

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

CoRDI Conference 2023
2023-09-14

Benedikt Heinrichs

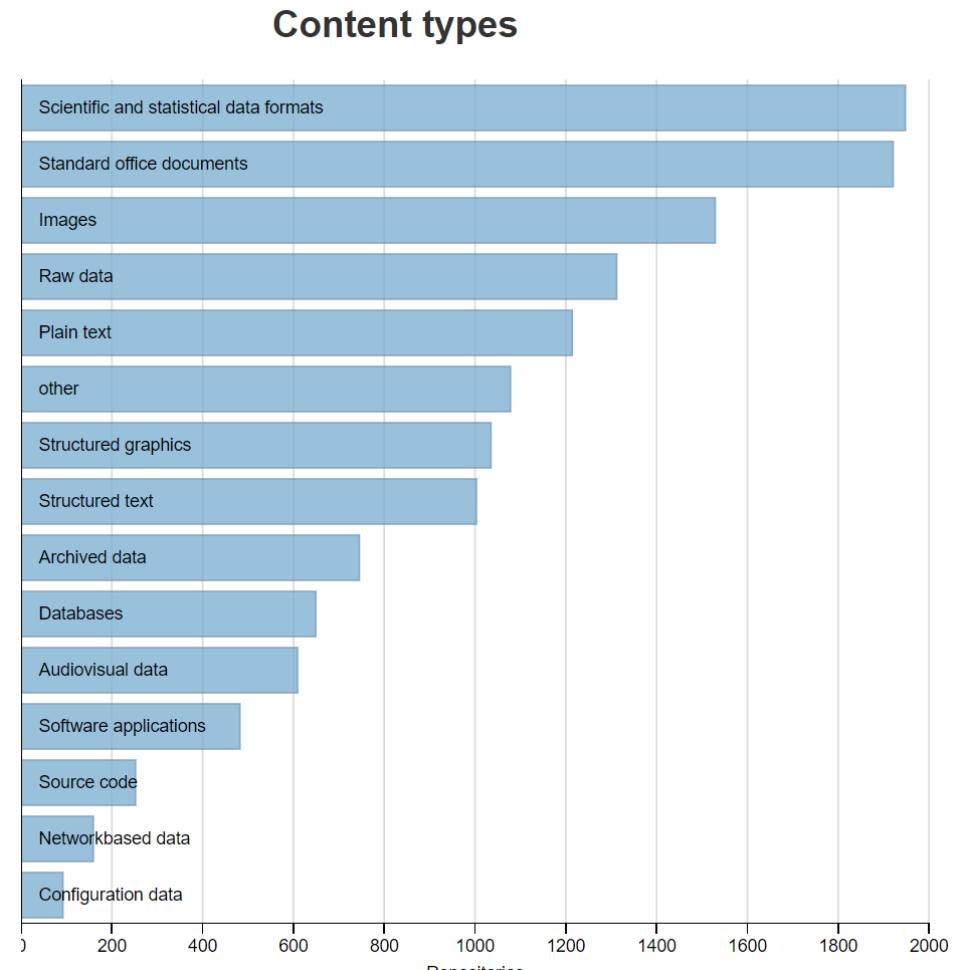


This work is licensed under a [Creative Commons Attribution 4.0 International License](#).



Comparing Research Data is Hard

- Research Data can be in many different formats
- This results in difficulties comparing research data
- With textual formats this might be easier since text comparison is easily doable
- With binaries, this becomes more difficult
- Especially comparing across formats is a difficult task



