1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.297

Published: 07 Sept. 2023

Digital Twin-Based Concept for Reliable Research Data Management

Integrating Proprietary Data Sources for Hyperspectral Imaging

Alessa Rache[1,2] [https://orcid.org/0000-0001-7598-8672],
Tim Häußermann[1,2] [https://orcid.org/0000-0002-4020-4089],
Joel Lehmann[1,2] [https://orcid.org/0000-0001-8261-8362], and
Julian Reichwald[1,2] [https://orcid.org/0000-0002-4809-5710]

[1]Center for Mass Spectrometry and Optical Spectroscopy

[2]Mannheim University of Applied Sciences

Abstract: In data-intensive research, reliable management of research data is a major challenge. In the field of Mass Spectrometry Imaging, vast amounts of data are being acquired from mostly proprietary data sources. Consequently, hindering seamless data integration into Research Data Management systems. Without a data repository, the continuous generation of scientific knowledge and innovative research based on existing information is limited. Moreover, to maintain the value of data to researchers throughout and beyond its lifecycle, FAIR principles for reliable data management approaches must be applied. To enable the required data transmission, the Digital Twin paradigm can be considered a reliable solution. The conceptual implementation of a heterogeneous mass spectrometer generating hyperspectral images leverages the Digital Twin to overcome common data management problems in data-intensive research.

Keywords: Research Data Management, Research 4.0, FAIR, Digital Twin, Container Digital Twin, Cyber-Physical System, Knowledge Graph, Ontology

1 Introduction

Reliable management of research data is becoming increasingly important, especially when dealing with data-intensive research [1]. As such, Mass Spectrometry Imaging (MSI) uses a combination of molecular mass analysis and spatial distribution to study the allocation of molecules present in the samples examined, visually mapped by hyperspectral images [2], [3]. This allows the generation of mass spectra for each single spot and consequently the acquisition of thousands of individual mass spectra per examination [4]. Accordingly, laboratories utilizing MSI generate vast amounts of data, causing an urgent need for extensive efforts regarding Research Data Management (RDM) [5], [6].

The primary objective of RDM is to pave the way for new scientific knowledge and enable innovative research based on existing information [7]. In response to emerging

efforts to reform research communication systems, Force 11 established the FAIR Principles in 2016. These principles are intended to serve as a guide seeking to improve the reusability of data, according to which data is expected to be Findable, Accessible, Interoperable, and Reusable to maintain its intrinsic value to researchers throughout the entire data lifecycle and beyond [8].

In reality, research data is commonly stored on local computers or on offline data repositories, which raises major concerns about the reproducibility of scientific research results [9]. The lack of standardization among the numerous software and hardware vendors contributes to the prevailing handling of data, complicating seamless integration. In practice, proprietary data formats and a high degree of heterogeneity across different devices are common [8]. This reality constrains the establishment of reliable RDM and the subsequent development of a knowledge base of relevance for the research community [10].

Research is increasingly adopting techniques raised by Industry 4.0 gearing itself up for Research 4.0 [11]. As an innovative technology for data transmission, the Digital Twin (DT) can be seen as a secure data source as it mirrors physical devices into the digital world through a bilateral communication stream [12], thus enabling the digital use and management of data [13].

2 Conceptual Architecture

In the preliminary work [14], [15] we proposed an RDM infrastructure offering extensive capabilities for storing and preserving research data in accordance with FAIR criteria utilizing DTs as well as agent-based DTs.

Figure 1 depicts an extension of the concept to overcome the limitation of proprietary measuring devices impeding automated data aggregation. The concept is divided into four segments: Physical Twin Space, Digital Twin Space, RDM Core Space and Smart Application Space.

