RESEARCH PAPER

The Specimen Data Refinery: A Canonical Workflow Framework and FAIR Digital Object Approach to Speeding up Digital Mobilisation of Natural History Collections

Alex Hardisty[1†], Paul Brack[2], Carole Goble[2], Laurence Livermore[3], Ben Scott[3], Quentin Groom[4], Stuart Owen[2] & Stian Soiland-Reyes[2,5]

[1]School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, UK

[2]The Department of Computer Science, The University of Manchester, Manchester M13 9PL, UK

[3]The Natural History Museum, London SW7 5BD, UK

[4]Meise Botanic Garden, 1860 Meise, Belgium

[5]Informatics Institute, Faculty of Science, University of Amsterdam, 1090 GH Amsterdam, The Netherlands

Keywords: Digital Specimen; Workflow; FAIR; Digital Object; RO-Crate

Citation: Hardisty, A., et al.: The specimen data refinery: A canonical workflow framework and FAIR digital object approach to speeding up digital mobilisation of natural science collections. Data Intelligence 4(2), 320-341 (2022). doi: 10.1162/dint_a_00134
Received: September 12, 2021; Revised: November 22, 2021; Accepted: February 4, 2022

ABSTRACT

A key limiting factor in organising and using information from physical specimens curated in natural science collections is making that information computable, with institutional digitization tending to focus more on imaging the specimens themselves than on efficiently capturing computable data about them. Label data are traditionally manually transcribed today with high cost and low throughput, rendering such a task constrained for many collection-holding institutions at current funding levels. We show how computer vision, optical character recognition, handwriting recognition, named entity recognition and language translation technologies can be implemented into canonical workflow component libraries with findable, accessible, interoperable, and reusable (FAIR) characteristics. These libraries are being developed in a cloud-based workflow platform—the 'Specimen Data Refinery' (SDR)—founded on Galaxy workflow engine, Common Workflow Language, Research Object Crates (RO-Crate) and WorkflowHub technologies. The SDR can be applied to specimens' labels and other artefacts, offering the prospect of greatly accelerated and more accurate data capture in computable form. Two kinds of FAIR Digital Objects (FDO) are created by packaging

† Corresponding author: Alex Hardisty (Email: hardistyar@gmail.com; ORCID: 0000-0002-0767-4310).

outputs of SDR workflows and workflow components as digital objects with metadata, a persistent identifier, and a specific type definition. The first kind of FDO are computable Digital Specimen (DS) objects that can be consumed/produced by workflows, and other applications. A single DS is the input data structure submitted to a workflow that is modified by each workflow component in turn to produce a refined DS at the end. The Specimen Data Refinery provides a library of such components that can be used individually, or in series. To cofunction, each library component describes the fields it requires from the DS and the fields it will in turn populate or enrich. The second kind of FDO, RO-Crates gather and archive the diverse set of digital and real-world resources, configurations, and actions (the provenance) contributing to a unit of research work, allowing that work to be faithfully recorded and reproduced. Here we describe the Specimen Data Refinery with its motivating requirements, focusing on what is essential in the creation of canonical workflow component libraries and its conformance with the requirements of an emerging FDO Core Specification being developed by the FDO Forum.

1. INTRODUCTION

A key limiting factor in organising and using information from physical specimens curated in natural history collections is making that information computable ('machine-actionable') and extendable. More than 85% of available information currently resides on labels attached to specimens or in physical ledgers [1]. Label data are commonly transcribed manually with high cost and low throughput, rendering such a task constraining for many institutions at current funding levels. However, the advent of rapid, high-quality digital imaging has meant that digitizing specimens, including their labels, is now faster and cheaper [2]. With initiatives such as Advancing Digitization of Biological Collections (ADBC), integrated Digitized Biocollections (iDigBio) and the Distributed System of Scientific Collections (DiSSCo) [3, 4, 5, 6] aiming to increase the rate and accuracy of both mass and on-demand digitization of natural history collections, the gap between expectations of what should be digitally available and computable, and what can be achieved using traditional transcription approaches is widening. Modern, highly efficient workflow tools and approaches can play a role to address this.

Collection digitization began towards the end of the 20th century by typing basic data from labels into the collection (asset) management systems of collection-holding institutions such as natural history museums, herbaria and universities. Initially, this was to facilitate indexing and cataloguing and locating the physical specimens, but with the addition of photographic images of specimens and the public availability of specimen data records, through data portals of the institutions themselves as well as international data infrastructures like the Global Biodiversity Information Facility (GBIF), such bodies of data have been rapidly exploited for research [7, 8]. It has become clear that widespread digitization of data about physical specimens in collections and the advent of high-throughput digitization processes [9, 10, 11, 12, 13] is transforming and will radically further transform the range of scientific research opportunities and questions that can be addressed [14, 15]. Scientific conclusions and policy decisions evidenced by digital specimen data enhance humankind's ability to conserve, protect, and predict the biodiversity of our world [16, 17].

Harnessing technologies developed to harvest, organise, analyse and enhance information from sources such as scholarly literature, third-party databases, data aggregators, data linkage services and geocoders and reapplying these approaches to specimens' labels and other artefacts offers the prospect of greatly accelerated data capture in a computable form [18]. Tools of particular interest span the fields of computer vision, optical character recognition, handwriting recognition, named entity recognition and language translation.

Workflow technologies from the ELIXIR Research Infrastructure [19], including Galaxy [20], Common Workflow Language [21], Research Object Crates (RO-Crates) [22, 23] and WorkflowHub [24], and selected tools are integrated in a cloud-based workflow platform for natural history specimens—the 'Specimen Data Refinery' [1] that will become one of the main services to be offered by the planned DiSSCo research infrastructure [5]. The tools themselves, implemented with findable, accessible, interoperable, and reusable (FAIR) characteristics [25] are packaged into canonical workflow component libraries [26], rendering them reusable, and interoperable with one another. FAIR Digital Objects are adopted as the common input/output pattern, fully compatible with digital objects at the core of DiSSCo data management [27].

The Refinery brings together domain-specific workflows for processing specimen images and extracting text and data from images with canonical forms for components and interactions between components that can lead to improved FAIR characteristics for both the workflows themselves and the data resulting from workflow execution.

