






Cooking With DOIs

A Recipe for FAIR Digital Specimens Using FDOs

Soulaine Theocharides^{1,*} , Wouter Addink¹ , Sharif Islam¹ , Matt Buys² ,
and Sara El-Gebali² 

¹Naturalis Biodiversity Center, Netherlands

²DataCite, Germany

*Correspondence: Soulaine Theocharides, soulaine.theocharides@naturalis.nl

Abstract. In the global biodiversity crisis, robust data management is growing increasingly important in the sector. The Digital Specimen is a response to this need, facilitating important connections between infrastructures. The FAIR Digital Object (FDO) paradigm offers a further machine-actionable solution to data management. One FDO component is assigning of unique, persistent identifiers to the object, and enriching these identifiers with a machine-readable “FDO Record” to further describe the referent. This conference presentation discussed Digital Specimens as FDOs, and how assigning persistent identifiers with FDO records was achieved through a partnership between DiSSCo and DataCite.

Keywords: FAIR, Digital Specimen, Persistent Identifiers, FAIR Digital Objects

1. Digital Specimens as FAIR Digital Objects

Digital Specimens are FAIR Digital Objects (FDO) derived from data about specimens in natural science collections. These resources act as a digital surrogate for the physical specimen over the Internet, enabling modern analytical research practices using machines and Artificial Intelligence [1]. These Digital Specimens are linked to related data such as genetic, morphological, chemical, environmental and taxonomical data available elsewhere on the Internet. These linkages provide important context for the specimen and empower interdisciplinary research [2].

In order to reliably create links between specimen data and data held in other infrastructures, this data needs to be equipped with globally unique, resolvable and persistent identifiers (PIDs). To create and manage such PIDs at scale, a robust infrastructure is needed. PIDs are a key component of FAIR data principles (“F1. (Meta)data are assigned a globally unique and persistent identifier” [3]), facilitating discovery and ensuring the unambiguous reference to a resource of interest. They are also a key component of the Digital Specimen paradigm, as illustrated by Figure 1.

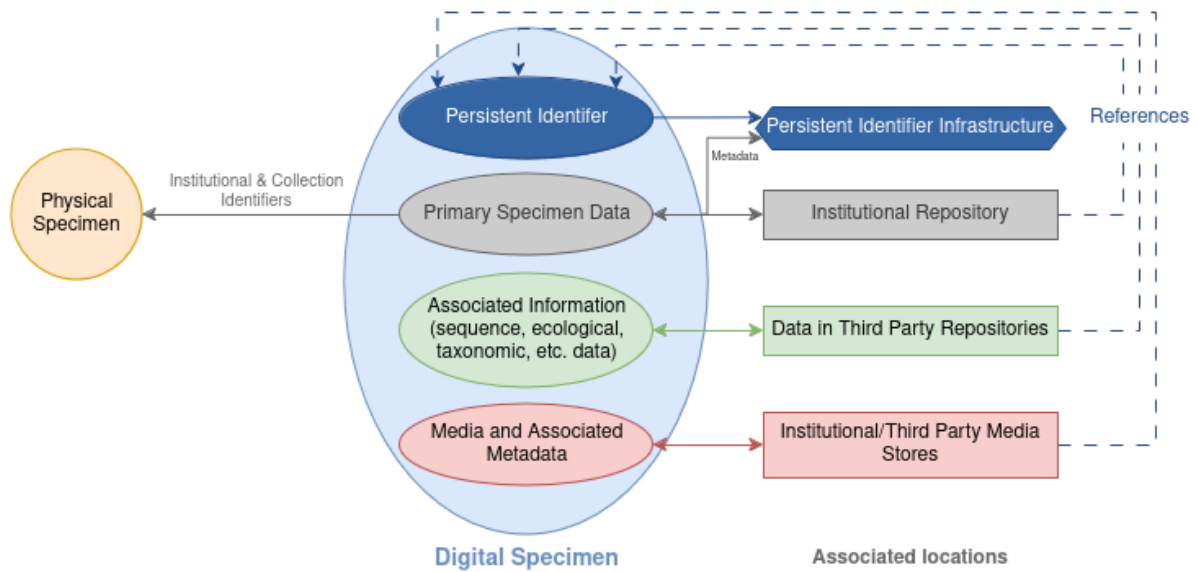


Figure 1. The Digital Specimen as a surrogate for a physical specimen, with linkages facilitated through use of a Persistent Identifier Resolution System. At the center of the diagram is the Digital Specimen and its core components. The PID is stored in the PID Infrastructure, and is informed by metadata (including institutional identifiers) from the primary specimen data. The primary specimen data is derived from the physical specimen, and is linked to the Digital Specimen through institutional and collection identifiers. Associated data and media objects are stored in third party or institutional repositories. Institutional and third party repositories should reference the Persistent Identifier, creating bi-directional linkages, though this functionality has yet to be realized. Adapted from [4].

When a PID is associated with the location of the digital resource, as is the case with Handles and DOIs, the identifier becomes resolvable and improves findability across infrastructures and literature. When the PID is associated with additional, structured metadata, machines can better interact with the digital resource as the metadata can provide context allowing the machine to make decisions before resolving the resource [5]. PID metadata records can be structured into one of the foundational components of FAIR Digital Objects: the FDO Record.