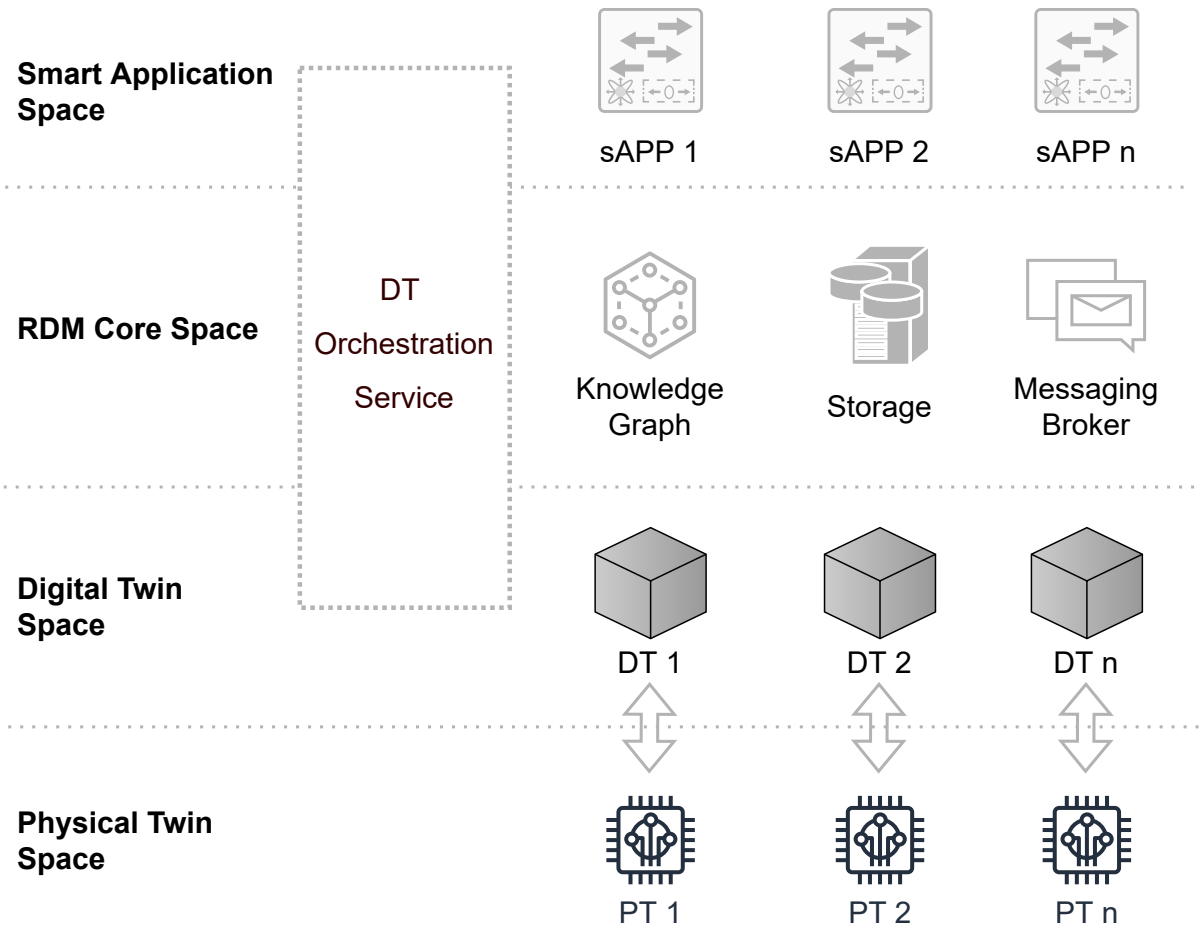

Figure 1. Conceptual RDM Architecture targeting proprietary devices without an open interface

The Physical Twin Space holds physical devices such as mass spectrometers (PT_1 - PT_n), each of having a unique DT (DT_1 - DT_n) which is provided using the existing infrastructure in the Digital Twin Space. Due to the proprietary nature of the devices, containerized DTs are utilized. These are distinguished by their portability, distributability and cloud independence from other types of DTs. As a result, they prevent interface- and accessibility-issues. For this purpose, the container-based DT is movable and thus can be run on the workstation which is associated to the corresponding physical device. Based on the lack of open interfaces, DTs cannot be instantiated by PTs themselves but must be obtained using a interface for specification. This interface is provided through the microservice DT Orchestration Service (DTOS). It provides the foundation for the execution of cross-layer tasks and communication. The DTOS deploys a container-based DT bound to the associated workstation, which then monitors the corresponding paths of relevant research data and handles the further storage procedures. The additions made to the DTOS and the use of the container-based DT constitute the extension of the architecture enabling connectivity and mapping of proprietary devices.

The workflow for the acquisition of hyperspectral data is as follows. Once the container-based DT is instantiated, the paths of the workstation are scanned continuously as new data is created. The container-based DT then recognizes the generation and launches a graphical user interface (GUI) on the workstation. The GUI includes several inputs needed to save the data in the Storage enriched by FAIR compliant metadata. After

completing all required entries, the container-based DT independently stores the data within the RDM architecture.

3 Conclusion

Conservatively decentralized handling of research data is attributed to prevailing proprietary data sources without open interfaces. Data-intensive research generates massive amounts of data which needs to be managed properly. To address this issue, a holonic infrastructure for reliable RDM using container-based DTs to integrate proprietary data sources is illustrated and exemplified by the research area of MSI.

The concept is generally applicable to any research area where seamless data transfer and reliable access to research data from proprietary physical data sources pose a major challenge such as environmental or physical sciences. The heterogeneous devices are mapped by container-based DTs that crawl and store research data into a central and structured FAIR-compliant RDM. Considering scientific knowledge generation and reusability, this concept opens up entirely new possibilities for exploiting the extensive potential of digitized RDM in research. The next logical step is to extend the concept to interdisciplinary research and the associated challenges, such as collaborative data use and related security concerns. To fully exploit the potential of the concept, a comprehensive implementation in the future is essential to evaluate its usability.

4 Appendix

Data availability statement

Not applicable.

Author contributions

Conceptualization, A.R., T.H., J.L.; methodology, A.R.; investigation, A.R., T.H, J.L.; writing-original draft preparation, A.R., T.H, J.L.; supervision, J.R.; project administration, A.R.; All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

Parts of this work presented in this paper were supported by a grant from the German Ministry of Education and Research (BMBF), grant number 16FDFH125.