FAIR Digital Objects (FDO) are created by packaging outputs of workflows and workflow components as digital objects with metadata, a persistent identifier, and a specific type definition against which operations can be executed [28]. The Refinery uses two kinds of FDOs:

- **Computable Digital Specimen (DS) objects** [29] from DISSCo for the scientific input/output data that can be consumed/produced by workflows and other applications.
- **Workflow objects, implemented as RO-Crates** [23], from ELIXIR gather and archive the diverse set of workflow process data—the digital and real-world resources, configurations and actions (the provenance) contributing to a unit of digitization or other work producing the Digital Specimen digital objects, allowing that work to be scrutinised and faithfully reproduced if necessary.

We first summarise related work before describing the problem to be addressed by the Specimen Data Refinery. We then explain our Canonical Workflows for Research (CWFR) approach using these FDOs in the design of the SDR, the experimental setup, and results so far from the work in progress. While future work will clarify full results and challenges of implementing a robust, reliable, and easy-to-use production-capability SDR, in this early report following SDR prototyping and conceptualization, we focus on what we found to be essential in the use of FDOs and CWFR canonical step libraries, and on the compliance of canonical workflow (component) inputs and outputs with the requirements of the FDO Framework [30].

2. RELATED WORK

2.1 Workflows for Processing Specimen Images and Extracting Data

While natural history collections are heterogeneous in size and shape, often they are mass digitized using standardised workflows [9, 10, 11, 12, 13]. In pursuit of higher throughput at lower cost, yet with higher accuracy and richer metadata, further automation will increasingly rely on techniques of object detection and segmentation, optical character recognition (OCR) and semantic processing of labels, and automated taxonomic identification and visual feature analysis [1, 18].

Although there is a great deal of variety among images of different kinds of collection objects that are digitized (see figure 1) there are visual similarities between them. Most images contain labels, scale bars and often, colour charts as well as the specimen itself. This makes them amenable to improved approaches to object detection [31] and segmentation into 'regions of interest' [32] as precursive steps for multiple kinds of workflows.

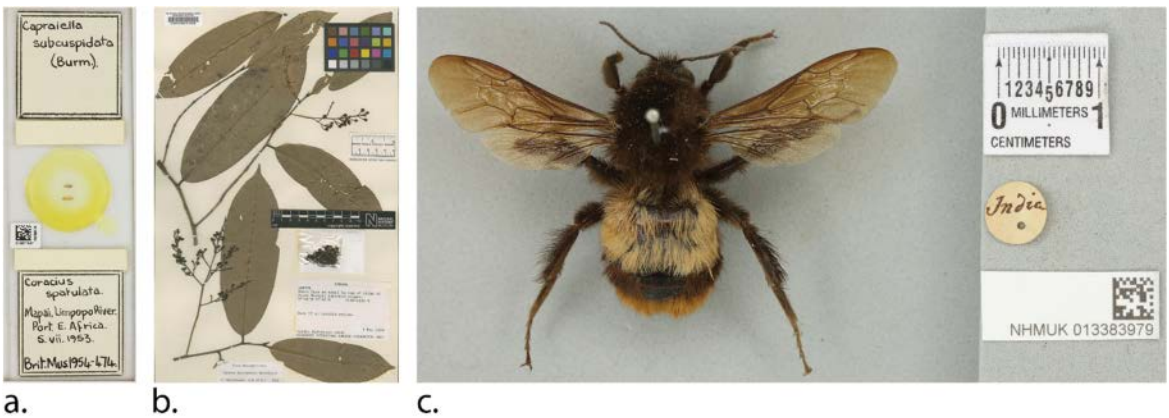


Figure 1. A range of specimen images from the Natural History Museum, London, demonstrating the diversity of collection objects, which include handwritten, typed, and printed labels. (a) Microscope slide (NHMUK010671647[1]), (b) Herbarium specimen (BM000546829[2]), and (c) pinned insect (NHMUK013383979[3]). CC-BY 4.0 © The Trustees of the Natural History Museum, London.

Segmentation, specifically, can be employed as an early step in a workflow to send just the relevant region(s) of interest from an image to later workflow steps. Not only does this decrease data transfer time and minimise computational overheads but it can also substantially increase the accuracy of subsequent OCR processing and semantic recognition steps [18].

[1] https://data.nhm.ac.uk/dataset/collection-specimens/resource/05ff2255-c38a-40c9-b657-4ccb55ab2feb/record/8005500/1646092800000

[2] https://data.nhm.ac.uk/object/be595f07-73c5-4764-a96c-8b377e3d1507/1646092800000

[3] https://data.nhm.ac.uk/object/745febc7-8222-498a-9969-5f6b12f85ef3/1646092800000

Much of the data about specimens is stored on their handwritten, typed or printed labels or in registers/ledgers [33]. Direct manual transcription into local databases with manual georeferencing is the primary method used today to capture this data. Potentially, OCR can significantly increase transcription speeds whilst reducing cost; although it sacrifices accuracy and disambiguation that are today achieved with specialist knowledge provided by humans during the process. Returning character strings from OCR is useful, but semantically placing this data in its context as information specific to natural history specimens and linking that back to the original physical specimens is of much higher value, improving the utility of natural history collections. Shortfalls in accuracy and disambiguation can be made up by exploiting Natural Language Processing (NLP) advances such as named entity recognition to identify text segments belonging to predefined categories (for example, species name, collector, locality, date) [18]. Nevertheless, this only works well on a small proportion of captured data in the absence of 'human in the loop' input. To better automate disambiguation of people's names, for example, access to other contextual 'helper' data are needed (e.g., biographical data in Wikidata) as well as cross-comparison with other data from the specimen, such as the date of collection and location [34].

Automated identification of species from images of living organisms has achieved impressive levels of accuracy [35, 36, 37, 38, 39, 40] with techniques translated to an increasing range of enthusiastically received consumer applications for plant and animal identification using mobile phones (e.g., Plantsnap[④], PictureThis[⑤], iNaturalist SEEK[⑥]). Automated identification of *preserved specimens*, however, presents different challenges. Although identification might be made more accurately because a specimen is presented in a standard manner, separated from other organisms and the complexity of a natural background, the loss of colour and distortion of the shape of the organism arising from preparation and preservation processes can lead to the loss of important identification clues that might be present on a living example.