Content types from <https://www.re3data.org/>

Comparing Binary Research Data is Hard

0A27AF0	2CFC	2DC7	479C	F23F	B90D	5995	999D	F23F	,ü-çG	ò?¹.Y	ò?
0A27B00	2CFC	2DC7	479C	F23F	E1E2	D72A	A499	F23F	,ü-çG	ò?áâ×*¤	ò?
0A27B10	E1E2	D72A	A499	F23F	E1E2	D72A	A499	F23F	áâ×*¤	ò?áâ×*¤	ò?
0A27B20	8D22	8463	EB9E	F23F	2CFC	2DC7	479C	F23F	" cë	ò?,ü-çG	ò?
0A27B30	E1E2	D72A	A499	F23F	A53A	AF31	3DA0	F23F	áâ×*¤	ò?¥: ^1=	ò?
0A27B40	24DB	AC5C	5298	F23F	E1E2	D72A	A499	F23F	\$Û¬\R	ò?áâ×*¤	ò?
234EF0	2D30	342D	3031	2031	353A	3435	3A30	302E	-04-01	15:45:00.	
234F00	3000	0000	0000	0000	0000	3C00	0000	88E6	0.....	<... æ	
234F10	9BA1	A353	0440	08C2	1C16	3FB6	0240	54E2	¡£S. @. Ä... ?¶. @Tâ		
234F20	A401	96AC	0640	50C7	6AD9	76CB	973F	2FE4	¤. -. @PÇjÙvË	?/ä	
234F30	C48C	E026	943F	6F25	F5B6	26B7	A43F	3230	Ä à&	?o%ö¶&.¤?20	
234F40	3134	2D30	342D	3031	2031	353A	3436	3A30	14-04-01	15:46:0	

- How do you get a good idea on what is different when comparing two binaries (in this case HDF5 files)?

Possible ways to compare research data

- Use domain knowledge and use a similarity method for a specific type
- Compare the raw research data
- Use a representation of the research data

Representations of Research Data

- We talk a lot about creating metadata and extracting metadata
 - Why don't we make use of this metadata?
- A method has been proposed that can turn research data into (interoperable & content-based) metadata
 - This can act as a representation of research data
 - Making use of such a method could lead to format independent similarity computation
- The goal here is: Make use of the content-based representation and compare research data with it

Metadata Extraction Pipeline

Metadata Extraction



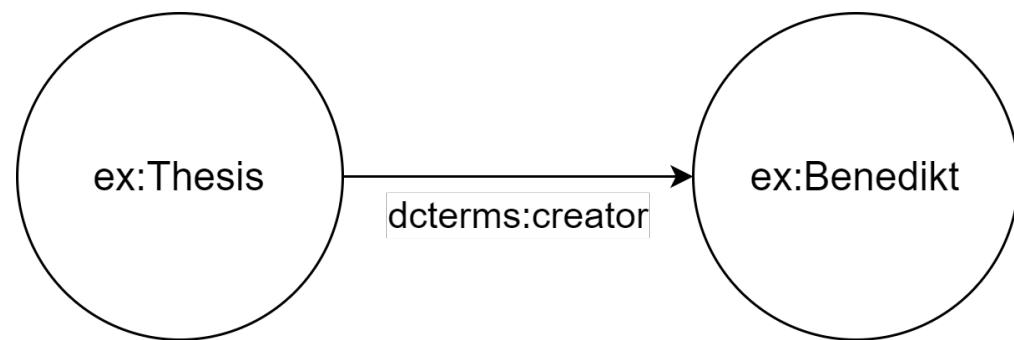
Heinrichs, B. ; Politze, M.

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, KDIR 2020, online, 2 Nov 2020 - 4 Nov 2020

Comparing Research Data with their Metadata

- The provided metadata is formulated in triples of subject, predicate and object pairs (RDF)
 - Follows ontologies and is interoperable
- This metadata can be represented as graphs where:
 - Subjects and Objects are the nodes
 - Predicates are the directed arcs that point from a subject to an object