References

[1] J. Gray, D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber, "Scientific data management in the coming decade," *ACM SIGMOD Record*, vol. 34, no. 4, pp. 34–41, Dec. 2005, ISSN: 0163-5808. DOI: 10.1145/1107499.1107503. [Online]. Available: https://doi.org/10.1145/1107499.1107503.

[2] E. R. Amstalden van Hove, D. F. Smith, and R. M. Heeren, "A concise review of mass spectrometry imaging," en, *Journal of Chromatography A*, vol. 1217, no. 25, Jun. 2010, ISSN: 00219673. DOI: 10.1016/j.chroma.2010.01.033. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0021967310000701 (visited on 12/13/2022).

[3] F. Grélard, D. Legland, M. Fanuel, B. Arnaud, L. Foucat, and H. Rogniaux, "Esmraldi: Efficient methods for the fusion of mass spectrometry and magnetic resonance images," *BMC Bioinformatics*, vol. 22, no. 1, p. 56, Feb. 2021, ISSN: 1471-2105. DOI: 10.1186/s12859-020-03954-z. [Online]. Available: https://doi.org/10.1186/s12859-020-03954-z.

[4] A.-M. Lahesmaa-Korpinen, S. M. Carlson, F. M. White, and S. Hautaniemi, "Integrated data management and validation platform for phosphorylated tandem mass spectrometry data," *PROTEOMICS*, vol. 10, no. 19, 2010, ISSN: 1615-9861. DOI: 10.1002/pmic.200900727. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200900727.

[5] P. Romano, A. Profumo, M. Rocco, R. Mangerini, F. Ferri, and A. Facchiano, "Geena 2, improved automated analysis of MALDI/TOF mass spectra," *BMC Bioinformatics*, vol. 17, no. 4, p. 61, Mar. 2016, ISSN: 1471-2105. DOI: 10.1186/s12859-016-0911-2. [Online]. Available: https://doi.org/10.1186/s12859-016-0911-2.

[6] O. J. R. Gustafsson, L. J. Winderbaum, M. R. Condina, *et al.*, "Balancing sufficiency and impact in reporting standards for mass spectrometry imaging experiments," *GigaScience*, vol. 7, no. 10, Oct. 2018, ISSN: 2047-217X. DOI: 10.1093/gigascience/giy102. [Online]. Available: https://doi.org/10.1093/gigascience/giy102.

[7] A. Whyte and J. Tedds, "Making the Case for Research Data Management," *Digital Curation Centre Jisc Briefing Paper*, Sep. 2011. [Online]. Available: https://www.researchgate.net/publication/252931138_Making_the_Case_for_Research_Data_Management.

[8] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," en, *Scientific Data*, vol. 3, no. 1, p. 160 018, Mar. 2016, Number: 1 Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. [Online]. Available: https://www.nature.com/articles/sdata201618.

[9] M. Diepenbroek, F. O. Glockner, P. Grobe, *et al.*, "Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio)," p. 11, [Online]. Available: https://www.researchgate.net/publication/267574356_Towards_an_Integrated_Biodiversity_and_Ecological_Research_Data_Management_and_Archiving_Platform_The_German_Federation_for_the_Curation_of_Biological_Data_GFBio.

[10] B. Mons, *Data Stewardship for Open Science: Implementing FAIR Principles*. New York: Chapman and Hall/CRC, Feb. 2018, ISBN: 978-1-315-38071-1. DOI: 10.1201/9781315380711.

[11] E. Jones, N. Kalantery, and B. Glover, "Research 4.0: Interim report," en, Demos, Report, Oct. 2019. [Online]. Available: https://apo.org.au/node/262636.

[12] M. Grieves, *Origins of the Digital Twin Concept*. Aug. 2016. DOI: 10.13140/RG.2.2.26367.61609. [Online]. Available: https://www.researchgate.net/publication/307509727_Origins_of_the_Digital_Twin_Concept.

[13] T. P. Raptis, A. Passarella, and M. Conti, "Data Management in Industry 4.0: State of the Art and Open Challenges," *IEEE Access*, vol. 7, pp. 97 052–97 093, 2019, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2929296. [Online]. Available: https://ieeexplore.ieee.org/document/8764545.

[14] J. Lehmann, S. Schorz, A. Rache, T. Häußermann, M. Rädle, and J. Reichwald, "Establishing reliable research data management by integrating measurement devices utilizing intelligent digital twins," *Sensors*, vol. 23, no. 1, 2023, ISSN: 1424-8220. DOI: 10.3390/s23010468. [Online]. Available: https://www.mdpi.com/1424-8220/23/1/468.

[15] J. Lehmann, A. Lober, T. Häußermann, *et al.*, "The anatomy of the internet of digital twins: A symbiosis of agent and digital twin paradigms enhancing resilience (not only) in manufacturing environments," *Machines*, vol. 11, no. 5, 2023, ISSN: 2075-1702. DOI: 10.3390/machines11050504. [Online]. Available: https://www.mdpi.com/2075-1702/11/5/504.