Under the auspices of the BiCIKL Project (<https://bicikl-project.eu/>) funded by the European Union, a PID infrastructure was developed and adopted by the Distributed System of Scientific Collections (DiSSCo) as the basis of its FDO-based Digital Specimen infrastructure [6]. DiSSCo is a Research Infrastructure that aims to unify natural science collections data across Europe. Through data aggregation, unification, and annotation, DiSSCo is leveraging the FAIR Digital Object paradigm to make Digital Specimens FAIR and machine-actionable [7] [8]. Within DiSSCo, each digital object benefits from an atomic description, ensuring a PID and a minimum set of metadata attributes, enhancing their discoverability and interoperability.

2. Persistent Identifier Infrastructure

The Handle System is the foundation for the Digital Specimen PID infrastructure. Developed by CNRI, the Handle System is a globally distributed infrastructure for minting and managing PIDs for digital resources [9]. One advantage of the Handle System over other infrastructures is the existence of a Handle Record that contains the location of the resource, which can be further extended to meet FDO requirements. According to FDO specifications, an FDO must have a PID that resolves to a PID record, "a structured record...compliant with a specified PID Profile which leads to resolution results that enable programmatic resolution from PID back to the FDO and its elements[.]" [10]. DiSSCo has leveraged the technical capabilities of the Handle system to store additional metadata in the Handle Record for each PID to become

FDO-compliant. Metadata in the Handle Record adheres to Type-specific PID Profiles developed under DiSSCo. Because these Profiles are tailored to different Types of digital objects, this FDO implementation enables extension and adaptability of different objects within the infrastructure.

To further improve trust and reliability of the Digital Specimen PID Infrastructure and to ensure persistence, DiSSCo is implementing Digital Object Identifiers (DOIs) for Digital Specimens. DOIs are built on the Handle resolution system (so they are also Handles), and they are used by both humans and machines to reliably refer to objects. Under the oversight of the DOI Foundation, DOIs guarantee persistence, even if the original registrant of the PID is no longer operational. The DOI Foundation governs the DOI system on behalf of the agencies who manage DOI registries and provide services to their respective communities. This ensures that committed organisations are in compliance with the DOI policies and best practices [11].

DataCite is one such organisation. It is a nonprofit organisation that provides open infrastructure and facilitates registration of DOIs for research outputs and resources, and creates links between these. Because of their aligned mission statements, DiSSCo and DataCite are partnering to explore registration of FDO-compliant DOIs at scale for potentially billions of Digital Specimens. Over the course of this partnership, a mapping has been developed between the DiSSCo PID profiles and the DataCite metadata schema, so the FDO record may be exposed more broadly and comply with DataCite requirements.

The FDO-compliant PID Infrastructure developed under the BiCIKL project is a step in the direction of an integrated network of biodiversity data across domains. In this session, we provide a technical overview of the PID Infrastructure, with a focus on the implementation of Digital Specimens as FAIR Digital Objects. Additionally, we shall discuss the developments of the DiSSCo-DataCite partnership and the state of DOIs for Digital Specimens.

Data availability statement

The infrastructure and deployment code of the described infrastructure is available on GitHub: <https://github.com/DiSSCo/doi-infrastructure>

Author contributions

Soulaine Theocharides: Writing – original draft, review & editing, Software. **Wouter Addink:** Writing – review & editing, Project administration. **Sharif Islam:** Writing – review & editing. **Matt Buys:** Writing – review & editing. **Sara El-Gebali:** Writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101007492 (BiCIKL project)

References

- [1] M. Webster, J. Buschbom, A. Hardisty, and A. Bentley, "The Digital Extended Specimen will Enable New Science and Applications," *Biodivers. Inf. Sci. Stand.*, vol. 5, p. e75736,

- Sep. 2021, doi: 10.3897/biss.5.75736.
- [2] J. Lendemer *et al.*, "The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education," *BioScience*, vol. 70, no. 1, pp. 23–30, Jan. 2020, doi: 10.1093/biosci/biz140.
- [3] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
- [4] A. R. Hardisty *et al.*, "Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet," *BioScience*, vol. 72, no. 10, pp. 978–987, Oct. 2022, doi: 10.1093/biosci/biac060.
- [5] T. Weigel *et al.*, "RDA Recommendation on PID Kernel Information," 2018, doi: 10.15497/RDA00031.
- [6] W. Addink, S. Islam, M. Dillen, A. Güntsch, and S. Theocharides, "Deliverable D7.1 Architecture Design for a pan-European PID system for Digital Specimens," *ARPHA Prepr.*, vol. 4, p. e107168, May 2023, doi: 10.3897/arphapreprints.e107168.
- [7] "FAIR Digital Objects and Machine-Actionability," DiSSCoTech. Accessed: Mar. 01, 2024. [Online]. Available: <https://dissco.tech/2022/07/21/fair-digital-objects-and-machine-actionability/>
- [8] S. Leeflang *et al.*, "DiSSCo Prepare D6.2 Implementation and construction plan of the DiSSCo core architecture," Jun. 2022, doi: 10.5281/zenodo.6832200.
- [9] Corporation for National Research Initiative, *HANDLE.NET (Ver. 9) Technical Manual*. 2018. [Online]. hdl: 20.1000/113
- [10] A. Ivonne *et al.*, "FDO Forum FDO Requirement Specifications," Jan. 2023, doi: 10.5281/zenodo.7782262.
- [11] "What Are Registration Agencies?" Accessed: Dec. 13, 2023. [Online]. Available: <https://www.doi.org/the-community/what-are-registration-agencies>