2.2 Workflow Management Systems and Canonical Workflows for Research

A workflow chains together atomised and executable components with the relationships between them to clearly define a control flow and a data flow. Their significant defining characteristics are (i) abstraction, through the separation of the workflow specification (the work to be done) from its execution (how it is done), and (ii) composition whereby the components can be cleanly combined and reused and workflows themselves can be neatly packaged as components [41]. Workflow management systems typically provide the necessary mechanisms for explicitly defining workflows in a reusable way together with a workflow engine that executes the workflow steps and keeps an accountable record of the processing—logging the codes executed and the data lineage of the results. In the past decade there has been a rise in popularity in both the development of WfMS and their use, driven by the increasing scales of data and the accompanying complexity of its processing [41].

① https://www.plantsnap.com/
② https://www.picturethisai.com/
③ https://www.inaturalist.org/pages/seek_app

Workflow management systems typically vary in the features they provide for supporting: workflow programming language and control flow expressivity; data type management; code wrapping, containerisation and integration with software management tools; exploitation of computational architectures; availability of development and logging tools; licensing and so on. Although several hundred kinds of such systems exist [42], communities tend to cluster around a few popular systems based on their "plugged-in" availability of data type specialist codes, the catered-for skills level of the workflow developers, and its documentation, community support and perceived sustainability. For the Specimen Data Refinery, the Galaxy workflow system [20] in conjunction with Common Workflow Language (CWL) [21] has been chosen. CWL is a workflow specification standard geared towards supporting interoperable and scalable production pipelines, abstracting away from the internal data structures of some of the language-specific workflow systems.

Originally designed for computational biology and with many available tool components, Galaxy [20] supports multiple domains. Workflows can be built by manually experimenting with data manipulations in a 'data playground' and subsequently converting histories of those to workflows, or by a more traditional drag-and-drop composition approach. New components can be created by wrapping existing programs, with in-built dependency management and automated conversion to executable containers. As such, Galaxy and CWL offer possibilities for a rich canonical workflow component landscape with a workflow management regime that can be both easily FAIR compliant and efficient internally [26]. The WorkflowHub, which facilitates CWL and enables workflows to be registered, shared and published, is mutually coupled with Galaxy so that workflows can be discovered in the Hub and immediately executed in a public-use Galaxy instance.

In the context of the SDR, users can construct institution or project-specific variants of digitization workflows to suit their specific needs. As collections are heterogeneous, different specimen types or specific sets of specimens are likely to have variations and idiosyncrasies in the digitization and processing needed. Tools for automated identification of specimens are likely to be taxon-specific, and as such it seems likely that taxon-specific workflows will become common. In addition, institutions have specific data exchange requirements for their individual collection management systems. Ensuring that workflows can be easily modified in a common environment bridges the gap between community contribution to shared tooling and the bespoke needs of specific institutions/collections.

Although computational workflows typically emphasize scalable automated processing, in practice many also combine automation with manual steps. This feature is also supported by Galaxy and CWL, allowing (for example) manual geocoding and verification during the digitization process of the locations where specimens were collected.

2.3 FAIR Digital Objects

Galaxy/CWL environments offer the possibility to integrate generic digital object methods [43, 44, 45] for the interactions between workflow components, thus making them able to meet the need and ease the burden of compiling FAIR compliant data throughout the research lifecycle [26].

A digital object exhibiting FAIR characteristics is a FAIR Digital Object [28] and is defined formally as "a unit composed of data and/or metadata regulated by structures or schemas, and with an assigned globally unique and persistent identifier (PID), which is findable, accessible, interoperable and reusable both by humans and computers for the reliable interpretation and processing of the data represented by the object".

Supporting 'FAIRness' internally and acting as glue between the steps of canonical workflows, FDOs record and can represent the state of a workflow, its inputs and outputs, and the component steps performed in a comprehensive manner [26]. Each FDO is anchored by a globally unique and resolvable, persistent identifier (PID) (such as a DOI®, for example) that clearly refers to one digital entity. The PID resolution offers persistent references to find, access and reuse all information entities that are relevant to access and interpret the content of an FDO. In doing so, the FDO creates a new kind of machine-actionable, meaningful and technology independent unit of information. This is both immediately available and amenable for further use, as well as being comparable to the role of the classical archival storage box when necessary.

2.3.1 Computable Digital Specimens as a Kind of FAIR Digital Object

Digital Specimens (DS) are a specific class of FDO that group, manage and allow processing of fragments of information relating to physical natural history specimens. On a one-to-one correspondence a DS authoritatively collates data about a physical specimen (i.e., information extracted and captured from labels by digitization workflows) with other data—often to be found from third-party sources—derived from analysis and use of the specimen.

openDS [46] is the developing specification for open Digital Specimens and other related object types, defining: i) the logical structure and content of Digital Specimen (DS), Basic Image Object (BIO) and other object types, and the operations permitted on them; ii) the handling rules and behaviors governing digital specimen object operations in general; and iii) serialization and packaging as JavaScript Object Notation (JSON) for lightweight data interchange between systems, sub-systems and components of systems (for which, read 'workflow components' [47]. openDS is essential to future FAIR digitization of natural history collections and to Digital Specimens as self-standing digital objects on the Internet, amenable to computer processing. It contributes to the new transformative generation of FAIR infrastructure and applications based on Digital Object Architecture that is planned for the Distributed System of Scientific Collections (DiSSCo) [6, 24, 29] European research infrastructure.

Henceforth we refer to these as openDS FDOs.

2.3.2 FAIR Packaging of Research/Workflow Objects with RO-crate

The useful outcomes of research are not just traditional publications nor data products but everything that goes into and supports an investigative work or production pipelining activity. This includes input and intermediate data, parameter settings, final outputs, software programs and workflows, and configuration information sufficient to make the work reproducible. Research objects [48] are a general approach to

describing and associating all of this content together in a machine-readable form so that it can be easily preserved, shared and exchanged. Workflow objects are a specific subclass of research objects.

RO-Crate [22, 23] has been established as a community standard to practically achieve FAIR packaging of research objects with their structured metadata. Based on well-established Web standards, RO-Crate uses JSON-LD [49] with Schema.org [50] for its common metadata representation. It is extensible with domain-specific vocabularies in a growing range of specializing RO-Crate profiles, e.g., for domains such as earth sciences [51], biosciences [52]; for object types such as data or workflow [53]; or for workflow runs). RO-Crate has been proposed for the implementation of FAIR Digital Objects on the World Wide Web as a common representation of the FDO Metadata objects foreseen by the FDO Framework [52, 30]. Combined with FAIR Signposting [54] for resolving persistent identifiers (PID) to FDOs on the World Wide Web, these RO-Crates are findable, accessible, interoperable, and reusable by machines to both create and obtain the information they need to function.

