Charles, H. Manganinhas, A. Isaac, N. Freire, S. Gordea, Designing a multilingual knowledge graph as a service for cultural heritage—some challenges and solutions, in: International Conference on Dublin Core and Metadata Applications, 2018, pp. 29–40.
- [8] P. Ristoski, H. Paulheim, Rdf2vec: Rdf graph embeddings for data mining, 2016, pp. 498–514. doi:10.1007/978-3-319-46523-4_30.
- [9] A. Grover, J. Leskovec, Node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 855–864. URL: <https://doi.org/10.1145/2939672.2939754>. doi:10.1145/2939672.2939754.
- [10] H. Turner, Cataloguing culture: legacies of colonialism in museum documentation, UBC Press, 2020.

- [11] H. Turner, Organizing knowledge in museums: A review of concepts and concerns., *Knowledge Organization* 44 (2017).
- [12] B. Haslhofer, A. Isaac, R. Simon, Knowledge graphs in the libraries and digital humanities domain, *arXiv preprint arXiv:1803.03198* (2018).
- [13] V. Alexiev, et al., Museum linked open data: Ontologies, datasets, projects, *Digital Presentation and Preservation of Cultural and Scientific Heritage* (2018) 19–50.
- [14] A. Hogan, E. Blomqvist, M. Cochez, C. d’Amato, G. d. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, *Synthesis Lectures on Data, Semantics, and Knowledge* 12 (2021) 1–257.
- [15] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, *Semantic Web* 11 (2020) 187–193.
- [16] W. Shen, J. Wang, J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, *IEEE Transactions on Knowledge and Data Engineering* 27 (2015) 443–460. doi:10.1109/TKDE.2014.2327028.
- [17] I. O. Mulang[†], K. Singh, C. Prabhu, A. Nadgeri, J. Hoffart, J. Lehmann, Evaluating the Impact of Knowledge Graph Context on Entity Disambiguation Models, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020) 2157–2160. URL: <http://arxiv.org/abs/2008.05190>. doi:10.1145/3340531.3412159, arXiv: 2008.05190.
- [18] A. Cetoli, M. Akbari, S. Bragaglia, A. D. O’Harney, M. Sloan, Named Entity Disambiguation using Deep Learning on Graphs, *arXiv:1810.09164 [cs]* 11438 (2019) 78–86. URL: <http://arxiv.org/abs/1810.09164>. doi:10.1007/978-3-030-15719-7_10, arXiv: 1810.09164.
- [19] J. Baas, M. M. Dastani, A. J. Feelders, Entity matching in digital humanities knowledge graphs, *Proceedings* <http://ceur-ws.org> ISSN 1613 (2021) 0073.
- [20] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, *AI Magazine* 17 (1996) 37. URL: <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>. doi:10.1609/aimag.v17i3.1230.
- [21] I. Tiddi, M. d’Aquin, E. Motta, Data patterns explained with linked data, in: A. Bifet, M. May, B. Zadrozny, R. Gavaldà, D. Pedreschi, F. Bonchi, J. Cardoso, M. Spiliopoulou (Eds.), *Machine Learning and Knowledge Discovery in Databases*, Springer International Publishing, Cham, 2015, pp. 271–275.
- [22] L. Zuckerman, Tracking looted art with graphs. (extended abstract) graphs and networks in the humanities 2022 conference, february 3-4, 2022 (????). URL: https://www.academia.edu/70298130/Tracking_Looted_Art_with_Graphs_extended_abstract_Graphs_and_Networks_in_the_Humanities_2022_Conference_February_3_4_2022.
- [23] M. G. Kirschenbaum, The remaking of reading: Data mining and the digital humanities, in: *The National Science Foundation symposium on next generation of data mining and cyber-enabled discovery for innovation*, Baltimore, MD, volume 134, 2007.
- [24] M. Falkenthal, J. Barzen, U. Breitenbücher, S. Brüggemann, D. Joos, F. Leymann, M. Wurster, Pattern research in the digital humanities: how data mining techniques support the identification of costume patterns 32 (????) 311–321. URL: <http://link.springer.com/10.1007/s00450-016-0331-6>. doi:10.1007/s00450-016-0331-6.