Humanities in the Center of Data Usability

Data Visualization in Institutional Research Repositories

Peukert, Hagen; Voges, Lucas Filipo; Asselborn, Thomas; Bender, Magnus; Möller, Ralf; Melzer, Sylvia

Published in:
Proceedings of the Workshop on Humanities-Centred Artificial Intelligence (CHAI 2024)

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Peukert, H., Voges, L. F., Asselborn, T., Bender, M., Möller, R., & Melzer, S. (2024). Humanities in the Center of Data Usability: Data Visualization in Institutional Research Repositories. In S. Melzer, H. Peukert, S. Thiemann, & E. Radisch (Eds.), *Proceedings of the Workshop on Humanities-Centred Artificial Intelligence (CHAI 2024): 4th Workshop at the 47th German Conference on Artificial Intelligence, 2024, Würzburg, Germany* (Vol. 3814, pp. 67-74). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3814/paper6.pdf

Humanities in the Center of Data Usability: Data Visualization in Institutional Research Repositories

Hagen Peukert[1], Lucas F. Voges[2,3], Thomas Asselborn[3,4,5], Magnus Bender[5], Ralf Möller[5] and Sylvia Melzer[3,4,5]

[1] University of Hamburg, Center for Sustainable Research Data Management, Monetastraße 4, 20146 Hamburg, Germany
[2] University of Hamburg, Institute of Food Chemistry, Grindelallee 117, 20146 Hamburg, Germany
[3] University of Hamburg, Cluster of Excellence 'Understanding Written Artefacts' (UWA), Warburgstraße 26, 20354 Hamburg, Germany
[4] University of Hamburg, Centre for the Study of Manuscript Cultures (CSMC), Warburgstraße 26, 20354 Hamburg, Germany
[5] University of Hamburg, Institute for Humanities-Centered AI (CHAI), Warburgstraße 28, 20354 Hamburg, Germany

Abstract

Research in the humanities is evolving with the introduction of data-driven methods and visualisation techniques that integrate different datasets such as images, videos, and texts. This change is supported by institutional research repositories that follow the FAIR principles (findability, accessibility, interoperability, and reusability) and fundamentally ensure that data is not only stored but also actively used for analyses. In contrast to conventional databases, which often lack interoperability, FAIR-compliant repositories generally improve the findability, reproducibility, and citation of individual data elements. In order to enable the reproducibility of data in special formats such as TEI (**T**ext **E**ncoding **I**nitiative), EpiDoc (**Epi**graphic **Doc**uments in TEI XML) or other project-specific formats according to the project-specific requirements for the visualization of the data, generic and user-friendly approaches are required, which an RDR (**R**esearch **D**ata **R**epository) should offer. This article demonstrates a new approach to data management using RDRs, offering the option to visualise data on a project-specific basis with just a click. Additionally, it explains how to cite not only the entire dataset within an RDR but also specific sections, ensuring clarity and precision by guiding readers to the exact information or argument referenced.

1. Introduction

Humanities research is undergoing a significant transformation as it embraces the potential of data-driven approaches and visualization techniques. The integration of different data sets, including research data with images, audios, videos and texts, opens up new ways of analysing and interpreting data in the humanities and other fields. The shift to data-driven and a new type of data management are exemplified by the growing importance of institutional research repositories that adhere to FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**euse) principles. These repositories are not mere data storage facilities but are intended to enable humanities scholars to use computer-aided methods and gain new insights. By combining textual data with materials science information or applying OCR (**O**ptical **C**haracter **R**ecognition) techniques to manuscript images, researchers can uncover patterns and connections previously hidden from view. However, this data-centred approach also brings challenges, particularly when it comes to ensuring that complex humanities data remains discoverable and accessible beyond simple keyword searches. The move towards FAIR-compliant repositories represents a departure from