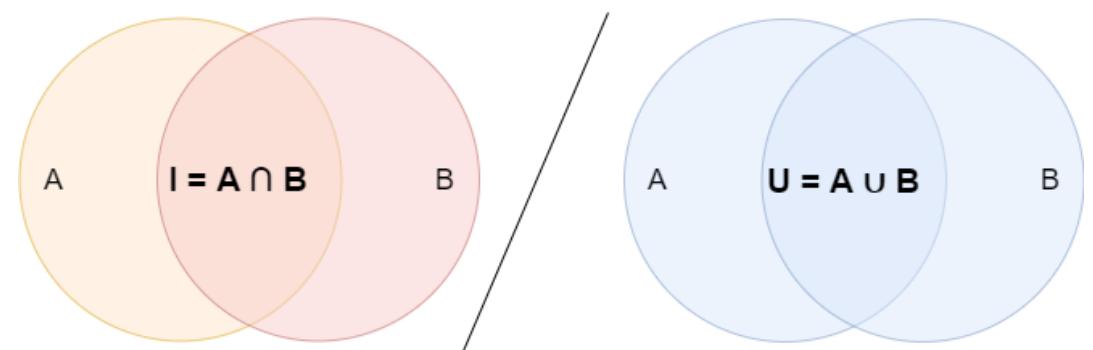


Comparing Research Data with their Metadata

- A multitude of graph similarity algorithms can be used to determine the similarity between metadata
- In this context, domain knowledge can be used as the properties of the interoperable metadata are known
- Known properties:
 - The subjects are usually unique and do not provide a lot of value in a similarity comparison, so they are *filtered* out
 - The *structure* of the interoperable metadata is similar
