Community Forum Topic: Converging Digital Specimens and Extended Specimens - Towards a global specification for data integration

Short Link to this document: http://bit.ly/esdsconsult
Contact email: alliance@gbif.org

Background and context

During 2020 there have been growing conversations about the exciting possibilities of digital representations on the Internet of the billions of specimens currently held in the world's natural science collections. In Europe, DiSSCo's Digital Specimen concept and in the USA, BCoN's Extended Specimen concept are aligned in a general vision of connecting and aligning all information related to a specimen. GBIF currently brings together potentially related biodiversity records by matching similar entries in individual fields across different datasets mediated at GBIF. This new clustering algorithm gives us a taste of the possibilities of fully integrated biodiversity data in the future.

These ideas have spurred discussions, notably through the TDWG 2020 conference about how our global community can work together to build infrastructure to achieve a fully integrated digital data infrastructure. Ideas in different discussions vary and it is beneficial when we can work together towards a single, robust global solution.

During TDWG 2020, a growing group of organizations and individuals signed a Letter of Intent to work collaboratively towards a global specification and interoperability for the digital/extended specimen. These signees now propose the present community consultation under the umbrella of the Alliance for Biodiversity Knowledge, using a mix of virtual meetings and online discussions via GBIF's Community Forum.

Problem statement

A growing consensus of the collections-based natural sciences community views further and deeper digital data integration as essential to making data more relevant, more easily findable, accessible, interoperable and reusable (FAIR) for the research and policy work needed to address global issues. There are complimentary but differing visions of a richer data model for the Natural Science Collections community to accomplish this goal. The community must discuss several critical areas with technical, financial, social, governance and professional implications and understand these together. This is necessary to achieve collective consensus on the vision, model and way forward from where we are today. It is important that the entire global community is engaged through consultation and has access to the latest technical and social developments and thinking.

This consultation's goals are to: i) expand participation in the process; ii) build support for further collaboration; iii) identify important driving use cases; iv) identify significant challenges and obstacles; and v) develop a comprehensive roadmap towards achieving the vision.

Discourse topics with moderators

Topics 1 - 5 below will be discussed in parallel during a first phase, followed by a later phase discussing topics 6 and 7. Details about the individual topics are provided in the sections about value streams and capabilities further below. The names of the persons moderating each topic discussion appear in brackets.

First phase:

1. Making FAIR data for specimens accessible (Alex Hardisty, Barbara Thiers, Wouter Addink. Lawrence Monda)
2. Extending, enriching and integrating data (Andy Bentley, Jen Zaspel, Mike Webster, Keping Ma)
3. Annotating specimens and other data (Joe Miller, Gil Nelson, James Macklin, Rich Rabeler)
4. Attributing work done (Data Attribution) (David Shorthouse, Nicky Nicolson, Andy Bentley)
5. Analysing/mining specimen data for novel applications (Andrew Young, Libby Ellwood, Anna Monfils, John Bates)

Later phase:

6. Well-founded access points and data cyberinfrastructure alignment (Jose Fortes, Tim Robertson, Sharif Islam, ALA rep.)
7. Persistent identifier (PID) scheme(s) (Alex Hardisty, Wouter Addink)

If you would like to co-moderate a thread please contact us at: alliance@gbif.org

Timeline

- Before New Year's Eve 2020, 'save the date' announcement.
- January 2021
    - Finalise and release this scoping document with renewed announcement
    - Convenors meet to finalize materials and organize virtual opening session(s)
    - GBIF sets up Discourse templates
    - Moderators develop the background materials and discussion seeds
    - Preparation of the introductory sessions
- Jan 29th - *moderator meeting*
- **9th February - Open view of the consultation threads**
- 12 February - *moderator meeting*
- **16th Feb Hold virtual introductory sessions** (x 2 for different world regions)
        - 6:00 UTC and 15:00 UTC
        - Consultation open for editing
- 19th & 26th February - *moderator meetings*
- **5 March - Officially close session**
    - Close consultations

- Mid March - convenors deliver summaries
- End March - Synthesized results available for roadmap development
- April (after Easter, date to be determined)
    - Initiate consultation threads for topics 6 - 7 on Discourse
    - Hold virtual introductory sessions (x 2 for different world regions)
- May (4 weeks later)
    - Hold virtual closing sessions
    - Close consultations
- End May - convenors deliver summaries
- Mid-June - Further synthesized results available for roadmap development

Value streams as discourse topics

The digital specimen / extended specimen framework enables multiple value streams that are the first five topics of the consultation.

A value stream is a sequence of activities that creates an overall result or outcome for a stakeholder (end-user). A stakeholder can be a scientist, a collection manager, a curator, etc. The result or outcome has a worth or usefulness to the stakeholder.