Humanities-Centred AI (CHAI), 4th Workshop at the 47th German Conference on Artificial Intelligence, September 23, 2024, Würzburg, Germany
hagen.peukert@uni-hamburg.de (H. Peukert); lucas.voges@uni-hamburg.de (L. F. Voges); thomas.asselborn@uni-hamburg.de (T. Asselborn); magnus.bender@uni-hamburg.de (M. Bender); ralf.moeller@uni-hamburg.de (R. Möller); sylvia.melzer@uni-hamburg.de (S. Melzer)
https://www.chai.uni-hamburg.de/~asselborn (T. Asselborn); https://www.chai.uni-hamburg.de/~bender (M. Bender); https://www.chai.uni-hamburg.de/~moeller (R. Möller); https://www.chai.uni-hamburg.de/~melzer (S. Melzer)
0000-0002-3228-316X (H. Peukert); 0000-0001-8433-8321 (L. F. Voges); 0009-0005-3011-7626 (T. Asselborn); 0000-0002-1854-225X (M. Bender); 0000-0002-1174-3323 (R. Möller); 0000-0002-0144-5429 (S. Melzer)

traditional database models, which often lack interoperability and do not support appropriate citation of individual data elements. Instead, a modern RDR (**R**esearch **D**ata **R**epository) aims to provide seamless access to datasets for external computational processes while maintaining data integrity and citation. This approach not only improves the reproducibility of research, but also promotes interdisciplinary collaboration and places the humanities at the centre of data use and visualisation in the digital age.

2. Related Work

The general idea of the FAIR principles came from the workshop "Jointly designing a Data FAIRPORT" experts held in Leiden, the Netherlands, in 2014. They were first published in 2016 in [1]. The principles were developed to provide guidelines for improving the management and use of scientific data, particularly with regard to the machine readability of data and its availability for widespread use [2]. They are used worldwide, especially in the field of scientific research.

The use of DOI (**D**igital **O**bject **I**dentifier)s [3] to cite datasets has become a widely accepted method to ensure that datasets are easily identifiable, accessible, and citable. Data citation via DOI supports the findability and accessibility aspects of these principles by providing a unique and persistent identifier for datasets. Multiple studies emphasize that without proper data citation practices, datasets can be difficult to locate or may become inaccessible over time, particularly when managed by individual researchers or institutions with limited resources [2]. DOIs ensure that data remains findable even if its physical location changes, as the DOI will always redirect users to the dataset's current location.

RDRs, such as Zenodo [4] and the University of Hamburg (UHH)'s RDR (RDR@UHH) [5], integrate DOI systems to enhance data management and promote the open sharing of research results. The challenges that still exist, however, are that the data is made findable and citable by means of DOI or RDRs, but the visualization of the data is not always guaranteed, especially when special formats are used for data representation such as TEI (**T**ext **E**ncoding **I**nitiative).

With regard to citations, a DOI can be used to refer to the entire entry, but for a precise citation, the book page or similar must be specified in addition to the DOI.

3. Data Viewer in Institutional Research Repositories

One of the significant challenges in reusability arises when researchers develop specialized software tools or viewers for data analysis. These custom tools are often tailored to specific datasets or formats, making the data difficult to reuse by others who do not have access to or cannot maintain these tools. Over time, these tools can become obsolete, leading to issues in data reanalysis or reinterpretation. This highlights the importance of using standardized, widely supported formats and generic approaches for ensuring long-term reusability. In addition, reliance on proprietary software or closed formats exacerbates this problem. If a dataset is locked into a proprietary system, it can be difficult for future users to access or edit the data without incurring costs or overcoming technical barriers. This risk increases over time if software vendors discontinue support for older versions or the software itself is no longer available. Therefore, institutional repositories should promote the use of open-source tools and widely accepted formats, such as CSV for tabular data, to ensure long-term accessibility. Alternatively, they can customize institutional RDRs to support the integration of data viewers tailored to project-specific formats.

4. Data Visualisation in Humanities Research

Data visualization in the humanities is increasingly essential as researchers seek to communicate their information effectively. This shift is particularly relevant given the growing use of structured data formats like EpiDoc (**Epi**graphic **Doc**uments in TEI XML), which allows for the encoding of textual data related to ancient documents and inscriptions. While EpiDoc facilitates the creation of machine-readable datasets, the challenge arises when these datasets are uploaded to RDRs and become less accessible

to a broader audience. To fulfil the FAIR principles, data needs to be clearly communicated to human users, among other things, so that its relevance and applicability for specific projects can be assessed among others. Therefore, effective visual representations of research data are essential, moving beyond mere directory listings or standard table displays. Furthermore, the presentation of datasets must guarantee that individual data points, which are consistently available in RDRs, can be properly cited by researchers and other users. Citations should be easy to trace, meaning that upon accessing a citation DOI, users should be able to effortlessly locate the referenced data item.

