1st Conference on Research Data Infrastructure

Harmonizing RDM

https://doi.org/10.52825/CoRDI.v1i.290

Published: 07 Sept. 2023

Determining the Similarity of Research Data by Using an Interoperable Metadata Extraction Method

Benedikt Heinrichs[1] [https://orcid.org/0000-0003-3309-5985], and
M. Amin Yazdi[1] [https://orcid.org/0000-0002-0628-4644]

[1]IT Center, RWTH Aachen University, Seffenter Weg 23, Aachen, Germany

Abstract: Determining the similarity of research data is not a simple task, as the formats can differ widely depending on the domain. Especially, since many formats are represented as binary files, the raw comparison of these will not yield good results. This makes it hard to accurately tell how similar certain research work is by comparing the data. With the emergence of extracted interoperable metadata, a form to describe data has been provided which is independent of the data format. Therefore, this work tries to use this extracted interoperable metadata and create a method to determine the similarity of research data based on their metadata. The produced method utilizes domain knowledge about the extracted metadata and the way they are formulated. A baseline is created, and further methods are created to compare to. The results show that our method outperforms all other methods, especially the ones which are focused on comparing the research data itself, not the metadata. Since the results are promising, we propose further investigations against other datasets and possible use cases.

Keywords: Research Data Management, Data Similarity, Metadata Similarity, Linked Data

1 INTRODUCTION

Typically, researchers accumulate large datasets while conducting scientific studies, which are often stored in various ways and locations depending on the domain. With the emergence of research data management and the FAIR Principles, scholars demanding to make research data Findable, Accessible, Interoperable, and Reusable [1], the current trend is to deal with this divergence. However, even though research data management (RDM) can provide platforms where research data can be stored and put forward recommendations for certain file types based on the domain, the reality is that there is no one-for-all solution. This is mainly because the scientific disciplines differ wildly and, with that, their data as well (e.g. texts, measurements, Excel sheets, and code). Especially since researchers aim to deliver novel studies, the choice for file types occasionally ought to be of a unique type to satisfy the requirements of a study. However, this creates a challenge when it is necessary to compare research data and assess the pairwise relevancy of files. While with texts, this might be an achievable task [2], with more complex file types represented in binary formats, this comparison

gets complicated very fast [3]. With different file types, this comparison can become impossible if one only has the raw binary format. Thankfully, work has been conducted to extract the most descriptive parts of research data and describe it as interoperable metadata [4]. With this metadata, a standard way exists to summarize research data, and it can be represented as a graph. Since with this, a uniform description of research data can be created, the task here is to extend the previous work and see how well this extracted metadata can be used to determine the similarity of research data, regardless of their type. The goal in mind is that such a method should outperform methods that try to directly compare two binaries of research data. Additionally, the created method should be provided as open-source software so that it can be easily integrated into RDM workflows.

2 CONTRIBUTION

For the contribution, we will shortly go over the background, discuss our approach, and finally present our results.

2.1 Background

The used metadata extraction method [5] [4] makes use of software like Apache Tika [6] to create interoperable metadata. This interoperable metadata is described using the Resource Description Framework (RDF) [7] and makes use of ontologies like DCAT [8]. Since RDF metadata can be formulated as a graph, graph similarity methods are viable candidates for the similarity approach [9]. Especially, works about structural similarity [10] and entity comparison [11] provide some building blocks for the proposed research goal.

2.2 Approach

Our proposed similarity method makes use of interoperable metadata as a graph. Additionally, with our domain-knowledge, we identify several optimizations that we can apply to improve upon standard similarity metrics. These optimizations aim to *Filter* out unique subjects, utilize the metadata *Structure* given by DCAT, and make the output *Simpler* by removing specific relationship triples (FSS). These optimizations were applied to similarity metrics like Jaccard, Cosine, and a customized metric. Our favored metric for this domain is Jaccard, so with the optimizations, our proposed method was called `FSS Jaccard`. For a comparison, additional methods were created that do not follow these optimizations. Lastly, methods were built that compare the research data binaries itself and not the interoperable metadata. All of these methods were compared on an evaluation dataset containing research data. The implementation of these methods is provided in this open-source repository: https://git.rwth-aachen.de/coscine/research/semanticsimilarity. The code can be run as a service so that it is easily integrable in RDM workflows.

2.3 Results

For evaluating the proposed similarity method, it was put against the other comparison methods. They all ran on the same evaluation dataset, and the results were evaluated using specific test characteristics and reliability measures. They were chosen based on the works of [12], and [13]. The results can be found in figure 1.

