Persistent Data, Sustainable Information

Simon Schiff[1,*], Ralf Möller[1]

[1]University of Lübeck, Institute of Information Systems, Ratzeburger Allee 160, 23562 Lübeck, Germany

Abstract

In almost all academic fields, results are derived from found evidence such as digitized objects stored at a repository. Deriving results from such repositories can be time and cost intensive, as data is often difficult to reuse. Guidelines such as FAIR (**F**indability, **A**ccessibility, **I**nteroperable, and **R**euse) are intended to disseminate the proper archiving of research data such that, among others, archived data is easy to reuse. However, we argue that even if one follows FAIR guidelines, it is still challenging to derive results by reusing data. First of all, before reusing data from any repository, one needs to find the data. Search engines can index data stored at repositories, as data is associated with metadata as proposed by the FAIR guidelines. Deciding whether the data found is relevant usually requires downloading, extracting and visualizing the entire dataset which is time-consuming and costly. We propose that data need to be archived by associating it with metadata that determines how data can be prepared to support decision making. Metadata links data with executable code that can create an information system from associated data on demand. With our solution, data is easier to reuse, as one can decide whether found data is relevant. In addition, if the data is found to be relevant, our information system allows researchers to clearly refer to specific regions in the data for data governance.

Keywords

persistent data, sustainable information, research data, data management

1. Introduction

Domain-independent guidelines such as FAIR (**F**indability, **A**ccessibility, **I**nteroperable, and **R**euse) [1] or domain-dependent ones such as CARE (**C**ollective benefit, **A**uthority to Control, **R**esponsibility, and **E**thics) [2] are intended to disseminate the proper use of research data. Many problems that can arise when dealing with research data should not arise in the first place if these guidelines are followed. However, researchers still spend a lot of time on deriving any evidence from research data that can be found at repositories, such as case studies, observations, experiments, or digitizations of objects, as authors of repositories do not always follow these guidelines. The reasons are manifold, among others, it costs to much time or one does not know why the effort would be worthwhile. As presented by Schiff et al. in [3], researchers can be supported with an extract transform load (ETL) pipeline to spend less time on producing results being valid with respect to the FAIR guidelines, while they still work with their preferred tools and document formats. Based on this pipeline, research data from a repository are extracted, transformed into formats to be interpretable by machines and loaded into another repository that

3rd Workshop on Humanities-Centred Artificial Intelligence (CHAI)

schiff@ifis.uni-luebeck.de (S. Schiff); moeller@ifis.uni-luebeck.de (R. Möller)

https://www.ifis.uni-luebeck.de/index.php (S. Schiff); https://www.ifis.uni-luebeck.de/index.php (R. Möller)

0000-0002-1986-3119 (S. Schiff); 0000-0002-1174-3323 (R. Möller)

CEUR Workshop Proceedings (CEUR-WS.org)

is linked with the repository from where the data was extracted from. As data is interpretable by machines, one can semi-automatically combine and link the data which reduces the time required to do that manually by several weeks or even months. With the ETL, data is valid with respect to the FAIR principles, however one needs to download the whole repository to decide whether data is relevant. And if data is found to be relevant, then one can only refer to the whole repository via a unique identifier that is usually associated with a repository instead of referring to specific regions in the data. We argue that one needs an information system, accessing data stored at a repository, to decide quickly whether found data is relevant and to uniquely refer to specific regions in the data. Thus, if one persists her or his data, one gets an information system *almost* for free.

We explain the problem of managing a mesh of data in Section 2, followed by Section 3, where we introduce an ETL pipeline that helps produce results that conform to FAIR principles and thus reduce the time spent searching for information in a mesh of data. In Section 4 we present InvenioRDM, a research data repository implementation, that we extend with an "information system button" to obtain an information system on demand, that we present in Section 5. In Section 6 we extend our information system with share buttons to allow others to uniquely refer to specific regions in the data uploaded at InvenioRDM, and finally conclude and give an outlook to further research directions in Section 7.