At the UWA (**U**nderstanding **W**ritten **A**rtefact) Cluster of Excellence, scholars collaborate with computer scientists to customize generic viewers to meet specific project needs or create unique viewers tailored to particular datasets. Scholars can compile their research data into archives known as CSMC files, which are comparable to DOCX files. A CSMC file is a new specialized format designed to manage and present research data within the context of humanities scholarship, particularly in the study of manuscripts and related materials. This format enables scholars from a wide range of disciplines to encapsulate their research data in a structured way that is both user-friendly and suitable for digital environments. These CSMC files can then be uploaded to the RDR by scholars, providing project-specific visualizations accessible directly through the UHH's RDR. In partnership with computer scientists, CSMC dataset generators were developed to assist researchers, enabling them to create CSMC datasets independently, without the need for direct assistance from computer scientists. Once a CSMC dataset file is submitted to the RDR, the generated RDRs facilitate easy access to the web presentations of the data.

The process for creating a CSMC file for a project operates on a web server and is developed in collaboration with a computer scientist. Once the initial setup is complete, scholars can independently generate new dataset packages. Using the software technology developed by UWA in Research Field F which is the CSMC App[1], humanities researchers can first review their data in CSMC files locally on their computers. This allows scholars without internet access—such as those working on excavation trips—to utilize UWA datasets if they have a CSMC file and the installer for the CSMC App (e.g., provided on a DVD).

Scholars can easily make their CSMC files publicly accessible by submitting them to the RDR system at the UHH. Once a dataset is submitted, it is automatically recognized in the RDR, and a "View Data" button will appear, allowing any user to access the data online. Thus, by submitting their data to the UHH's RDR, scholars receive an online representation of their data in a format they have defined, thanks to initial collaboration with computer scientists to develop project-specific views based on generic components and to make the CSMC file generator available.

5. Application and Results

We have practically implemented the new approach to visualising research data from the RDR in a wide variety of projects. Three projects NETamil [6], EDAK [7], and FTIR [8] are described below.

NETamil

The project NETamil, under the title "Going From Hand to Hand: Networks of Intellectual Exchange in the Tamil Learned Traditions" is a research project dedicated to investigating the extensive intellectual history of Tamil literature and its transmission across centuries. Launched in March 2014 with funding from the European Union's Seventh Framework Programme, the project focuses on reconstructing the processes of interaction and knowledge exchange within Tamil intellectual traditions, particularly in the period preceding the so-called *Tamil renaissance* of the 19th century. [9]

At https://doi.org/10.25592/mdq0-7x79 a published CSMC dataset from the NETamil project is shown in RDR as an example of work in the UWA. Clicking on the DOI link above, an archived CSMC file can

[1] download for Windows https://csmc-view.chai.uni-hamburg.de/local-app/windows-latest, for macOS (Apple Silicon) https://csmc-view.chai.uni-hamburg.de/local-app/macos-latest