 - A couple of triples are only there to describe the internal structure and not the content, so the metadata can be *simplified*

Developed Method - FSS Jaccard

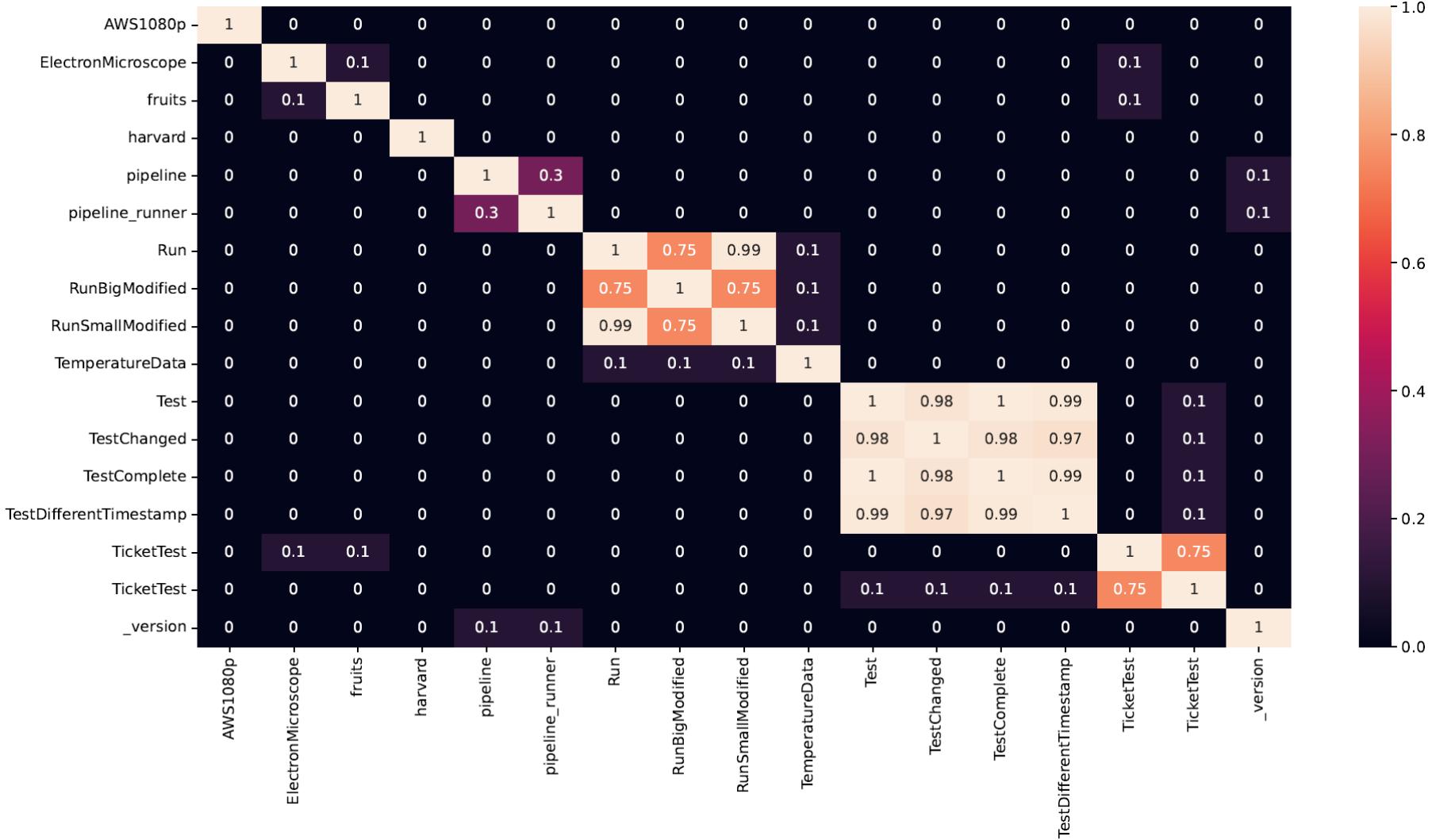
- The known properties lead to a method that follows:
 - a *filter* step
 - makes use of the *structure*
 - *simplifies* the metadata
- FSS
- We want to know the similarity between research data based on their metadata
 - This problem can be described when viewing research data as sets
 - Similar parts are in the intersections between sets
- Similarity algorithm to compute this: Jaccard
- FSS Jaccard was created