Henceforth, we refer to RO-Crate FDOs.

3. PROBLEM DESCRIPTION

3.1 Automating Digitization and Capturing the Process

In the lengthy history of collectors and museums curating artefacts and specimens, we see that there have been and always will be ambiguities, uncertainties, and inaccuracies in interpretations of recorded information and attached labels [55]. The practices of different collectors and curators vary and change over time. There are constraints of the label medium itself arising from the specifics of accepted preparation and preservation processes (e.g., tiny, handwritten labels pinned to butterflies).

Although systematic digitization of label and other recorded data can help to unify otherwise diverse information (e.g., species names, locations) the digital process and the resulting digital specimen data carry their own assumptions, simplifications, inconsistencies, and limitations. Over time, tools and methods, workflows and data models all evolve and improve. In particular, increasing automation for throughput and accuracy often involves increased assistance from computers and software.

Just as manual curation and improvement work implies the need for good record keeping, so too does working digitally imply the importance of ensuring that sufficient records are captured about the computer-assisted digitization and curation processes (provenance). These justify the produced digital specimen data and propagate credit for work done to their analogue equivalents, and also allow retrospective review, revision or recomputation of the produced data as future needs, practices or knowledge change.

Globally, there is underinvestment and missing technical expertise for wide-scale automated mass digitization. Sharing proven digitization workflows via a repository or registry linked to an individual published journal article presents significant barriers to re-use. Exploiting hosted community environments—in this case Galaxy and WorkflowHub—for the deployed tooling lowers barriers and provides rapid and

easy access for institutions with limited capabilities and capacities for digitization. Hosted workflows represent "primacy of method" for a community evolving towards a new research culture that is becoming increasingly dependent on working digitally and collaboratively [56, 57].

3.2 Users, User Stories and Specimen Categories

Initially, two kinds of users must be supported: digitizer technicians and collections managers/curators. Five high-level user stories describe and broadly encompass the functionality these users need:

1. As a digitizer, I want to construct a workflow from a set of predefined components, so I can use that workflow to digitize specimens to a predefined specification.
2. As a digitizer, I want to run one or many specimen images through a workflow so I can create new digital specimens.
3. As a collection manager/curator, I want to run one or many digital specimens through a workflow to enrich my digital specimens with further data.
4. As a collection manager/curator, I want to view the metadata of a digitization workflow run so I can understand what happened on that run.
5. As a digitizer, I want to export the output of a digitization run, so I can consume the output of a digitization run into my institution's collection management system.

To prove the SDR concept, three categories of preserved specimen types have been selected to be supported initially: herbarium sheets, microscope slides and pinned insects (Figure 1).

4. THE FDO AND CWFR APPROACH IN THE SPECIMEN DATA REFINERY

Workflows will be designed to support the user categories and stories given above. The performance of the SDR will be evaluated against these specimen types, eventually using several thousand different specimen and label images. This is in anticipation of SDR becoming part of the pivotal technology to achieve high rates of mass FAIR digitization expected through the planned DiSSCo research infrastructure [6, 24, 29].

4.1 FDO Types

In the Specimen Data Refinery (see figure 2) the role of openDS FDOs is planned as the basis for the primary workflow inputs and outputs, and for data transfer and interactions between components within SDR workflows. A single openDS FDO submitted to the beginning of the workflow (or a *de novo* digitization that is immediately wrapped as a new openDS FDO) becomes modified by each workflow component to produce an incrementally refined openDS FDO. FDOs are acting as the unit of data communication between canonical workflow components, in that each step is immediately creating an FDO with associated FAIR compliant documentation.

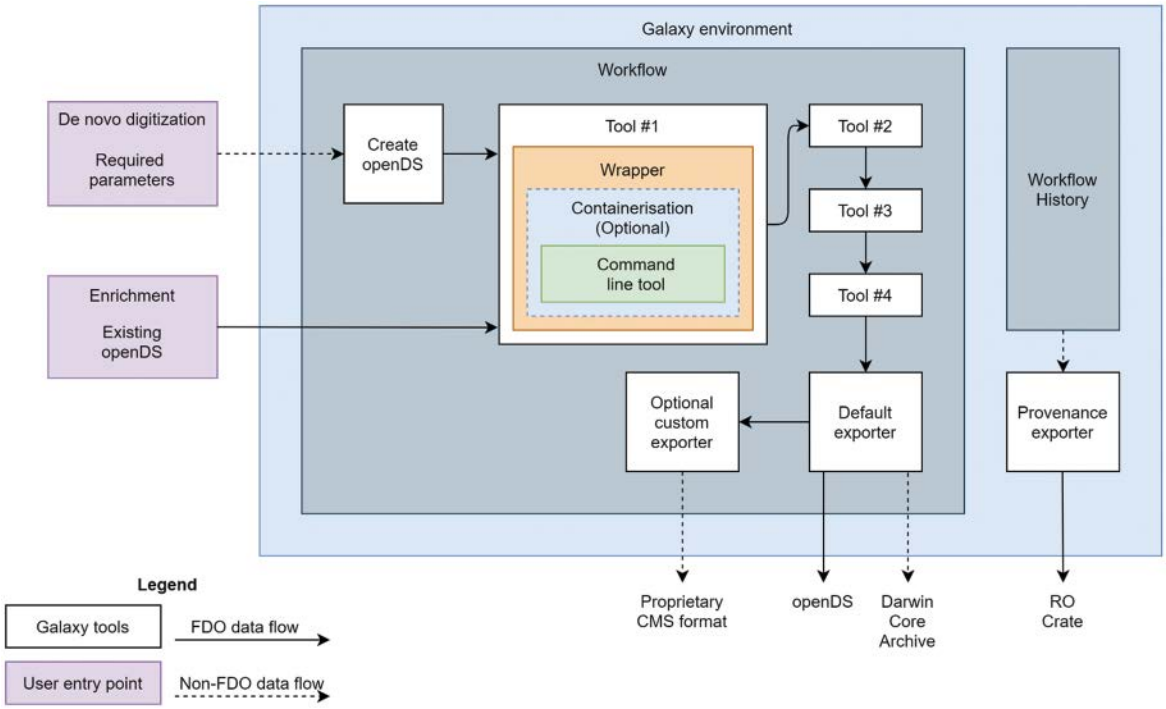


Figure 2. The CWFR approach adopted for the SDR as a Galaxy workflow management system implementation with '*de novo* digitization' and 'enrichment' entry points.