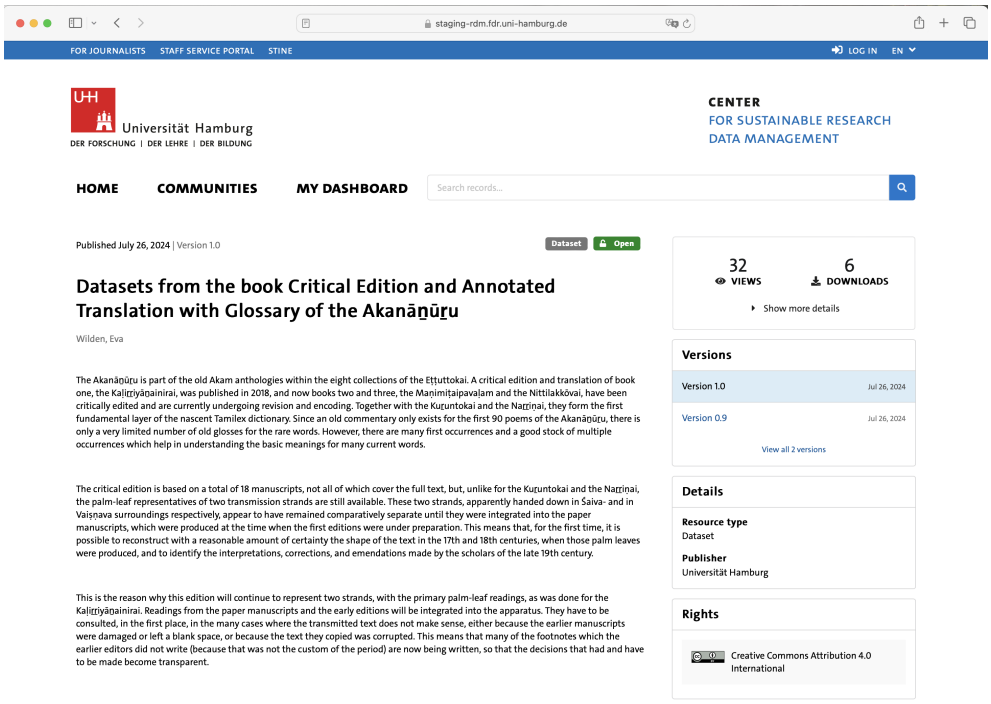

Figure 1: Archived research data in the RDR from the NETamil project

be seen, which was submitted to the RDR@UHH (see Figure 1). After clicking on View Data (below the feather icon, please scroll down and look under Files, if required) the data will be displayed as intended by the scholars who created the dataset (see Figure 2). The link behind the View Data button can also be put on a web page as shown here.