2. Mesh of Data

Independent of the research field, researchers make observations, such as finding written artefacts from the past, a disease that can not be treated at the moment, or problems that occur during a specific process to be optimized. Observations need to be kept so that they can be reviewed, examined and used as evidence later by creating a digitized version of them in any format. Problems that occur during the digitization are manifold, have various reasons, and are the fault for a mesh of data locally at each repository. For example, a mesh of data is depicted in Figure 1. At the bottom middle, hundreds of described artifacts are the subject of research findable in a library that is difficult to access. It is almost impossible to uniquely refer to these artefacts via a unique identifier as the library has no bookshelf numbers. Pictures of written artefacts can be made only once as access to the library is highly restricted and it might be impossible to transcribe all texts from pictures that where made. Additionally, a material analysis of the artefacts is promising, however either too costly or not permitted. Hours of audio transcriptions of the contents of the written artefacts are available as well. All observations, where it was possible to collect them on site, are stored in a repository.

The task of a researcher is then to produce results based on data stored at the repository, such as transcribing texts from pictures of described artefacts to create a critical edition. However, data was collected on site in a hurry without associating it with additional metadata. Metadata contains a description, image resolution, length of a recording, a unique identifier, a title and many more and it is very time-consuming to create it. An information retrieval (IR) system, supporting a researcher to find data quickly, can not index the data if it is not associated with metadata, researchers can not point explicitly to any data point without a unique identifier part of the metadata, and one needs to download data from the repository to decide whether it is

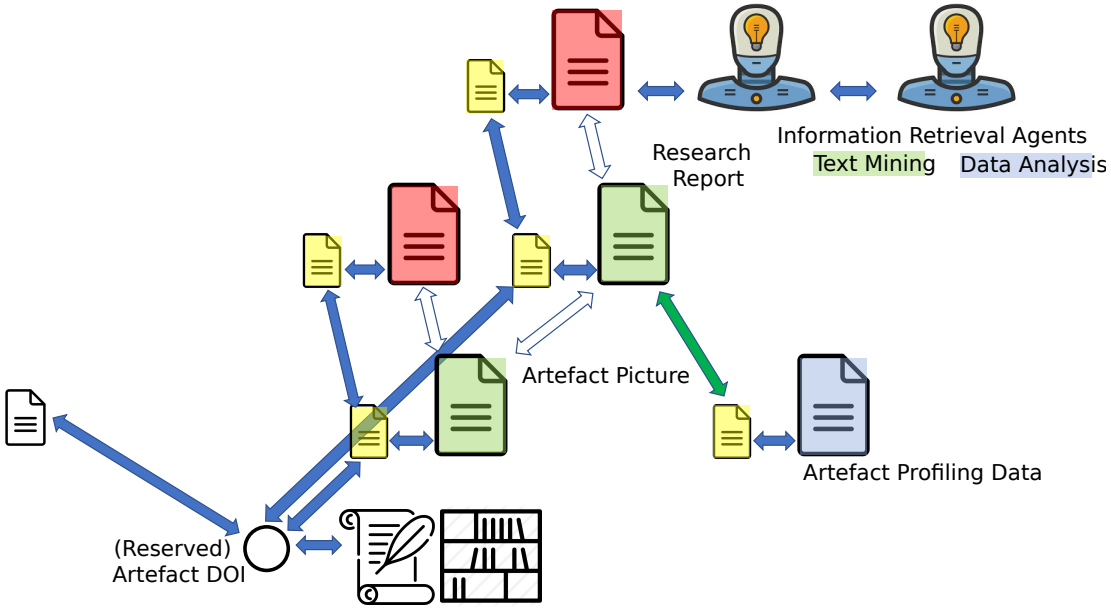

Figure 1: Illustration of a Mesh of Data

relevant or not.

Hence, many work needs to be done before deriving results from observations, such as producing metadata about the artefacts. However that is mostly not done caused by lack of time, of knowledge in IT, or of knowledge why the effort would be worthwhile.

Results then do not contain any explicit link to made observations. Instead descriptions pointing to the origin of the results are written in natural language. Other local repositories containing related observations or results may exist, as depicted in Figure 1 on the middle left. Data across repositories are possibly linked with each other implicitly, explicitly or not at all. In the latter case, however, the lack of links through identically but differently named individuals cannot be excluded. A faceted search engine can only index data stored at a repository with available metadata. Published results, containing among others implicit links, are less likely to be reused or further processable by machines.