Developed Comparison Methods

- Based on interoperable metadata:
 - FSS Cosine – FSS with cosine similarity
 - FSS Similarity – FSS with a custom similarity
 - Jaccard Similarity – Jaccard similarity directly applied on the metadata
 - Filter Similarity – Similarity only on predicate & object pairs
- Based on research data:
 - Jaccard Binary – Jaccard similarity applied on research data
 - Cosine Binary – Cosine similarity applied on research data

Evaluation Dataset



Evaluation

Table 1

Diagnostical Analysis

	Error Values		Reliability	Test Characteristics			Inter-rater reliability
	Mean	SD		true score	Sensitivity	Specificity	
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).

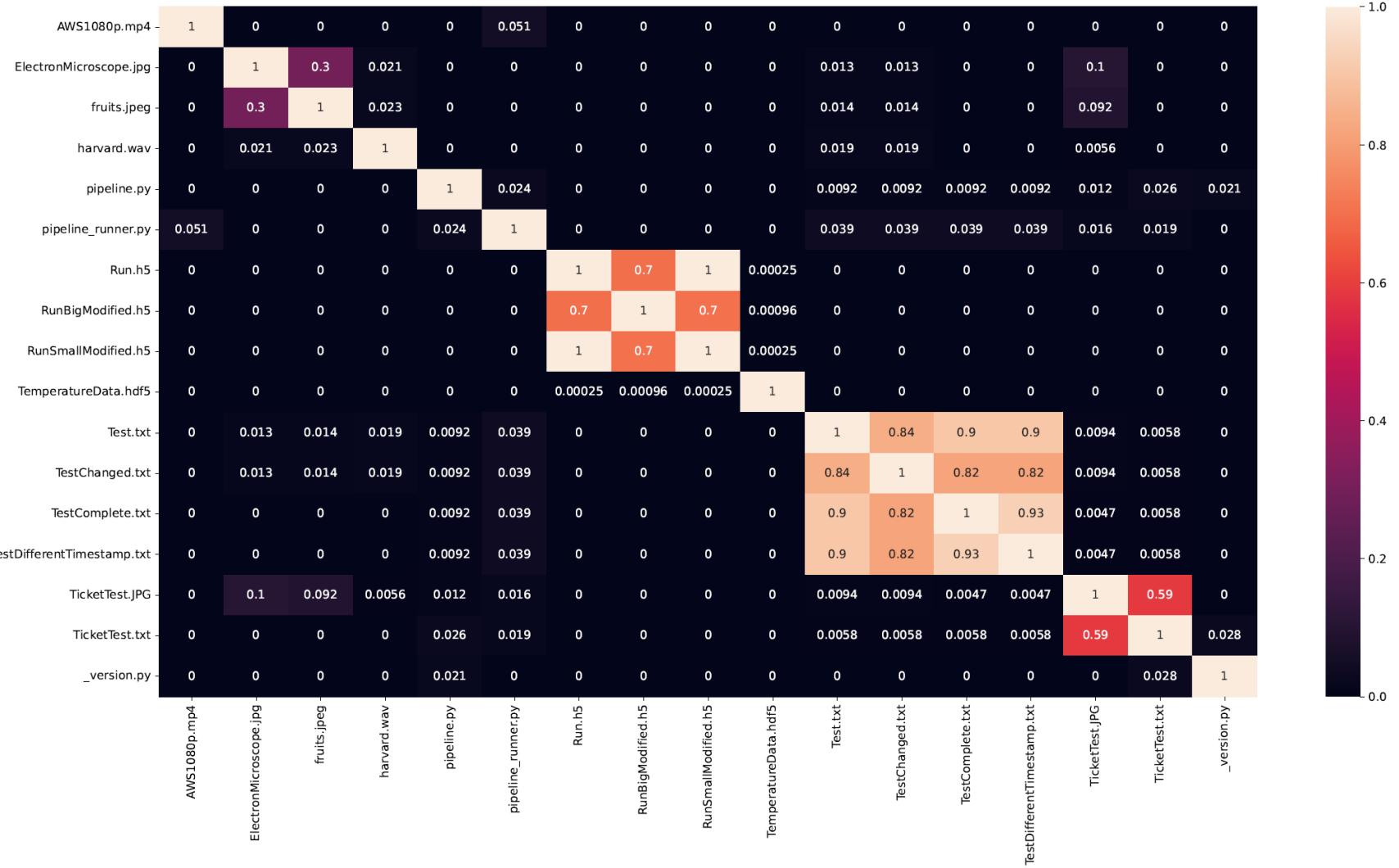
Evaluation – Ranking

1. FSS Jaccard
2. FSS Cosine & FSS Similarity
3. Jaccard Similarity
4. Filter Similarity
5. Jaccard Binary
6. Cosine Binary

Table 1
Diagnostical Analysis

	Inter-rater reliability
	Cohen's Kappa
FSS Similarity	0,6677
FSS Cosine	0,6677
FSS Jaccard	0,7437
Filter Similarity	0,4386
Cosine Binary	0,2514
Jaccard Binary	0,3853
Jaccard Similarity	0,6601