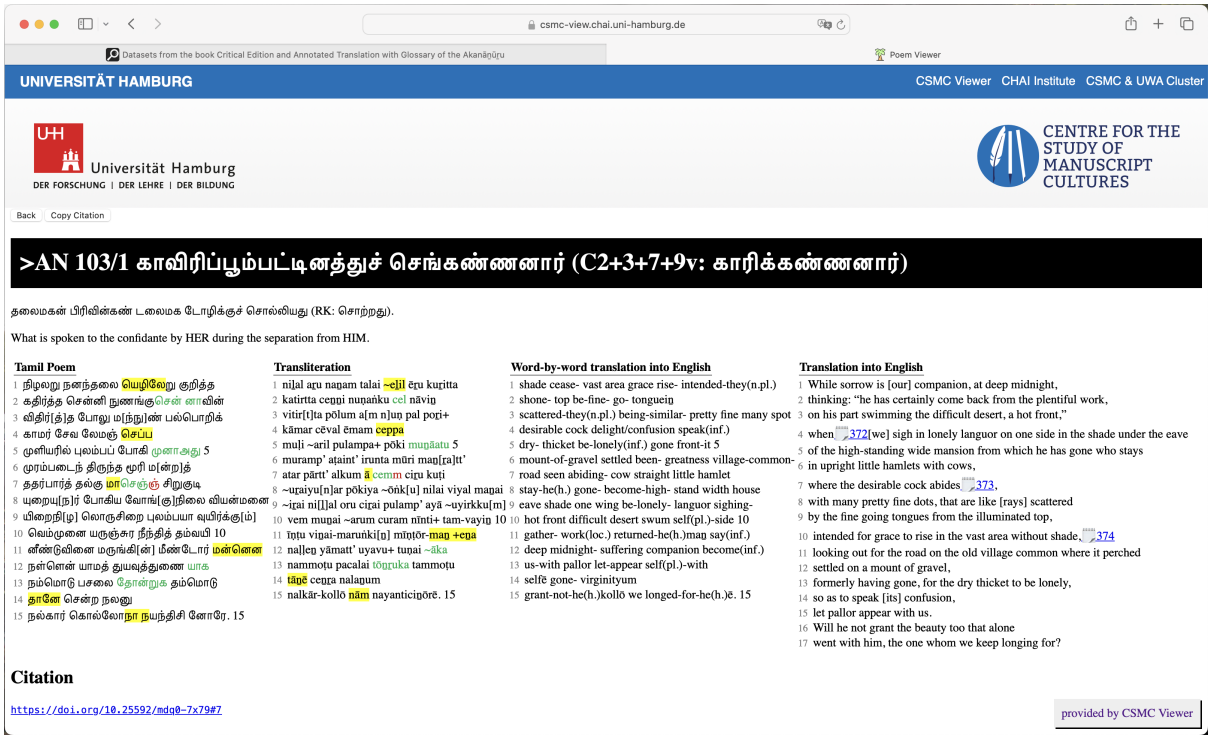

Figure 2: Data Viewer of archived research data from the NETamil project

EDAK

The EDAK (Epigraphische Datenbank zum Antiken Kleinasien) project at the UHH, focuses on creating a comprehensive digital epigraphic database for inscriptions from ancient Asia Minor. The project's primary goal is to document and preserve the rich epigraphic heritage of this region by providing detailed, accessible records of inscriptions, which are crucial for the study of ancient languages, cultures, and history. [7]

An example of relational data from this DOI https://doi.org/10.25592/5mx6-1k15, see Figure 3, is available for the EDAK project. After clicking on "View Data" (see Figure 3 below the feather icon), the

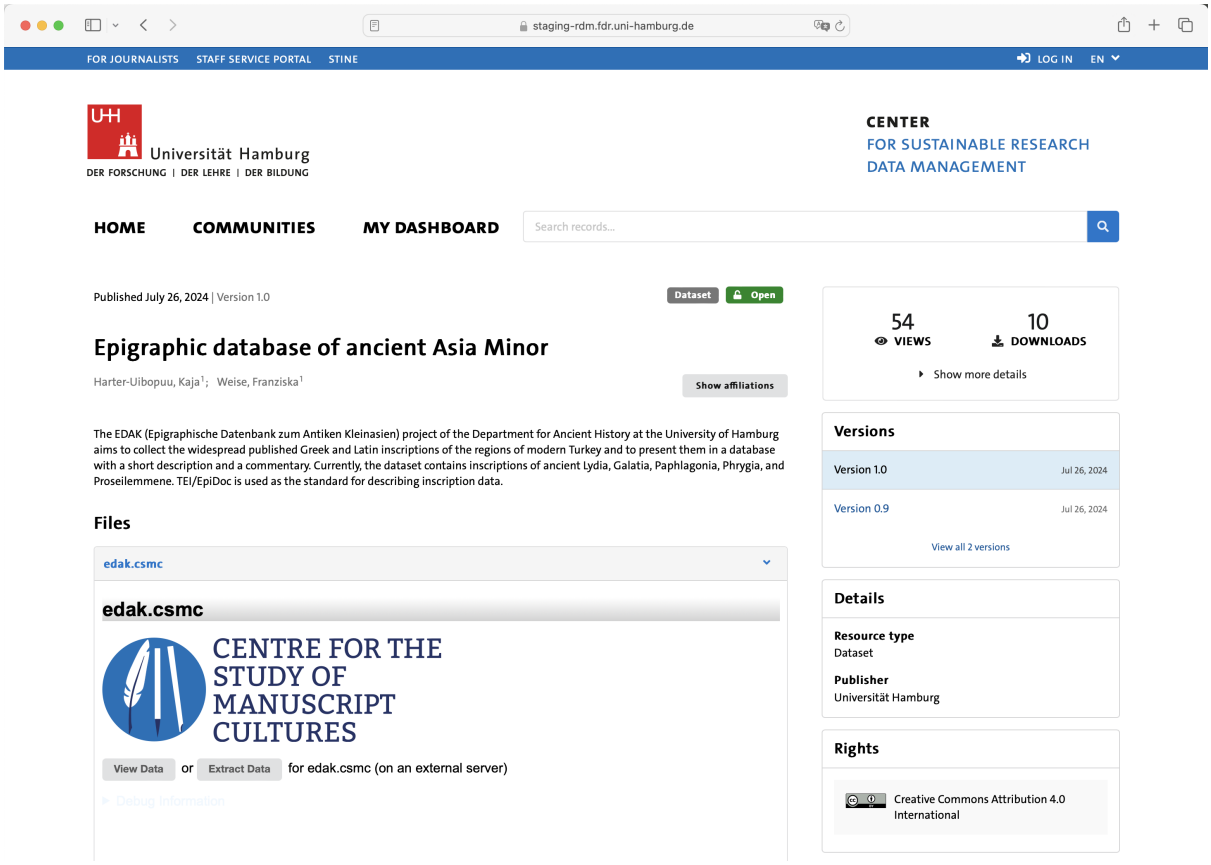

Figure 3: Archived research data in the RDR from the EDAK project

data will be displayed as the creating scholars intended (see Figure 4). Filtering and navigation facilities for the dataset are provided in this case. Clicking on a row shows a detailed view of the clicked data item. The link behind the "View Data" button can also be integrated into a web page.