3. Multi-Target Publishing

Schiff et al. in [3] aim to enable researchers to spend less time on publishing reusable results while they still work with preferred tools and document formats. Mostly, researchers format results such that they can be printed later by using so called what-you-see-is-what-you-get (WYSIWYG) editors. Documents formatted for later printing should be able to be read by humans, not necessarily interpreted by machines, and are therefore not easily reusable. Hence, as depicted in Figure 2, a new ETL pipeline is proposed that enables researchers to publish reusable results without changing their preferred tools and document formats. Results to be printed later, such as a Microsoft Word DOCX document, is stored at a research data repository

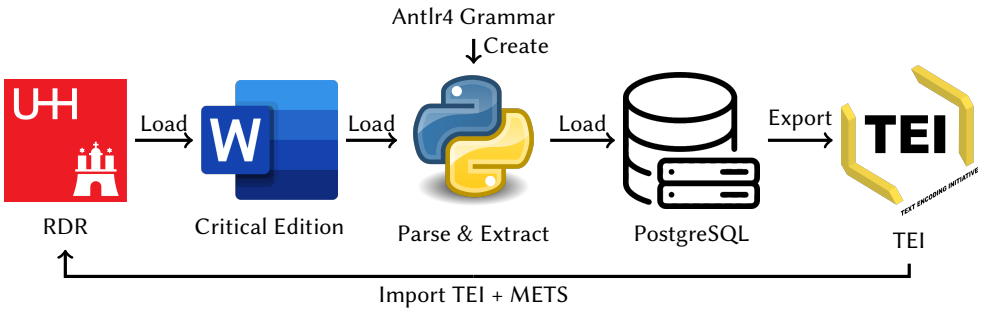

Figure 2: Extract transform load (ETL) pipeline

(RDR) and associated with metadata containing, among others, author names and a unique identifier such as a digital object identifier (DOI). Word DOCX documents are loaded from a RDR, parsed and contents are extracted, by an automatically generated parser from a manually written Antlr4 grammar [4]. Everything that is, from the perspective of the author, semantically different is separated after parsing the document. Parsed document contents are loaded into a PostgreSQL RDBMS (relational database management system) to, among others, export the data in any other desired format, such as Text Encoding Initiative (TEI). TEI is a XSD schema for XML documents to encode results produced primarily by humanities scholars. By our observations, humanities scholars produce such XML formatted documents often manually which is a time intensive laborious task and possibly error prone. Documents are available in a structured format, with the help of the ETL pipeline, and thus interpretable by machines. This is beneficial not only for exporting documents to any format, but also for semi-automatic combination and linking of document content which can reduce the time spent on manual processing by several weeks or even months. Finally, exported documents are stored at a RDR and linked with the original document from where they stem via a unique identifier. With that solution, researchers spend less time on producing results to be reusable and being valid with respect to the FAIR guidelines. However, many time is still spend on deciding whether data stored at a repository is relevant. Hence, we extend InvenioRDM, a research data repository implementation, with an "information system button" to allow those who are interested in data stored at a repository to obtain an information system on demand.

4. InvenioRDM

The RDR is hosted at the Universität Hamburg and managed by the Center for Sustainable Research Data Management (RDM). It is a repository containing research data created and uploaded by, among others, humanities scholars, and is based on Zenodo. Zenodo itself is an open-source platform, based on the Invenio digital library that is also open source and developed by European Organization for Nuclear Research (CERN) [5]. The platform enables researchers worldwide to share their research data with others associated with a license and a DOI.