	Error Values		Reliability	Test Characteristics			Inter-rater reliability
	Mean	SD	true score	Sensitivity	Specificity	Accuracy	Cohen's Kappa
FSS Similarity	0,021	0,0442	0,9633	0,6491	0,9655	0,9031	0,6677
FSS Cosine	0,0218	0,0445	0,9628	0,6491	0,9655	0,9031	0,6677
FSS Jaccard	0,0195	0,0436	0,9628	0,6721	0,9912	0,9239	0,7437
Filter Similarity	0,0495	0,0494	0,9404	0,8182	0,7607	0,7716	0,4386
Cosine Binary	0,1304	0,2061	0,6772	0,9459	0,5952	0,6401	0,2514
Jaccard Binary	0,0504	0,1253	0,8433	0,6596	0,8182	0,7924	0,3853
Jaccard Similarity	0,0264	0,0704	0,9451	0,5738	0,9912	0,9031	0,6601

Note. Error values are determined by the absolute distance of the values from the validation values. Reliability is defined as the proportion of true variance, measured as covariance between the validation values and the test values, of the overall variance of the test values. Values are considered true positive if the data and validation value is larger than 0, but the data value is not bigger than 1.3 times the validation value (false positive) or less than 77% of the validation value. The data is still considered true negative if the validation value is 0 and the data value is less than 0.083 (average standard deviation of all algorithms).

Figure 1. Diagnostic Analysis

What can be seen from figure 1 is that `FSS Jaccard` (our proposed method) performs better than all the other methods when looking at the error values, reliability and inter-rater reliability. Especially interesting is that all methods based on the interoperable metadata perform much better than the ones based on the binary itself (`Jaccard Binary` and `Cosine Binary`). We discuss that this is the case because the interoperable metadata has more useful data to compare to, while binaries are oftentimes very hard to compare. The results for our proposed method `FSS Jaccard` are visualized in figure 2.

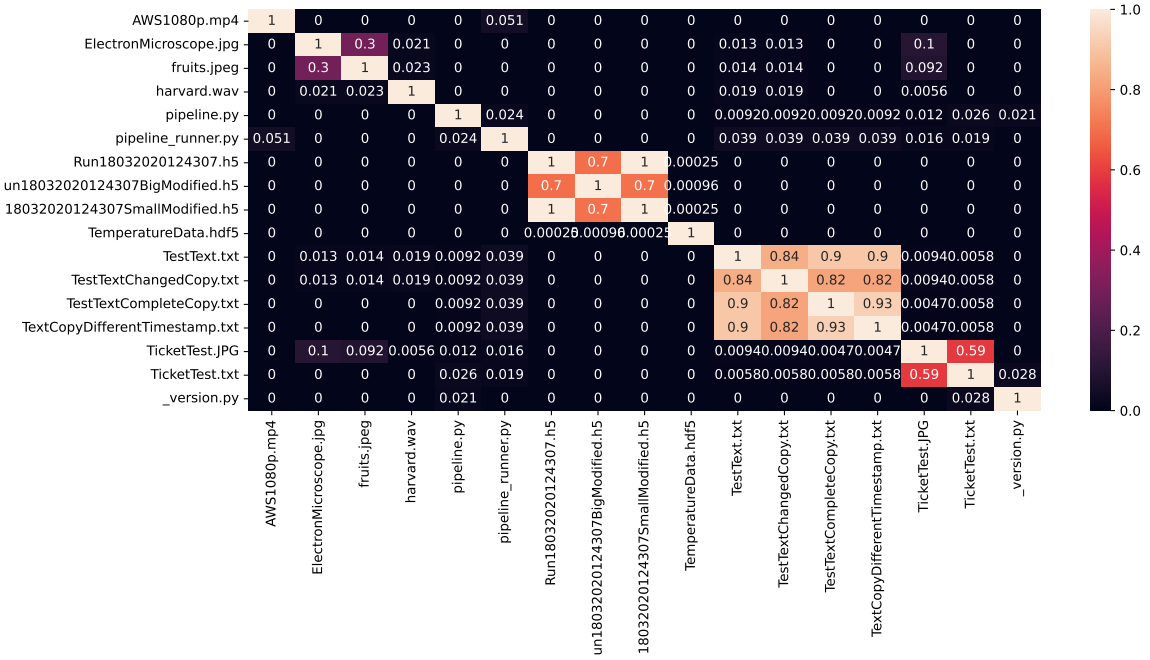

Figure 2. FSS Jaccard Results

The results in figure 2 show that there are not many outliers, and similar research data is usually identified as such. Especially interesting cases like smaller modifications between research data and research data which convey the same information but are in different file types (text and image) are accurately detected.

