From data pipelines to FAIR data infrastructures: A vision for the new horizons of bio- and geodiversity data for scientific research

Sharif Islam[‡,§], Claus Weiland[|], Wouter Addink[‡,§]

‡ Naturalis Biodiversity Center, Leiden, Netherlands
§ Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
| Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main, Germany

Corresponding author: Sharif Islam (sharif.islam@naturalis.nl)

Abstract

Natural science collections are vast repositories of bio- and geodiversity specimens. These collections, originating from natural history cabinets or expeditions, are increasingly becoming unparalleled sources of data facilitating multidisciplinary research (Meineke et al. 2018, Heberling et al. 2019, Cook et al. 2020, Thompson et al. 2021). Due to various global data mobilization and digitisation efforts (Blagoderov et al. 2012,Nelson and Ellis 2018), this digitised information about specimens includes database records along with two/three-dimensional images, sonograms, sound or video recordings, computerised tomography scans, machine-readable texts from labels on the specimens as well as media items and notes related to the discovery sites and acquisition (Hedrick et al. 2020, Phillipson 2022).

The scope and practice of specimen gathering are also evolving. The term *extended specimen* was coined to refer to the specimen and associated data extending beyond the singular physical object to other physical or digital entities such as chemical composition, genetic sequence data or species data. Thus the specimen becomes an interconnected network of data resources that have incredible potential to enhance integrative and data-driven research (Webster 2017,Lendemer et al. 2019,Hardisty et al. 2022). These practices also reflect the role of data and the curatorial data life-cycle starting from the initial material sampling process to the downstream analysis. We are also seeing growing acknowledgement that disparate and domain specific data elements prevent interdisciplinarity which is crucial for a holistic understanding of biodiversity and climate crisis (Hicks et al. 2010, Craven et al. 2019, Folk and Siniscalchi 2021).

Thus the data elements are not just records or rows in a database or data pipelines going from one repository to another. They have the potential to become self-describing digital artefacts that can revolutionise how machines interpret and work with specimen data. Within this context, the Distributed System of Scientific Collections (DiSSCo), a new

European Research Infrastructure for natural science collections, envisions an infrastructure based on FAIR Digital Objects (FDO) that can unify more than 170 European natural science collections under common and FAIR-compliant (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al. 2016) access and curation policies and practices. DiSSCo's key element in achieving FAIR is the implementation of Digital Specimen (a domain specific FDO) that closely aligns with the extended specimen practices. The idea behind Digital Specimen – an FDO that acts as a digital surrogate for a specific physical specimen in a natural science collection – was influenced by global conversations around the implementation of the Digital Object Architecture for biodiversity data (De Smedt et al. 2020, Islam et al. 2020,Hardisty et al. 2020).

The main purpose of this talk is to explain the vision of how FAIR and FDO can create a data infrastructure that can not only take advantage of existing databases and repositories but at the same time provide support for innovative services such as AI and digital twinning. With scientific use cases in mind, the talk will highlight a few key FAIR and FDO components (persistent identifiers, metadata, ontologies) within the collaborative modelling activity of Digital Specimen specification. These components provide the template for specifying how a Digital Specimen should look so DiSSCo can build a FAIR service ecosystem based on FDOs (Addink et al. 2021). We will also give examples of envisioned services that can help with image feature extraction, and model training (Grieb et al. 2021, Hardisty et al. 2022) and digital twinning (Schultes et al. 2022). We believe this is an exciting new paradigm powered by FAIR and FDO that can help both humans and machines to accelerate the use of specimen data. From physical objects curated over hundred years, we have developed data pipelines, aggregators and repositories (Barberousse 2021). Now is the time to look for solutions where these data records can become FAIR Digital Objects to enable wider access and multidisciplinary research.

Keywords

FAIR data infrastuctures, biodiveristy data, interdisciplinarity, digital specimen, digital twinning, FAIR Digital Objects, DiSSCo