Uploading at the RDR is done by accessing it via a web browser and then uploading research data from local storage. Uploading files requires providing additional metadata that is associated

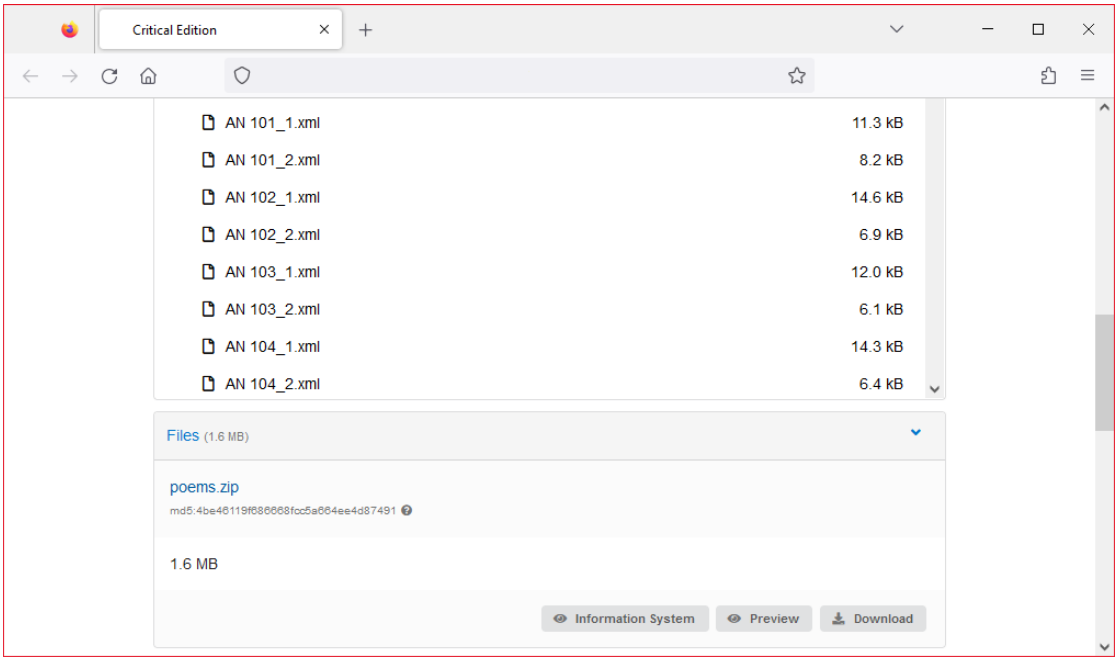

Figure 3: Uploaded data at our modified InvenioRDM

with the uploaded research data. Metadata is encoded in XML and its schema is standardized by DataCite.org [6]. Some metadata information is mandatory, while other data points are optional. Mandatory metadata is used by providers of repositories, such as RDM, for representing uploaded datasets and to enable others to find them. Each uploaded dataset at the RDR has an automatically created DOI that is associated with it. It is impossible to change the uploaded data afterwards without changing the DOI. Thus, a DOI is always associated with the same immutable data and therefore citable. Otherwise, published results based on public research data are not verifiable.

The uploaded data at the RDR is archived to be available in the long term later, but researchers are still interested in working with the data for now. For instance, a RDR contains a ZIP archive full of TEI documents. Each TEI document contains a transcription of a described artifact in a language used three centuries ago. Researchers are interested in these transcriptions, as they provide additional context about the words they aim to translate into English. First uploaded data at a RDR needs to be found and to find data, one needs to use an IR system. Among other repositories, the repository containing the ZIP archive full of TEI documents is returned by the IR system, as depicted in Figure 3. However, one cannot decide whether the ZIP archive contains relevant information, even though the RDR provides a preview of uploaded ZIP archives, as visualized at the top of Figure 3, as the previewer lists only the names of the TEI files. The whole ZIP archive needs to be downloaded, extracted and the contents of the files visualized for decision making. We modified the RDR by adding an “information system button”, that redirects a user directly to an information system that can be used to explore archived data at the RDR.