3 CONCLUSION

Our work established a new way to determine the similarity of research data based on interoperable metadata. First, we clarify our motivation, which ranges from the gen-

eral interest of figuring out how similar research data is to the problem of different data types. Then, we discuss the formal background and describe our similarity method. This method makes use of domain knowledge to craft a metric that filters the metadata, builds on top of the pre-existing structure, and simplifies it. The resulting algorithm is based on the Jaccard similarity metric, and we call it `FSS Jaccard`. We determine other similarity metrics to which we want to compare our similarity metric to on an evaluation dataset. The results strongly suggest that `FSS Jaccard` outperforms the other similarity metrics on our limited evaluation dataset. This is due to the use of interoperable metadata and the utilization of domain knowledge. Especially, the similarity metrics based on the binary itself perform worse than the ones based on the interoperable metadata. With this, we conclude that it has some merit to compare research data using interoperable metadata. This comparison has additionally the benefit of being type-agnostic and can even achieve better similarity results, at least in our case.

Competing interests

The authors declare that they have no competing interests.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, vol. 3, p. 160 018, 2016. DOI: 10.1038/sdata.2016.18.

[2] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—a survey," *ACM Comput. Surv.*, vol. 54, no. 2, Feb. 2021, ISSN: 0360-0300. DOI: 10.1145/3440755. [Online]. Available: https://doi.org/10.1145/3440755.

[3] S. Kim, Y. J. Yoo, J. So, J. G. Lee, J. Kim, and Y. W. Ko, "Design and implementation of binary file similarity evaluation system," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, no. 1, pp. 1–10, 2014. DOI: 10.14257/ijmue.2014.9.1.01.

[4] B. Heinrichs, N. Preuß, M. Politze, M. S. Müller, and P. F. Pelz, "Automatic General Metadata Extraction and Mapping in an HDF5 Use-case," in *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR,*, INSTICC, SciTePress, 2021, pp. 172–179, ISBN: 978-989-758-533-3. DOI: 10.5220/0010654100003064.

[5] B. Heinrichs and M. Politze, "Moving Towards a General Metadata Extraction Solution for Research Data with State-of-the-Art Methods," 12th International Conference on Knowledge Discovery and Information Retrieval, Nov. 2, 2020. DOI: 10.18154/RWTH-2020-12385. [Online]. Available: https://publications.rwth-aachen.de/record/809129.

[6] C. Mattmann and J. Zitting, *Tika in action*, 2011.

[7] D. Wood, M. Lanthaler, and R. Cyganiak, "RDF 1.1 Concepts and Abstract Syntax," W3C, W3C Recommendation, Feb. 2014, https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[8] A. Perego, A. G. Beltran, R. Albertoni, S. Cox, D. Browning, and P. Winstanley, "Data Catalog Vocabulary (DCAT) - Version 2," W3C, W3C Recommendation, Feb. 2020, https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/.

[9] J. Carroll, "Matching rdf graphs," May 2002, pp. 5–15, ISBN: 978-3-540-43760-4. DOI: 10.1007/3-540-48005-6_3.

[10] P. Maillot and C. Bobed, "Measuring structural similarity between rdf graphs," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ser. SAC '18, Pau, France: Association for Computing Machinery, 2018, pp. 1960–1967, ISBN: 9781450351911. DOI: 10.1145/3167132.3167342. [Online]. Available: https://doi.org/10.1145/3167132.3167342.

[11] A. Petrova, E. Sherkhonov, B. Cuenca Grau, and I. Horrocks, "Entity comparison in rdf graphs," in *The Semantic Web – ISWC 2017*, C. d'Amato, M. Fernandez, V. Tamma, *et al.*, Eds., Cham: Springer International Publishing, 2017, pp. 526–541, ISBN: 978-3-319-68288-4. DOI: 10.1007/978-3-319-68288-4_31.

[12] M. Eid, M. Gollwitzer, and M. Schmitt, *Statistik und Forschungsmethoden, Lehrbuch* (Grundlagen Psychologie), ger, 3., korrigierte Auflage, Online-Ausgabe. Weinheim ; Basel: Beltz, 2013, 1 Online–Ressource (XXXII, 1024 Seiten), ISBN: 978-3-621-27524-8. [Online]. Available: https://content-select.com/index.php?id=bib_view&ean=9783621278348.

[13] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, pp. 613–619, 1973. DOI: 10.1177/001316447303300 30.