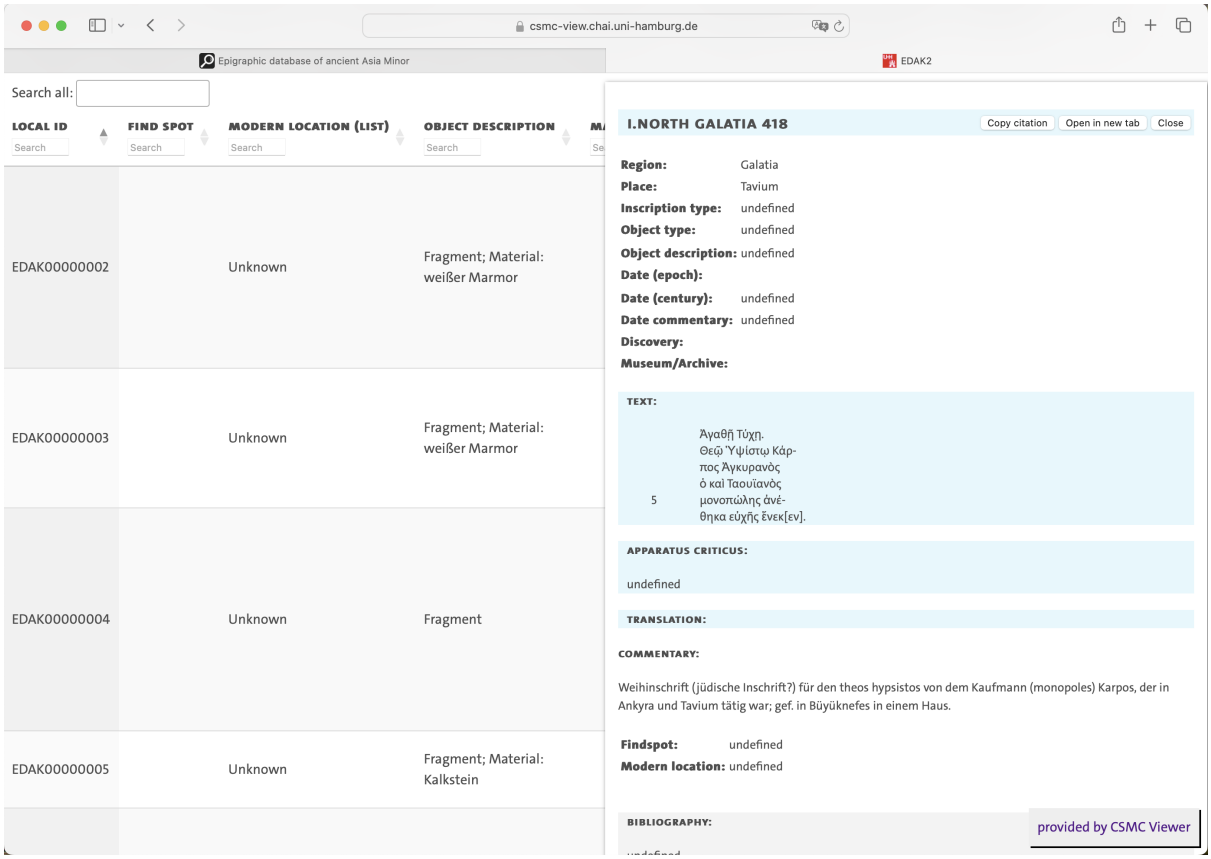

Figure 4: Data Viewer of archived research data from the EDAK project

FTIR

Application of infrared spectroscopy in combination with chemometrics as a fingerprint technique can be utilized to study the materiality of manuscripts. Within the Palm-Leaf Profiling Initiative (PLMPI) we demonstrated that Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS) can be used to obtain manuscript specific information and that a differentiation regarding the taxonomic species of the palm leaves as well as their geographical origin in South and Southeast Asia was possible.

