The Specimen Data Refinery: Using a scientific workflow approach for information extraction

Laurence Livermore[‡], Paul Brack[§], Ben Scott[‡], Stian Soiland-Reyes[§], Oliver Woolland[§]

‡ The Natural History Museum, London, United Kingdom
§ The University of Manchester, Manchester, United Kingdom

Corresponding author: Laurence Livermore (l.livermore@nhm.ac.uk)

Abstract

Over the past three years, we have been developing the Specimen Data Refinery (SDR) to automate the extraction of data from specimen images as part of the SYNTHESYS project (Walton et al. 2020). The SDR provides an easy to deploy, open source, web-based interface to multiple workflows that enable a user to create new or enhance existing natural history specimen records. The SDR uses the Galaxy workflow platform as the basis for managing data analysis, and where possible, using existing Galaxy community tools and approaches (Jalili et al. 2020, Hardisty et al. 2022). We have developed a library of domain-specific tools including semantic segmentation, optical character recognition, hand-written text recognition, barcode reading and natural language processing. These tools have been designed to work on standardised images of specimens, specifically herbarium sheets, pinned insects and microscope slides.

In this presentation, we provide our technical approach in developing the SDR, including the Galaxy workflow platform, application deployment, and tool interoperability, using FAIR digital objects (e.g., RO-Crates and openDigital Specimen objects (Soiland-Reyes et al. 2022, Addink and Hardisty 2020)). We present an evaluation of the tools, including segmentation, text recognition, and others, and the new challenges in using the resulting data from both a technical and social perspective.

Keywords

Galaxy workflow platform, automation, natural history specimens, digitisation

Presenting author

Laurence Livermore

Presented at

TDWG 2022

Funding program

H2020-EU.1.4.1.2. - Integrating and opening existing national and regional research infrastructures of European interest

Grant title

SYNTHESYS PLUS – "Synthesis of systematic resources", Grant Agreement No. 823827

Author contributions

Author contributions to this article according to the Contributor Roles Taxonomy CASRAI CrEDiT:

- **Laurence Livermore**: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Resources, Writing – review & editing.
- **Paul Brack**: Conceptualization, Software.
- **Ben Scott**: Data curation, Software, Validation.
- **Stian Soiland-Reyes**: Investigation, Methodology, Supervision, Writing – review & editing.
- **Oliver Woolland**: Data curation, Resources, Software, Visualization, Writing – review & editing.

Conflicts of interest

References

- Addink W, Hardisty A (2020) 'openDS' – Progress on the New Standard for Digital Specimens. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59338
- 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 FAIR Digital Object Approach to Speeding up Digital Mobilisation of Natural History Collections. Data Intelligence 4 (2): 320-341. https://doi.org/10.1162/dint_a_00134

- Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, Nekrutenko A (2020) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Research 48 https://doi.org/10.1093/nar/gkaa434
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. Data Science1-42. https://doi.org/10.3233/ds-210053
- Walton S, Livermore L, Bánki O, Cubey R, Drinkwater R, Englund M, Goble C, Groom Q, Kermorvant C, Rey I, Santos C, Scott B, Williams A, Wu Z (2020) Landscape Analysis for the Specimen Data Refinery. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e57602