1. **Making FAIR data for specimens accessible** – Everyone wants widened/improved access to specimen data through digitization. Everyone wants data to be findable, accessible, interoperable and reusable. Standardized open Digital Specimens are FAIR by design.
2. **Extending, enriching and integrating data** – People want to be able to add references to other (third-party) data that relates to or is derived from specimens, e.g. genetic sequence, isotope, CT scan, or publication/citation. They want to be able to find such data when using a specimen (or vice-versa).
3. **Annotating specimens*** – Common uses of annotations are to bring the scientific names of specimens up-to-date to conform with current classification concepts, to dispute the identification of a specimen and/or to make comments and correct locality, georeference or other specimen information. Scientists and curators want to annotate specimens with the latest opinions and determinations, and they want to see what has been annotated. Collection managers want to be able to review and optionally accept such annotations back into their systems as updates to information already there.
4. **Attributing work done** – People want their efforts to be acknowledged and recognised. Other people want to know who did the work and when, and for that information to be unambiguous. Collections want to gain attribution for their contribution to scientific endeavor though specimen and data use in end-products of research. Standardized mechanisms and metrics facilitate this.
5. **Analysing/mining specimen data for novel applications** – People want to be able to exploit specimen and derived data for analysis, to use it for specific tasks and to find patterns in it. They desire '*higher level functionalities of e.g., a spatial portal*' than just being able to discover, inspect and retrieve. They want to re-use data and they want data to be interoperable.

* Feeding back to source is a subcomponent of annotating specimens. It could be separated out into a distinct value stream as it requires different capabilities than those for annotating specimens. However, both annotating and feeding back are reliant on a common annotatable/annotated Digital Specimen resource and its provenance (history).

Infrastructure capabilities

Seen from a provider's perspective, a capability is an ability and capacity to fulfill or achieve a specific business outcome. Capabilities generally map to and support stages in value streams but new capabilities (most likely towards different value streams) can also be generated by value streams. Some initial capabilities are:

Community curation is a capability to work with digitized specimens that map to several value stream concepts (above) that include annotating specimens, attributing work, extending and enriching the data. Curation begins after digitization has been done, when the Digital Specimen data has been shared and made publicly available.

Citing/referring to work done is a capability made possible by attributing and keeping track of work done in an unambiguous way. It includes being able to manage different versions as the data of Digital Specimens evolves.

6. **Well-founded access points and data cyberinfrastructure alignment** – Being able to access digital specimen data globally through a single or small number of well-founded access points (a capability) is valuable insofar as user efficiency and confidence are improved. Confidence and trust in the data can be higher when it is known that the activities behind this capability aim at maintaining the integrity of persistent identifier (PID) links, keeping metadata up to date (through, for example continuous two-way synchronisation) and encouraging/enforcing standardised data representations of open Digital Specimens. The present day context and present data cyber infrastructure is the starting point. We must consider what is needed to align the multiple elements we have today towards a single, robust global solution, and what is involved in the journey to get there. Existing standards and capabilities cannot just be thrown away. Professional practices don't change overnight. What is implied in terms of the new standards and professional practices needed? How do we do this?

7. **Persistent identifier (PID) scheme(s)** – PIDs are foundational elements of data cyber infrastructure, not only as identifiers but also as connectors of one thing to another. The ability of machines to process digital/extended specimen data depends on robust, reliable PIDs. PIDs act in several layers. What is used at the level of the institution can and often will be different from what is used in aggregation, federation and integration. The challenge is not the choice of scheme(s) for digital/extended specimens, which is quite straightforward but is in translation to actionable steps the community can align to and support through smooth, non-disruptive transitions. Multiple stakeholders must frame an agreement that includes technical, ownership, authority, governance and financial elements.

Moderation

Duties of the moderator:

- Attend weekly coordinating session with all moderators during consultation
- Seed the discussion with an introduction
- Perform periodic summaries
- Coordinate across threads with other moderators e.g., weekly
- Provide reference materials
- Stimulate and lead discussion throughout consultations
- Moderate virtual sessions
- Summarize threads at the end of consultation

Virtual Kickoff session
    Coordinate the five threads

Virtual sessions (flexible)
    Summarize consultation

Reference materials for orientation

The following items 1 - 4 are recommended introductory reading/viewing for orientation going into the consultation:

1. Webster, MS. (ed.) 2017. The extended specimen: emerging frontiers in collections-based ornithological research. CRC Press. doi: 10.1201/9781315120454-1.
   Especially Chapter 1: https://www.taylorfrancis.com/chapters/extended-specimen-1-michael-webster/e/10.1201/9781315120454-1 *explaining the idea of extending specimens from the scientific perspective*.