RO-Crate FDOs capture two aspects of a workflow:

a) A **Workflow-RO-Crate** contains the workflow definition, the computational tools and configuration, graphical image of the workflow, etc; this is the *method* registered in the WorkflowHub and activated in the Specimen Data Refinery for execution.
b) A **Workflow-Run-RO-Crate** references (a) and records the details of a specific computational workflow run and its runtime information, with relations to the used and generated FDOs. This captures the digitization provenance that is generated as the openDS FDO makes its journey through a workflow.

The final step in SDR workflows can be a data exporter tool, allowing users to export the entire openDS object as is, or to convert and export it in another format, such as CSV, DarwinCore Archive, etc. This reconfigurable nature of data export allows users to define their own transformer function to allow export to match formats specific to specific collection management systems in use by their institution, such that refined data can be repatriated. The provenance exporter transforms Galaxy workflow history data into a Workflow-Run-RO-Crate FDO.

All FDO types are serialized as JSON.

4.2 Canonical Components

Each workflow component from a canonical library (to be built, illustrated as tools #1–#4 in Figure 2) describes what attributes it requires from the openDS FDO to be able to function, and the attributes it will in turn populate or enrich. The interface between the component and the openDS FDO is formed by the wrapper (orange in Figure 2) around the (optionally) containerized command-line tool (blue/green in figure 2). Canonical components can be used individually, or in series. The openDS FDO data flows between components will always be of the same type, being modified as the workflow proceeds.

This allows tools to function both as standalone components and as part of any sequence of chained tools, provided that the specific openDS FDO attributes required for a tool to work are pre-populated. This keeps the SDR flexible and customisable for different digitization pipelines.

Two entry points are provided for users of the SDR. One is named as the '*de novo* digitization' entry point (figure 2) fulfilling the needs of user story (2) where specimens are being newly digitized for the first time. The second entry point, named as the 'enrichment' entry point (Figure 2) fulfils the needs of user story (3) where an existing openDS FDO (or reference to it) can be provided to the SDR as part of the input data.

In parallel to manipulating openDS FDOs, the Refinery gathers the minimum inputs and workflow components required to produce deterministic output and produces a Workflow-Run-RO-Crate FDO.

5. EXPERIMENTS AND ANALYSIS

5.1 Experimental Workflows

The workflows of the SDR compose different functional components according to specific need: image segmentation, barcode reading, optical character recognition, text contextualisation/entity recognition, geocoding, taxonomic linkage, people linkage, specimen condition checking, automated identification, and data export/conversion. Broadly speaking there are two main kinds of workflow: i) specimen workflows, where the specimen itself is analysed for morphological traits, colour analysis, condition checking and automated identification; and ii) text and label workflows, where handwritten, typed or printed text from the image is read, named entities are classified, then linked to identifiers or enhanced through post processing.

Both kinds of workflow can begin with initial openDS object creation based on the submission of specimen image files and accompanying input parameters through a forms-based user interface (*de novo* digitization entry point); or, alternatively, a pre-existing openDS object with accompanying image object(s) can be supplied as the input (enrichment entry point). Both kinds of workflow also rely on the image segmentation component as the precursor for subsequent workflow steps. Similarly, and if needed both kinds might use a format conversion and export component as their final step; for example, if an openDS FDO is not a natively compatible output for the next consuming application.

Although not within the scope of the present proof-of-concept, other more precise workflows for enhancing specific aspects of existing records can be foreseen. There are many specimen records, for example where

locality text, although digitally available, is not yet geocoded. There are records with unlinked or ambiguous collector names that could be linked/disambiguated; and records where unknown specimens still need identifying.

5.2 Experimental Data and Evaluation

5.2.1 Evaluation Images Datasets

The Refinery will be evaluated using sets of images, each composed of at least 1,000 unique specimens for each of the three categories of preserved specimen types: herbarium sheets, microscope slides and pinned insects. For herbarium sheet images we will reuse an existing benchmark dataset of 1,800 herbarium specimen images with corresponding transcribed data [58]. For microscope slide and pinned insect specimen images similar evaluation datasets will be prepared against the same label characteristics: written in different languages; printed or handwritten; covering a wide range of dates; both type specimens and general collections and will provide specimens from different families and different parts of the world. Each test dataset set will be composed of images from different institutions to ensure representation of heterogeneity. For the present proof of concept, we limit the scope to Latin alphabet languages. These datasets will also be used to train Refinery models for use in tools (e.g., segmentation, named entity recognition, object/feature detection). All the datasets will be made publicly available with documentation.

5.2.2 Component Functional Tests

Galaxy has a built-in functional test framework. Tools intended to become components of an SDR canonical library (actually, a Galaxy ToolShed repository) will need to pass previously defined tests within this framework. These tests, based on pre-supplied openDS FDO input and output files containing the properties expected to be populated by the tool, include validating a tool's own openDS FDO outputs by comparison against the expected output file. It will be necessary to register openDS FDOs as Galaxy custom data types.

6. RESULTS

openDS FDOs are the core data object at the heart of the SDR, playing not only the workflow input/output role but acting also as a common data structure between tool steps within the workflow. Users can launch the workflow with either an openDS object, for further augmentation by the SDR, or they can complete a form with the specimen information, which is then converted to an openDS object before the workflow proper begins.

Each SDR Galaxy tool defines the properties it requires in JSONPath syntax [59]. The wrapper validates that these properties exist in the openDS object, plucks them from the openDS JSON, makes them available as named parameters, and passes these through to the tool processing (via either a Docker or Python command line). The wrapper validates the input openDS against the openDS schema, the tool performs its

processing and updates the openDS, and the wrapper validates the changed openDS against the schema before writing to disk. For the prototypical SDR, a static, local version of the openDS schema is used. Future iterations will use referenceable versions of the openDS schema, allowing for schema changes and for tools to validate their data input and outputs against versions of the schema.

On ingestion, every openDS is assigned a persistent identifier, ensuring unambiguity and referential integrity for every processed object. In production, DOIs will be minted by the DiSSCo service; for the proof-of-concept Handles with prefix 20.5000.1025/ will be used.