Results – FSS Jaccard



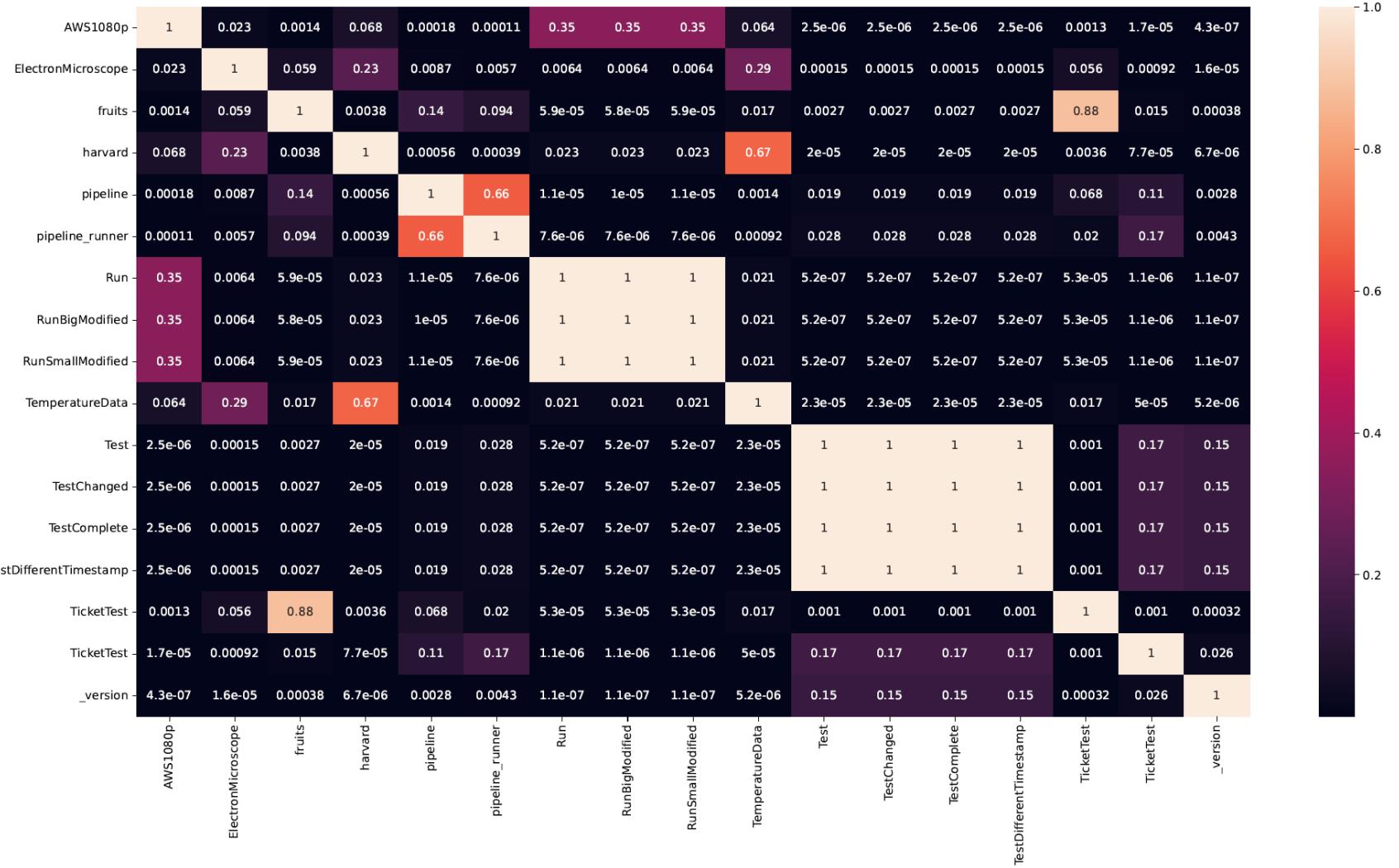
Results – FSS Jaccard

Evaluation Dataset

1	0	0	0	0	0.051	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0.3	0.021	0	0	0	0	0	0	0	0.013	0.013	0	0	0.1	0	0
0	0.3	1	0.023	0	0	0	0	0	0	0.014	0.014	0	0	0.092	0	0	0
0	0.021	0.023	1	0	0	0	0	0	0	0.019	0.019	0	0	0.0056	0	0	0
0	0	0	0	0	1	0.024	0	0	0	0.0092	0.0092	0.0092	0.0092	0.012	0.026	0.021	0
0.051	0	0	0	0	0.024	1	0	0	0	0.039	0.039	0.039	0.039	0.016	0.019	0	0
0	0	0	0	0	0	0	1	0.7	1	0.00025	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.7	1	0.7	0.00096	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0.7	1	0.00025	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.00025	0.00096	0.00025	1	0	0	0	0	0	0	0
0	0.013	0.014	0.019	0.0092	0.039	0	0	0	0	1	0.84	0.9	0.9	0.0094	0.0058	0	0
0	0.013	0.014	0.019	0.0092	0.039	0	0	0	0	0.84	1	0.82	0.82	0.0094	0.0058	0	0
0	0	0	0	0.0092	0.039	0	0	0	0	0.9	0.82	1	0.93	0.0047	0.0058	0	0
0	0	0	0	0.0092	0.039	0	0	0	0	0.9	0.82	0.93	1	0.0047	0.0058	0	0
0	0.1	0.092	0.0056	0.012	0.016	0	0	0	0	0.0094	0.0094	0.0047	0.0047	1	0.59	0	0
0	0	0	0	0.026	0.019	0	0	0	0	0.0058	0.0058	0.0058	0.0058	0.59	1	0.028	1
0	0	0	0	0.021	0	0	0	0	0	0	0	0	0	0	0.028	1	1