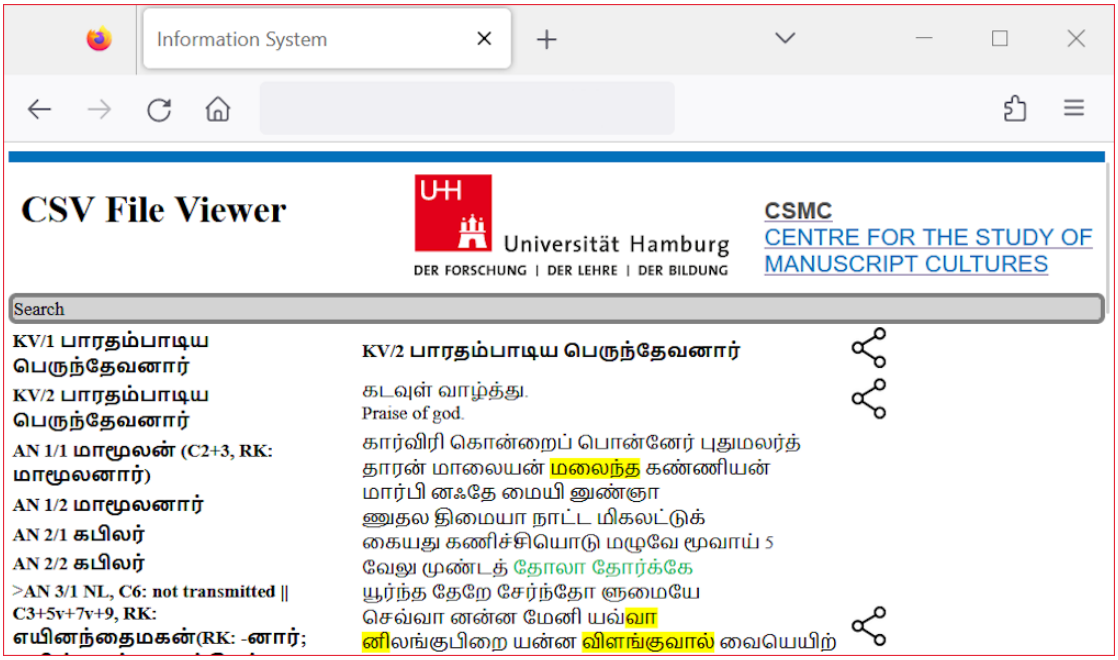

Figure 4: Generic CSV viewer

5. Information System

As we argue at the end of Section 1, if one persists her or his data, then one gets an information system almost for free. To obtain an information system by clicking on the "information system button", as depicted on the bottom middle of Figure 3, some requirements need to be fulfilled: (i) The data needs to be formatted to be interpretable by machines (optionally with the help of our ETL) and zipped as an archive, (ii) the data needs to be associated with a Metadata Encoding and Transmission Standard (METS) file encoded in XML, (iii) executable code needs to be linked with the data it visualizes via the METS file. The first requirement is easy to implement using our ETL and data then no longer violates the FAIR principles. Combining all the data that belongs together into one archive reduces the storage requirements and the time needed to upload or download the data.

Metadata that needs to be added to the archive is encoded in XML and its schema is specified according to METS for interoperability. The METS standard is developed by the METS board in collaboration with the Library of Congress [7]. METS is recommended, among other scenarios, when research data is uploaded as archives at an RDR for long-term preservation and is mainly composed of seven sections [7]: (i) METS header, (ii) descriptive metadata, (iii) administrative metadata, (iv) file section, (v) structural map, (vi) structural links, and (vii) a behaviour section. The METS header contains metadata about the METS file, such as the last modification date, the author name of the file, and one or more unique identifier. Descriptive and administrative metadata sections are optional. The former contains metadata encoded in any form about archived metadata and the latter, among others, intellectual property rights, or digital prove-

nance metadata. Data is hierarchically grouped in the file section and each file is associated with a unique id. A structural map section contains a hierarchy of nested `<div>` elements, each having a unique ID. It determines how data is presented to users, for instance, as a file browser, and for navigating through the data that is associated with the METS file. All `<div>` elements part of the structural map section can be linked with each other in the structural links section. The last section in the METS file determines the behavior when a user clicks on a file with a file browser, for example.

Among others, a JS script can be executed that creates an information system, as depicted in Figure 4, when a user clicks on a generic CSV file. On the left hand side of Figure 4, each row of the CSV is listed and represented by a specific column and on the right hand side a selected row from the list. Not all rows are shown on the left hand side, as some contain only database specific columns, such as a sequence number. Which entries are shown and which column represents all other ones in the list on the left hand side is determined in the METS file. Additionally, we added a sophisticated search engine, as depicted on top of Figure 4, which can be used to filter all rows in the CSV file. Hence, if one persists her or his data, then one gets an information system almost for free. Almost as data needs to be only associated with a METS file, archived as a ZIP file, and finally uploaded at our modified RDR. Generating a METS file is supported by our web application and predefined viewer can be used for, among others, CSV or TEI files.