7. DISCUSSION

7.1 What is Being Achieved?

The design of the Specimen Data Refinery uses two kinds of FAIR Digital Object—openDS FDOs and RO-Crate FDOs. Each plays a role to ensure 'FAIRer' automated digitization for natural history specimens and associated provenance capture:

- openDS FDOs act both as the input/output interface of a workflow and as the common intermediary pattern (canonical state) between steps within a workflow. They comply with DiSSCo data management principles and needs as outlined in the DiSSCo Data Management Plan [27] allowing specimen data to be processed and extended in a fully FAIR manner [6].
- RO-Crate FDOs record both the workflow definition and information about its configuration (shared as a method object) together with the details and context of the work done during a workflow run; details that are captured proprietarily within the adopted Galaxy environment and transformed to a common pattern (as another kind of canonical state) of provenance for later scrutiny and reproducibility of the work. These kinds of Research Objects [48] are an established mechanism whereby computational methods become first-class citizens alongside data, to be easily shared, discussed, reused and repurposed [56].

Both kinds of FDO are essential. They complement one another to support implementation of the FAIR principles, especially the interoperable and reusable principles by making workflows self-documenting. This renders automated whole processes (or fragments thereof) for digitizing and extending natural history specimens' data as FAIR without adding additional load to the researchers that stand to benefit most from that [26]. Each FDO type originates from different Research Infrastructures (ELIXIR, DiSSCo) with different implementation frameworks. Yet, they interoperate effectively due to their clear roles, common conceptual model and separation of concerns.

7.2 Different FDO Implementations Working Together

openDS FDOs have their heritage in distributed digital object services [45] and are implemented through Digital Object Architecture (DOA) [60] with Digital Object Interface Protocol (DOIP) [61], Digital Object Identifier Resolution Protocol (DO-IRP) [62], and recommendations of the Research Data Alliance [63].

Serialized as JSON, they are machine-actionable and compatible with established protocols of the World Wide Web.

RO-Crates are native to the World Wide Web, based on established web protocols, machine-readable metadata using Linked Data Platform methods [64], JSON-LD and Schema.org [48], and community-accepted packaging mechanisms such as BagIt. This makes RO-Crates straightforward to incorporate into pre-existing platforms such as Galaxy and data repositories such as Zenodo and DataVerse.

Both kinds of FDO use Persistent identifiers (PID), allowing instances to be both uniquely identified and their location to be determined; RO-Crates, as web natives, use URIs whereas openDS, as DOA objects, use Handle PIDs. Instances of both kinds are described by metadata and contain or reference data.

RO-Crates are self-describing using a metadata file and use openly-extensible profiles to type the Crates (profile-typing) to set out expectations for their metadata and content. openDS uses an object-oriented object typing and instance approach to define the structure and content of data/metadata. Complex object types are constructed from basic types, and an extension-section basic type. Both approaches seek to avoid locking objects into repository silos, ensuring that FDO instances can be interpreted outside of the contexts in which they were originally created/stored.

Structurally and semantically openDS FDOs and RO-Crate FDOs are potentially isomorphic, although at different granularity levels. Their main difference is in method calling. As a DOA object, openDS would expect to respond to type-specific method calls if these were implemented. RO-Crates delegate actionability to applications that interpret their self-describing profile.

Within the SDR the two kinds of FDO fulfill distinct and interlocking roles for data (openDS) and self-documented method (RO-Crate) so their different forms is not an issue. In future there may be a need to map and convert between the approaches (e.g., for reconstructing past processing), which would be assisted by the common FDO conceptual model [30].

7.3 Key Domain Challenges Ahead

For a digitized specimen to conform to FAIR principles, its data must be linked to a vocabulary of terms, but choosing a single vocabulary is likely to cause interoperability issues when cross-linking to resources using another vocabulary, for example Darwin Core, Schema.org, or Access to Biological Collection Data (ABCD). Whilst concepts can be mapped across vocabularies (for example, using Simple Knowledge Organization System (SKOS) matching), such an effort might rapidly become overly complex and cumbersome, as the challenge of the Unified Medical Language System (UMLS) demonstrates. The challenge remains—how is such a mapping exercise maintained at a 'just enough' level?

Different Earth Science domains have different use cases for digital records. A digital record produced for biodiversity research is likely to have different granularity, understanding and focus to one produced for climate science. It remains to be seen if a single FAIR Digital Object definition could be produced to satisfy

multiple domains, and if different objects could be produced for different domains, what would they look like; and would this hinder future cross-compatibility?

The openDS FDO type produced by the SDR is a new object format for the natural history domain that is foreseen to become an adopted standard over time. Institutional collection management systems will need to be upgraded before they can consume the FDO outputs from the SDR. Early adopters may need assistance to produce SDR exporters matching proprietary ingestion formats. For an interim period, there may be a need for the SDR to output today's widely used Darwin Core Archives format in parallel.

As the functional requirements of the SDR are emergent, a minimum viable product has been scoped, but this should be contrasted with the notion of a useful product. An MVP is a prototype; a tool to get a project off the ground with enough features to be usable by early adopters, and to build on to learn user requirements. But it is not intended to meet the day-to-day requirements of all users. To nurture future development, care must be taken to continue involving key stakeholders in eliciting further requirements to make the SDR useful for the widest range of users, and from there, develop a rich, configurable tool to allow simple uptake and provide utility for resource-poor collections.

8. CONCLUSION AND FUTURE WORK

The Specimen Data Refinery is likely to garner widespread interest across the Natural History community. Whilst the promise of a scalable, community-driven digitization platform is tantalising for many natural history professionals, the Specimen Data Refinery project is still in its early stages, and, as discussed above key challenges lie ahead.

Although natural history collections are generally catalogued by the taxonomic identity of the curated object, there remains a large historical backlog of unidentified specimens. The Meise Botanic Garden (BE), for example, has an estimated 4 million specimens with at least 11% not yet identified to species level. Furthermore, it is calculated that half of the World's estimated 70,000 plant species yet to be described have already been collected and are waiting in collections still to be 'discovered' [21]. The same is likely to be true for other groups of organisms, especially insects. Unnamed specimens tend to have lowest priority for digitization and their data are rarely shared. Machine learning as canonical steps in SDR workflows presents a tremendous opportunity to put an identification on these specimens and potentially, to triage them for further taxonomic investigation.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 823827 (SYNTHESYS Plus), 871043 (DiSSCo Prepare), 823830 (BioExcel-2), 824087 (EOSC-Life).