FSS Jaccard

Results – Jaccard Binary



Results – Jaccard Binary

1	0.023	0.0014	0.068	0.00018	0.00011	0.35	0.35	0.35	0.064	2.5e-06	2.5e-06	2.5e-06	2.5e-06	0.0013	1.7e-05	4.3e-07
0.023	1	0.059	0.23	0.0087	0.0057	0.0064	0.0064	0.0064	0.29	0.00015	0.00015	0.00015	0.00015	0.056	0.00092	1.6e-05
0.0014	0.059	1	0.0038	0.14	0.094	5.9e-05	5.8e-05	5.9e-05	0.017	0.0027	0.0027	0.0027	0.0027	0.88	0.015	0.00038
0.068	0.23	0.0038	1	0.00056	0.00039	0.023	0.023	0.023	0.67	2e-05	2e-05	2e-05	2e-05	0.0036	7.7e-05	6.7e-06
0.00018	0.0087	0.14	0.00056	1	0.66	1.1e-05	1e-05	1.1e-05	0.0014	0.019	0.019	0.019	0.019	0.068	0.11	0.0028
0.00011	0.0057	0.094	0.00039	0.66	1	7.6e-06	7.6e-06	7.6e-06	0.00092	0.028	0.028	0.028	0.028	0.02	0.17	0.0043
0.35	0.0064	5.9e-05	0.023	1.1e-05	7.6e-06	1	1	1	0.021	5.2e-07	5.2e-07	5.2e-07	5.2e-07	5.3e-05	1.1e-06	1.1e-07
0.35	0.0064	5.8e-05	0.023	1e-05	7.6e-06	1	1	1	0.021	5.2e-07	5.2e-07	5.2e-07	5.2e-07	5.3e-05	1.1e-06	1.1e-07
0.35	0.0064	5.9e-05	0.023	1.1e-05	7.6e-06	1	1	1	0.021	5.2e-07	5.2e-07	5.2e-07	5.2e-07	5.3e-05	1.1e-06	1.1e-07
0.064	0.29	0.017	0.67	0.0014	0.00092	0.021	0.021	0.021	1	2.3e-05	2.3e-05	2.3e-05	2.3e-05	0.017	5e-05	5.2e-06
2.5e-06	0.00015	0.0027	2e-05	0.019	0.028	5.2e-07	5.2e-07	5.2e-07	2.3e-05	1	1	1	1	0.001	0.17	0.15
2.5e-06	0.00015	0.0027	2e-05	0.019	0.028	5.2e-07	5.2e-07	5.2e-07	2.3e-05	1	1	1	1	0.001	0.17	0.15
2.5e-06	0.00015	0.0027	2e-05	0.019	0.028	5.2e-07	5.2e-07	5.2e-07	2.3e-05	1	1	1	1	0.001	0.17	0.15
2.5e-06	0.00015	0.0027	2e-05	0.019	0.028	5.2e-07	5.2e-07	5.2e-07	2.3e-05	1	1	1	1	0.001	0.17	0.15
0.0013	0.056	0.88	0.0036	0.068	0.02	5.3e-05	5.3e-05	5.3e-05	0.017	0.001	0.001	0.001	0.001	1	0.001	0.00032
1.7e-05	0.00092	0.015	7.7e-05	0.11	0.17	1.1e-06	1.1e-06	1.1e-06	5e-05	0.17	0.17	0.17	0.17	0.001	1	0.026
4.3e-07	1.6e-05	0.00038	6.7e-06	0.0028	0.0043	1.1e-07	1.1e-07	1.1e-07	5.2e-06	0.15	0.15	0.15	0.15	0.00032	0.026	1

Evaluation Dataset

Jaccard Binary

Conclusion

- Making use of extracted interoperable metadata to compare research data shows promising results!
 - Methods based on interoperable metadata outperformed methods based on the research data itself in this experiment
 - The proposed method FSS Jaccard performed best out of the presented methods
- The results can be used when wanting to determine the similarity between research data when interoperable metadata is available (or extractable)
 - Applications could be provenance tracking
(e.g., interoperable metadata of an older version is present and has suddenly changed)
 - Insights can be gained when applying it to use cases
- Code available to test it out yourself:
 - <https://git.rwth-aachen.de/coscine/research/semanticsimilarity>



**Thank you
for your attention!**

Results – Use Case – Clusters inD