6. Collections of Unique Identifiers

Recent trends show that publishing research data is considered good research practice and is increasingly mandated by funding agencies. Publishing research data allows for verification of results, reuse for other purposes, combination with other data, competitions and many other uses. Platforms such as Kaggle [8] or Zenodo [6, 9] allow sharing of research data with others. Sharing data first requires creating an account, then setting up a repository, and finally uploading the data. Uploaded data is associated with metadata to index it with an IR system and to allow users to decide whether the data fits their needs. Zenodo assigns each version of a repository a DOI that enables everyone to uniquely refer to it. This is important, as otherwise, other users are unable to verify results based on published research data, as the research data may have changed, leading to a potential change in the findings. A repository can contain research data of arbitrary format, ideally standardized, any number of files, and stem from one or more domains. Citing a repository of research data is only possible with a unique identifier, such as a DOI. In some cases, results depend on only a small subset of research data stored at a repository and thus citing the complete repository is not precise enough. One solution is to create a new repository that contains the subset of research data on which the results depend and which can be cited. This can lead to a collection of many small repositories containing subsets of research data from larger repositories that can only be used to produce results by the author of the repository. Many small repositories cause additional overhead in a mesh of data and poor comparability of sources used. Instead, we argue that one should be able to directly refer to subsets of research data within a repository. This allows everyone to comprehend how results were produced from actual data. As repositories mainly contain research data that is in

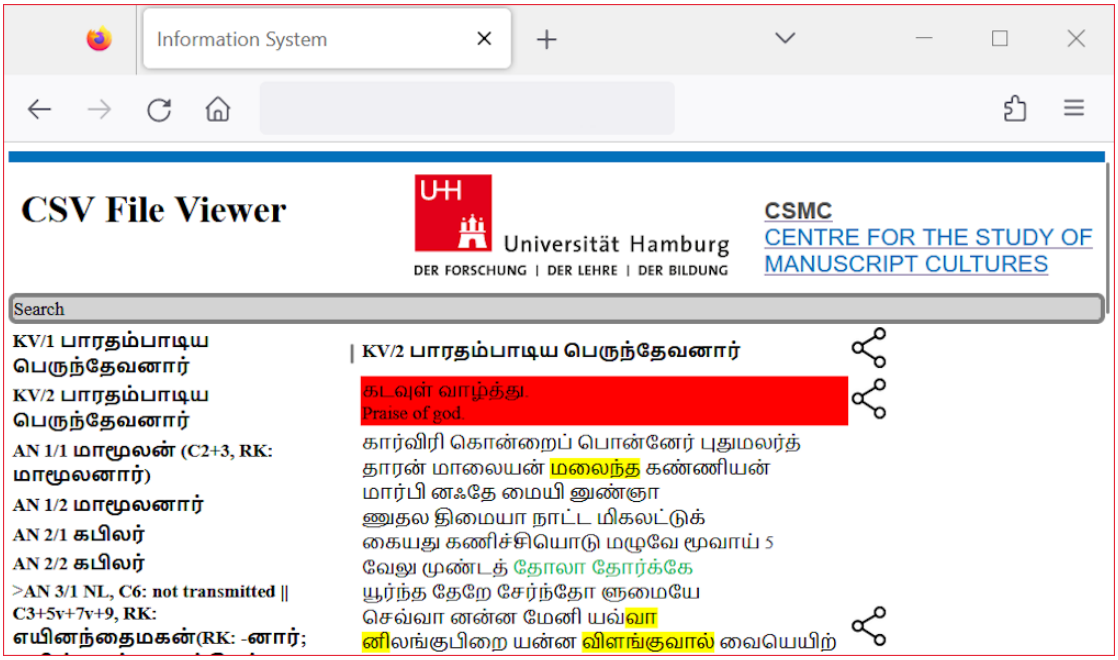

Figure 5: Selected region in research data

a format not primarily designed to be read by humans, our information system visualizes the data appropriately and allows for referencing to any parts of the data regardless of its format. Previewers for repositories hosted by Kaggle or Zenodo, on the other hand, only show a small section of the uploaded data.