AUTHOR CONTRIBUTIONS

A. Hardisty (hardistyar@gmail.com): Conceptualization, Investigation, Supervision, Validation, Writing—original draft, Writing—review & editing, Approval.

P. Brack (paul.brack@manchester.ac.uk): Investigation, Writing—original draft, Writing—review & editing.

C. Goble (carole.goble@manchester.ac.uk): Conceptualization, Supervision, Writing—review & editing.

L. Livermore (l.livermore@nhm.ac.uk): Conceptualization, Funding acquisition, Investigation, Writing—original draft, Writing—review & editing.

B. Scott (b.scott@nhm.ac.uk): Investigation, Writing—original draft, Writing—review & editing.

Q. Groom (quentin.groom@plantentuinmeise.be): Funding acquisition, Investigation, Writing—original draft, Writing—review & editing.

S. Owen (stuart.owen@manchester.ac.uk): Investigation, Writing—original draft, Writing—review & editing.

S. Soiland-Reyes (soiland-reyes@manchester.ac.uk): Investigation, Writing—original draft, Writing—review & editing.

REFERENCES

[1] Walton, S., et al.: Landscape analysis for the specimen data refinery. Research Ideas and Outcomes 6, e57602 (2020)
[2] Thiers, B.M., Tulig, M.C., Watson, K.A.: Digitization of the New York Botanical Garden herbarium. Brittonia 68(3), 324–333 (2016)
[3] Nelson, G., Ellis, S.: The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B: Biological Sciences 374, 20170391 (2019)
[4] Nelson, G., Paul, D.L.: DiSSCo, iDigBio and the future of global collaboration. Biodiversity Information Science and Standards 3, e37896 (2019)
[5] Addink, W., Koureas, D., Rubio, A.: DiSSCo as a new regional model for scientific collections in Europe. Biodiversity Information Science and Standards 3, e37502 (2019)
[6] Lannom, L., Koureas, D., Hardisty, A.R.: FAIR data and services in biodiversity science and geoscience. Data Intelligence 2(1–2),122–30 (2020)
[7] GBIF Secretariat: GBIF Science Review 2020. Available at: https://doi.org/10.35035/bezp-jj23. Accessed 9 September 2021
[8] Heberling, J.M., et al.: Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118(6), e2018093118 (2021)
[9] Sweeney, P.W., et al.: Large–scale digitization of herbarium specimens: Development and usage of an automated, high–throughput conveyor system. Taxon 67, 165–178 (2018)
[10] Allan, E.L., et al.: A novel automated mass digitisation workflow for natural history microscope slides. Biodiversity Data Journal 7, e32342 (2019)

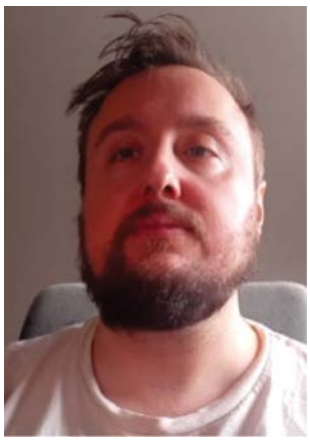
[11] Hereld, M., Ferrier, N.: LightningBug ONE: An experiment in high-throughput digitization of pinned insects. Biodiversity Information Science and Standards 3, e37228 (2019)
[12] Price, B.W., et al.: ALICE: Angled label image capture and extraction for high throughput insect specimen digitisation. arXiv preprint OSFPreprints:10.31219/osf.io/s2p73 (2018)
[13] Tegelberg, R., et al.: Mass digitization of individual pinned insects using conveyor-driven imaging. In: 2017 IEEE 13th International Conference on E-Science (e-Science), pp. 523–527 (2017)
[14] Heberling, J.M., Prather, L.A., Tonsor, S.J.: The changing uses of herbarium data in an era of global change: An overview using automated content analysis, BioScience 69(10), 812–822 (2019)
[15] Heather, M., et al.: Using insect natural history collections to study global change impacts: challenges and opportunities. Philosophical Transactions of the Royal Society B 374(1763) (2019)
[16] Watanabe, M.E.: The evolution of natural history collections: New research tools move specimens, data to center stage. BioScience 69(3), 163–169 (2019)
[17] Nic Lughadha, E.M., et al.: Harnessing the potential of integrated systematics for conservation of taxonomically complex, megadiverse plant groups. Conservation Biology 33, 511–522 (2019)
[18] Owen, D., et al.: Towards a scientific workflow featuring natural language processing for the digitisation of natural history collections. Research Ideas and Outcomes 6, e58030 (2020)
[19] Harrow, J., et al.: ELIXIR-EXCELERATE: Establishing Europe's data infrastructure for the life science research of the future. EMBO Journal 40(6), e107409 (2021)
[20] Afgan, E., et al.: The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46, W537–W544 (2018)
[21] Crusoe, M.R., et al.: Methods included: Standardizing computational reuse and portability with the common workflow language. arXiv preprint arXiv: 2105.07028 (2021)
[22] Carragáin, E.Ó., et al.: A lightweight approach to research object data packaging. In: Bioinformatics Open Source Conference (BOSC 2019). Available at: https://doi.org/10.5281/zenodo.3250687. Accessed 9 September 2021
[23] Soiland-Reyes, S., et al.: Packaging research artefacts with RO-Crate. Data Science (accepted). arXiv preprint arXiv: 2108.06503 (2021)
[24] Goble, C., et al.: Implementing FAIR digital objects in the EOSC-Life workflow collaboratory. (2021). Available at: https://doi.org/10.5281/zenodo.4605654. Accessed 9 September 2021
[25] Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, Article No. 160018 (2016)
[26] Wittenburg, P., et al.: Canonical Workflows to Make Data FAIR. Data Intelligence 4(2), 286–305 (2022)
[27] Hardisty, A.: Provisional data management plan for DiSSCo infrastructure. Deliverable D6.6. ICEDIG (2019). Available at: https://doi.org/10.5281/zenodo.3532937. Accessed 9 September 2021
[28] De Smedt, K., Koureas, D., Wittenburg, P.: FAIR digital objects for science: From data pieces to actionable knowledge units. Publications 8(2), Article No. 21 (2020)
[29] Hardisty, A., et al.: Conceptual design blueprint for the DiSSCo digitization infrastructure—DELIVERABLE D8.1. Research Ideas and Outcomes 6, e54280 (2020)
[30] FDO Coordination Group (2020) FDO Framework. Available at: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF. Accessed 10 August 2021
[31] Triki, A., et al.: Objects detection from digitized herbarium specimen based on improved YOLO V3. In: VISAPP 2020—15th International Conference on Computer Vision Theory and Applications, pp. 523–529 (2020)