Presenting author

Sharif Islam

Presented at

First International Conference on FAIR Digital Objects, presentation

Funding program

H2020-INFRADEV-2019-2020 – Grant Agreement No. 871043

Grant title

DiSSCo Prepare

Conflicts of interest

References

- Addink W, Islam S, Alonso J (2021) DiSSCo e-Services to Serve Global Community Needs. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.73903
- Barberousse A (2021) Biodiversity databanks and scientific exploration. Lato Sensu: Revue de la Société de philosophie des sciences 8 (2): 32-43. https://doi.org/10.20416/lsrsps.v8i2.4
- Blagoderov V, Kitching I, Livermore L, Simonsen T, Smith V (2012) No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209: 133-146. https://doi.org/10.3897/zookeys.209.3178
- Cook JA, Arai S, Armién B, Bates J, Bonilla CAC, Cortez MBdS, Dunnum JL, Ferguson AW, Johnson KM, Khan FAA, Paul DL, Reeder DM, Revelez MA, Simmons NB, Thiers BM, Thompson CW, Upham NS, Vanhove MPM, Webala PW, Weksler M, Yanagihara R, Soltis PS (2020) Integrating Biodiversity Infrastructure into Pathogen Discovery and Mitigation of Emerging Infectious Diseases. BioScience 70 (7): 531-534. https://doi.org/10.1093/biosci/biaa064
- Craven D, Winter M, Hotzel K, Gaikwad J, Eisenhauer N, Hohmuth M, König-Ries B, Wirth C (2019) Evolution of interdisciplinarity in biodiversity science. Ecology and Evolution 9 (12): 6744-6755. https://doi.org/10.1002/ece3.5244
- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2). https://doi.org/10.3390/publications8020021
- Folk R, Siniscalchi C (2021) Biodiversity at the global scale: the synthesis continues. American Journal of Botany 108 (6): 912-924. https://doi.org/10.1002/ajb2.1694
- Grieb J, Weiland C, Hardisty A, Addink W, Islam S, Younis S, Schmidt M (2021) Machine Learning as a Service for DiSSCo's Digital Specimen Architecture. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.75634
- Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul D, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e54280
- Hardisty A, Brack P, Goble C, Livermore L, Scott B, Groom Q, Owen S, Soiland-Reyes S (2022) The Specimen Data Refinery: A Canonical Workflow Framework and FAIRDigital Object Approach to Speeding up Digital Mobilisation of Natural

HistoryCollections. Data Intelligence 4 (2): 320-341. https://doi.org/10.1162/dint_a_00134

- Hardisty AR, Ellwood ER, Nelson G, Zimkus B, Buschbom J, Addink W, Rabeler RK, Bates J, Bentley A, Fortes JAB, Hansen S, Macklin JA, Mast AR, Miller JT, Monfils AK, Paul DL, Wallis E, Webster M (2022) Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. BioScience https://doi.org/10.1093/biosci/biac060
- Heberling JM, Prather LA, Tonsor SJ (2019) The Changing Uses of Herbarium Data in an Era of Global Change: An Overview Using Automated Content Analysis. BioScience 69 (10): 812-822. https://doi.org/10.1093/biosci/biz094
- Hedrick BP, Heberling JM, Meineke EK, Turner KG, Grassa CJ, Park DS, Kennedy J, Clarke JA, Cook JA, Blackburn DC, Edwards SV, Davis CC (2020) Digitization and the Future of Natural History Collections. BioScience 70 (3): 243-251. https://doi.org/10.1093/biosci/biz163
- Hicks C, FITZSIMMONS C, POLUNIN NC (2010) Interdisciplinarity in the environmental sciences: barriers and frontiers. Environmental Conservation 37 (4): 464-477. https://doi.org/10.1017/s0376892910000822
- Islam S, Hardisty A, Addink W, Weiland C, Glöckler F (2020) Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. Data Science Journal 19 https://doi.org/10.5334/dsj-2020-050
- Lendemer J, Thiers B, Monfils AK, Zaspel J, Ellwood ER, Bentley A, LeVan K, Bates J, Jennings D, Contreras D, Lagomarsino L, Mabee P, Ford LS, Guralnick R, Gropp RE, Revelez M, Cobb N, Seltmann K, Aime MC (2019) The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. BioScience 70 (1): 23-30. https://doi.org/10.1093/biosci/biz140
- Meineke E, Davies TJ, Daru B, Davis C (2018) Biological collections for understanding biodiversity in the Anthropocene. Philosophical Transactions of the Royal Society B: Biological Sciences 374 (1763). https://doi.org/10.1098/rstb.2017.0386
- Nelson G, Ellis S (2018) The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B: Biological Sciences 374 (1763). https://doi.org/10.1098/rstb.2017.0391
- Phillipson T (2022) Collections development in hindsight: a numerical analysis of the Science and Technology collections of National Museums Scotland since 1855. Science Museum Group Journal 12 (12). https://doi.org/10.15180/191205
- Schultes E, Roos M, Bonino da Silva Santos LO, Guizzardi G, Bouwman J, Hankemeier T, Baak A, Mons B (2022) FAIR Digital Twins for Data-Intensive Research. Frontiers in big data 5: 883341. https://doi.org/10.3389/fdata.2022.883341
- Thompson C, Phelps K, Allard M, Cook J, Dunnum J, Ferguson A, Gelang M, Khan FAA, Paul D, Reeder D, Simmons N, Vanhove MM, Webala P, Weksler M, Kilpatrick CW (2021) Preserve a Voucher Specimen! The Critical Need for Integrating Natural History Collections in Infectious Disease Studies. mBio 12 (1). https://doi.org/10.1128/mbio.02698-20
- Webster MS (Ed.) (2017) The extended specimen: emerging frontiers in collections-based ornithological research. CRC Press
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth

P, Goble C, Grethe J, Heringa J, ’t Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18