In the analysis pipeline we use the open-source scientific and technical publishing system "Quarto" [2] to document and publish the results. The interactive visualisation helps researchers to better understand the data, and we can communicate the results in a more conveniently accessible form than with a research paper written for a specialised field.

With the here described data viewer, we can now make the visualisation easily available to researchers and share the data at the same time. To especially uphold the "Reuse" of FAIR principles, the corresponding scripts (R, Python, …) can be made part of the shared data, so that the interested scientist can rebuild or even expand the visualisation and analysis by themselves or customize it to their needs.

In Figure 5 a CSMC file containing a dataset from FTIR spectroscopy is shown as another example using the CSMC App (see Figure 5). See https://staging-rdm.fdr.uni-hamburg.de/records/rp3g0-6zy30 for the RDR entry.

Citations

External users interested in specific data from RDR datasets can also cite an individual data record (and not just the entire archive via the RDR system). See, e.g., the citation links https://doi.org/10.25592/mdq0-7x79#8 or https://doi.org/10.25592/5mx6-1k15#1033 for citation DOIs. The DOI links can be

[2] https://quarto.org/

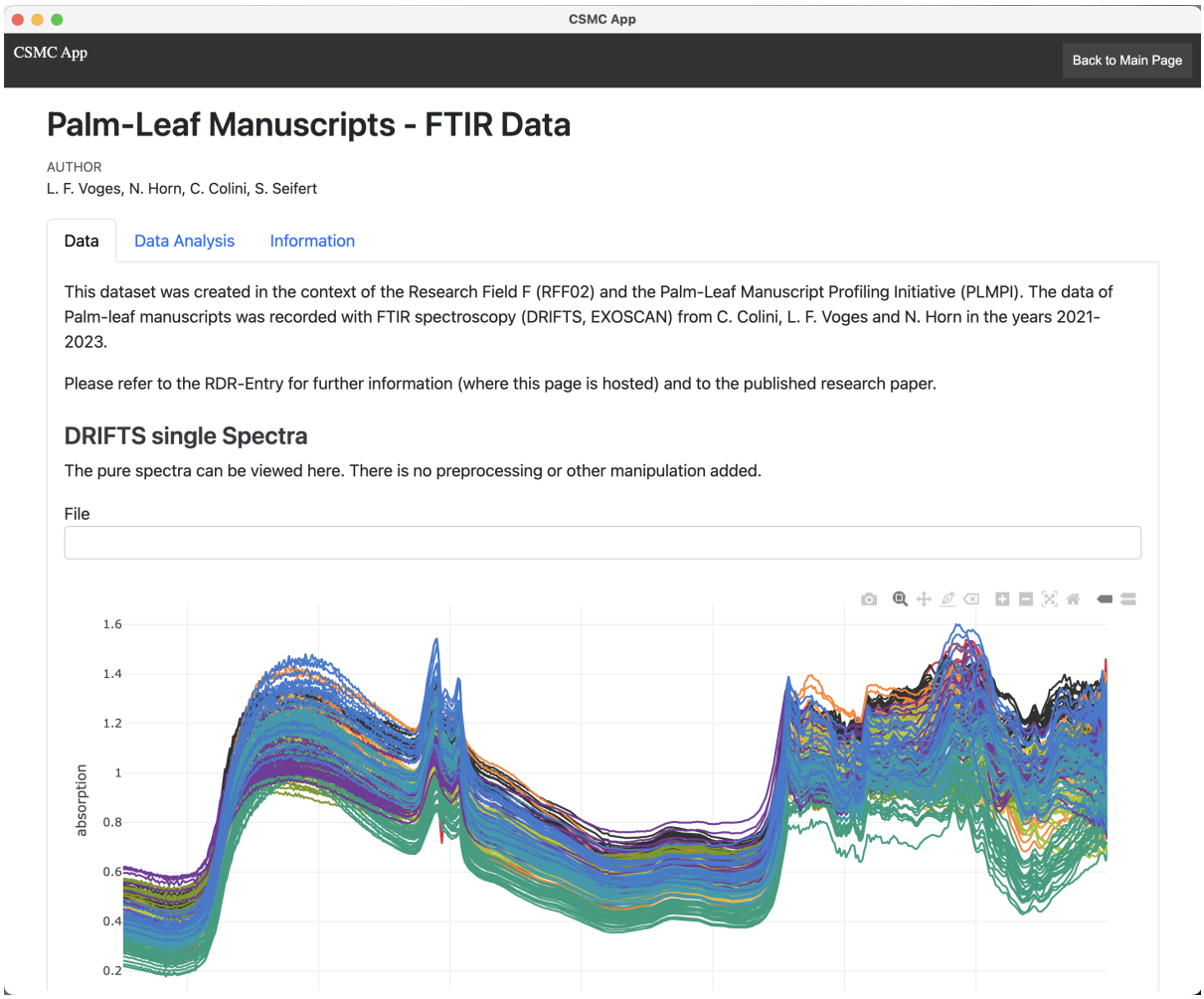

Figure 5: View research data in the CSMC APP the FTIR project

clicked or copied into the input window of a browser. Once the dataset is displayed in RDR, the "Show Citation" button, located just below the feather icon, should be clicked. Citation links can be obtained by clicking on a data item, which will then show the details, and the "Copy Citation" button can be clicked to copy the citation.

Keeping data in a database, however, as was often pursued in the past, and having a web interface built for data presentation, does not lead to citable presentations because databases can be changed, and thus, citations then become pointless. Citing datasets in the new way, as described above, makes sense because with RDR submission the data is persistent and cannot be changed (one can only upload new versions, but the old ones are retained).

6. Conclusion and Outlook

This article emphasises the importance of effective RDM (**R**esearch **D**ata **M**anagement). RDM focuses on making datasets accessible, citable and usable by integrating project-specific data presentations into an institutionalised RDR and providing tools such as the CSMC App to view the data locally. The institutional RDR@UHH allows researchers to create archives with unique views of their data, which can then be submitted to the RDR@UHH. The initiative aims to improve the transparency and usability of research data while ensuring correct citation through the persistent identifier DOI.