TEI documents stored at the RDR and displayed by our information system are read by those interested in the documents' contents for their own research and, thus, might want to refer to specific parts of the data. Using the DOI that is created for TEI documents that stem from the transformation of a Word DOCX document and are stored at the RDR is an option; however, it refers to a repository of TEI documents and not specific parts of the documents of interest. Hence, we extend our information system with share buttons, as depicted in Figure 4, that allow adding parts of documents to a collection. If a collection is complete, one can associate it with a DOI that refers uniquely to the collection as long as the collection does not change. Collections not associated with a DOI are visible only to the collection's creator. As illustrated in Figure 5, clicking on a reference part of a collection takes one to an information system that highlights the referenced area. On the other hand, those interested in a particular area can see in which context it has already been referenced together with other regions in the data.

7. Conclusion and Outlook

Researchers can decide quickly whether data stored at a repository is relevant with the help of our information system. If data is found to be relevant and reused to produce results, then one can uniquely refer to data from where results are derived via a unique identifier. Others can

verify how results where produced from data via unique identifer linking to our information system. Our information system then presents all data linked with results.

In a future work, we want to adapt solutions from platforms whose primary purpose is to share source code with others, like GitHub [10] or GitLab [11]. Such platforms allow to fork a whole repository to make a pull request. For each request, a forum can be used for discussions and the code can be automatically executed for tests. Research data is not executable code, however for each request, existing AI models based on uploaded data could be retrained and check whether there is an improvement. An improvement is an indication that it has improved the quality of the data with respect to a model to be trained. Security alerts are optionally available if source code uses a library that is found to be insecure. In case of research data, one could receive an alert if data is linked with sources that are unethical or have a new licence that forbids the reuse. Additionally, GitHub and OpenAI launched Copilot, an "AI pair programmer" [12]. Such a Copilot (e.g., agent) could support researchers producing new results based on uploaded research data at a repository by, for instance, searching for information, given the context a researcher is currently working on. Those who allow an agent to access their data to support others are more likely to be cited.

References

[1] FAIR, Go fair, 2022. URL: https://www.go-fair.org/.

[2] S. R. Carroll, I. Garba, O. L. Figueroa-Rodríguez, et al., The care principles for indigenous data governance, Data Science Journal 19 (2020). doi:10.5334/dsj-2020-043.

[3] S. Schiff, R. Möller, S. Melzer, Tei-based interactive critical editions, in: S. B. Uchida, V. Elisaand Eglin (Eds.), Lecture Notes in Computer Science, Springer, 2022, pp. 230–244. doi:10.1007/978-3-031-06555-2_16.

[4] T. Parr, The Definitive ANTLR 4 Reference, Pragmatic Bookshelf, 2013.

[5] European Organization For Nuclear Research, OpenAIRE, Zenodo, 2013. URL: https://www.zenodo.org/. doi:10.25495/7GXK-RD71.

[6] M.-A. Sicilia, E. García-Barriocanal, S. Sánchez-Alonso, Community curation in open dataset repositories: Insights from zenodo, Procedia Computer Science 106 (2017) 54–60. doi:10.1016/j.procs.2017.03.009.

[7] METS, Metadata Encoding and Transmission Standard: Prime and Reference Manual, 2022. URL: https://www.loc.gov/standards/mets/METSPrimer.pdf.

[8] Kaggle, Find open datasets and machine learning projects, 2022. URL: https://www.kaggle.com/datasets.

[9] J. Gorraiz, P. Kraker, E. Lex, et al., Zenodo in the spotlight of traditional and new metrics, Frontiers 2 (2017). doi:10.3389/frma.2017.00013.

[10] GitHub, The ai-powered developer platform to build, scale, and deliver secure software, 2023. URL: https://github.com/.

[11] GitLab, About gitlab, 2023. URL: https://about.gitlab.com/.

[12] N. Nguyen, S. Nadi, An empirical evaluation of github copilot's code suggestions, in: Proceedings of the 19th International Conference on Mining Software Repositories, MSR '22, Association for Computing Machinery, 2022, pp. 1–5. doi:10.1145/3524842.3528470.