2. What is a Digital Specimen? https://bit.ly/DigitalSpecimen. *A short blog post in DiSSCo Tech.*

3. TDWG 2020. BOF01 Birds of a Feather session on Converging Digital Specimens and Extended Specimens - Towards a global specification. 22 September 2020. Recording of the session: https://youtu.be/8ljokNRkjeo. Notes document: http://bit.ly/bof01notes.
   Especially the two presentations by Alex Hardisty and Andrew Bentley respectively on digital specimens and extended specimens, minutes 4 - 44 in the video.

4. Hardisty, A. 2021. ES/DS Framework - a technical explanation towards convergence. Video recording: *<to be added, early January 2021>*.
   A recently recorded talk highlighting the commonality of the digital specimen and extended specimen concepts, and explaining a technical framework for compatible implementation.

The items 5 - 9 cover next generation collections and the extended specimen network strategy:

5. NASEM 2020. Biological Collections: Ensuring Critical Research and Education for the 21st Century.

6. National Academies of Sciences, Engineering, and Medicine. 2020. Biological Collections: Ensuring Critical Research and Education for the 21st Century. Washington, DC: The National Academies Press. doi: 10.17226/25592.

7. Lendemer J, Thiers B, Monfils AK, Zaspel J, Ellwood ER, Bentley A, LeVan K, Bates J, Jennings D, Contreras D, Lagomarsino L. 2020. The extended specimen network: A strategy to enhance US biodiversity collections, promote research and education. BioScience:70(1):23-30. doi: 10.1093/biosci/biz140.

8. BCoN 2019. Extending U.S. Biodiversity Collections to Promote Research and Education. A report by the Biodiversity Collections Network (2019). URL: https://www.aibs.org/home/assets/BCoN_March2019_FINAL.pdf.

9. Schindel, D.E. and Cook, J.A., 2018. The next generation of natural history collections. PLoS Biology, 16(7), p.e2006
10. 125. doi: 10.1371/journal.pbio.2006125.

The items 10 - 12 cover work on the development of a specification for open Digital Specimens (openDS):

11. openDS github repository: https://github.com/DiSSCo/openDS. *In particular:*
    - *Positioning openDS in the landscape* (https://github.com/DiSSCo/openDS/blob/master/positioning-opends.md)
    - *Introduction to the data model* (https://github.com/DiSSCo/openDS/blob/master/data-model/data-model-intro.md)
    - *Introduction to the ontology* (https://github.com/DiSSCo/openDS/blob/master/ods-ontology/ods-ont-intro.md)

12. Addink W, Hardisty AR. 2020. 'openDS' – Progress on the New Standard for Digital Specimens. Biodiversity Information Science and Standards 4: e59338. doi: 10.3897/biss.4.59338. Recording of talk at TDWG 2020 SYM07 session, 20 October 2020: https://youtu.be/vBP0eYtcB5s. *On the current status of openDS development.*

13. Hardisty A, Ma K, Nelson G, Fortes J. 2019. ‘openDS’ – A New Standard for Digital Specimens and Other Natural Science Digital Object Types. Biodiversity Information Science and Standards 3: e37033. doi: 10.3897/biss.3.37033. *The original proposal.*

The items 13 - 18 cover work on using RDA outputs and FAIR Digital Objects in the design of the DiSSCo research infrastructure, and the FAIR Guiding Principles:

14. 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(50) pp.1-14. doi: 10.5334/dsj-2020-050.

15. Thessen, A.E., Woodburn, M., Koureas, D., Paul, D., Conlon, M., Shorthouse, D.P. and Ramdeen, S., 2019. Proper attribution for curation and maintenance of research collections: Metadata recommendations of the RDA/TDWG working group. Data Science Journal, 18(1): 54. doi: 10.5334/dsj-2019-054.

16. De Smedt K, Koureas D, Wittenburg P. 2020 FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8(2):21. doi: 10.3390/publications8020021.

17. Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul DL, 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: e54280. doi: 10.3897/rio.6.e54280.

18. Lannom, L., Koureas, D. and Hardisty, A. 2020. FAIR data and services in biodiversity science and geoscience. Data Intelligence 2(1-2), pp. 122-130. doi: 10.1162/dint_a_00034.

19. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3. doi: 10.1038/sdata.2016.18.

Miscellaneous items

20. TDWG 2020. PD03 Panel Discussion session on Enabling digital specimen and extended specimen concepts in current tools and services. 23 October 2020. Recording of the session: https://www.youtube.com/watch?v=i9IDCJp_aA4. Notes document: https://tinyurl.com/TDWG2020-PD03.

END.