This new procedure has been used for some projects, and we are campaigning for other RDRs to adopt this functionality in the future so that the public around the world has access to this functionality.

Funding Information

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796. The research was mainly conducted within the scope of the Centre for the Study of Manuscript Cultures (CSMC) at Universität Hamburg.

References

[1] FAIR Data Principles, Data FAIRport conference - JOINTLY DESIGNING A DATA FAIRPORT, 2016. URL: https://www.datafairport.org/component/content/article/8_news/9_item1/index.html, accessed: 2024-09-09.

[2] M. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. G. Dumon, S. C. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. N. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. A. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. M. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, B. Mons, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018–160018. doi:10.1038/SDATA.2016.18.

[3] DOI Foundation, DOI HANDBOOK, DOI Foundation, info@doi.org, 2023. URL: https://www.doi.org/doi-handbook/DOI_Handbook_Final.pdf.

[4] Zenodo, Zenodo - Research. Shared., 2023. URL: https://zenodo.org/, accessed: 2024-09-10.

[5] Universität Hamburg, Research Data Repository, Available: https://www.fdr.uni-hamburg.de/, 2024. Accessed September 9, 2024.

[6] Universität Hamburg, NETamil - Going From Hand to Hand: Networks of Intellectual Exchange in the Tamil Learned Traditions, Available: https://www.csmc.uni-hamburg.de/ends/projects/netamil.html, 2019. Accessed September 9, 2024.

[7] University of Hamburg, Epigraphische Datenbank zum antiken Kleinasien, 2013-2016. URL: https://www.epigraphik.uni-hamburg.de/content/index.xml.

[8] L. F. Voges, N. Horn, C. Colini, S. Seifert, Ftir spectra of 11 palm-leaf manuscripts, 2024. URL: https://doi.org/10.25592/uhhfdm.14774. doi:10.25592/uhhfdm.14774, The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796. The research was conducted within the scope of the Centre for the Study of Manuscript Cultures (CSMC) at Universität Hamburg.

[9] Centre for the Study of Manuscript Cultures, University of Hamburg, NETamil: Going From Hand to Hand: Networks of Intellectual Exchange in the Tamil Learned Traditions, 2014. URL: https://www.csmc.uni-hamburg.de/ends/projects/netamil.html, accessed: 2024-09-10.