[32] Nieva de la Hidalga, A., et al.: Cross-validation of a semantic segmentation network for natural history collection specimens (Accepted). Available at: https://doi.org/10.1007/s00138-022-01276-z. Accessed 2 March 2022
[33] Walton, S., et al.: A cost analysis of transcription systems. Research Ideas and Outcomes 6, e56211 (2020)
[34] Groom, Q., et al.: People are essential to linking biodiversity data. Database 2020, baaa072 (2020)
[35] Knyshov, A., Hoang, S., Weirauch, C.: Pretrained convolutional neural networks perform well in a challenging test case: Identification of plant bugs (Hemiptera: Miridae) using a small number of training images. Insect Systematics and Diversity 5(2), 3 (2021)
[36] Hussein, B.R., et al.: Application of computer vision and machine learning for digitized herbarium specimens: A systematic literature review. arXiv preprint arXiv: 2104.08732v1 (2021)
[37] Carranza-Rojas, J., et al.: Going deeper in the automated identification of herbarium specimens. BMC Evolutionary Biology 17, Article No. 181 (2017)
[38] Little, D.P., et al.: An algorithm competition for automatic species identification from herbarium specimens. Applications in Plant Sciences 8(6), e11365 (2020)
[39] Pryer, K.M., et al.: Using computer vision on herbarium specimen images to discriminate among closely related horsetails (Equisetum). Applications in Plant Sciences 8(6), e11372 (2020)
[40] Unger, J., Merhof, D., Renner, S.: Computer vision applied to herbarium specimens of German trees: Testing the future utility of the millions of herbarium specimen images for automated identification. BMC Evolutionary Biology 16, Article No. 248 (2016)
[41] Atkinson, M., et al.: Scientific workflows: Past, present and future. Future Generation Computer Systems 75, 216–227 (2017)
[42] Amstutz, P., et al.: Existing workflow systems. Common workflow language Wiki, GitHub. Available at: https://s.apache.org/existing-workflow-systems. Accessed 9 September 2021
[43] Hui, Y.: What is a digital object? Metaphilosophy 43, 380–395 (2012)
[44] Kallinikos, J., Aaltonen, A., Marton, A.: The ambivalent ontology of digital artifacts. MIS Quarterly 37, 357–370 (2013)
[45] Kahn, R., Wilensky, R.: A framework for distributed digital object services. International Journal on Digital Libraries 6, 115–123 (2006)
[46] openDS: Draft specification for open Digital Specimens (openDS). Available at: https://github.com/DiSSCo/openDS. Accessed 10 August 2021
[47] Bray, T.: The JavaScript Object Notation (JSON) data interchange format (Request for Comments No. RFC 8259). Internet Engineering Task Force (2017). Available at: https://doi.org/10.17487/RFC8259. Accessed 10 August 2021
[48] Bechhofer, S., et al.: Why linked data is not enough for scientists. Future Generation Computer Systems, Special section: Recent advances in e-Science 29, 599–611 (2013)
[49] Kellogg, G., Champin, P.A., Longley, D.: JSON-LD 1.1 A JSON-based serialization for linked data. W3C Recommendation. Available at: https://www.w3.org/TR/json-ld11/. Accessed 10 August 2021
[50] Schema.org—Schema.org. Available at: https://schema.org/. Accessed 10 August 2021
[51] Corcho, O., et al.: D5.1 RO model adapted to EOSC. RELIANCE deliverable, Zenodo. Available at: https://doi.org/10.5281/zenodo.4913285. Accessed 10 August 2021
[52] Goble, C., et al.: Implementing FAIR digital objects in the EOSC-Life workflow collaboratory. Available at: https://doi.org/10.5281/zenodo.4605654. Accessed 10 August 2021
[53] Bacall, F., Williams, A.R., Soiland-Reyes, S.: Workflow RO-Crate profile 1.0. WorkflowHub community. Available at: https://w3id.org/workflowhub/workflow-ro-crate/1.0. Accessed 17 November 2021

[54] Van de Sompel, H., et al.: FAIR signposting profile. 2021-04-20. Available at: https://www.signposting.org/FAIR/. Accessed 10 August 2021
[55] Lohonya, K., Livermore, L., Penn, M.G.: Georeferencing the natural history museum's Chinese type collection of plateaus, pagodas and plants. Biodiversity Data Journal 8, e50503 (2020)
[56] De Roure, D., Goble, C.: Anchors in shifting sand: The primacy of method in the Web of data. In: Web Science Conference, pp. 26–27 (2010). Available at: https://eprints.soton.ac.uk/270817/. Accessed10 August 2021
[57] Hardisty, A.R., et al.: BioVeL: A virtual laboratory for data analysis and modelling in biodiversity science and ecology. BMC Ecology 16, 49 (2016)
[58] Dillen, M., et al.: A benchmark dataset of herbarium specimen images with label data. Biodiversity Data Journal 7, e31817 (2019)
[59] Gössner, S., Normington, G., Bormann, C.: JSONPath: Query expressions for JSON. Internet-Draft draft-ietf-jsonpath-base-02. Internet Engineering Task Force. Available at: https://datatracker.ietf.org/doc/html/draft-ietf-jsonpath-base-02. Accessed 10 August 2021
[60] DONA Foundation: Digital object architecture. Available at: https://www.dona.net/digitalobjectarchitecture. Accessed 10 August 2021
[61] Digital Object Interface Protocol Specification, version 2.0, November 2018. Available at: https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf. Accessed 10 August 2021
[62] Sun, S., et al.: RFC 3652 Handle System Protocol (ver 2.1) Specification. RFC Editor, USA. Available at: https://doi.org/10.17487/RFC3652. Accessed 10 August 2021
[63] Islam, S., et al.: Incorporating RDA outputs in the design of a European research infrastructure for natural history collections. Data Science Journal 19(50), 1–14 (2020)
[64] Speicher, S., Arwe, J., Malhotra, A.: Linked data platform 1.0. W3C Recommendation. Available at: https://www.w3.org/TR/ldp/. Accessed 10 August 2021

AUTHOR BIOGRAPHY

Alex Hardisty was before his recent retirement Director of Informatics Projects in the School of Computer Science and Informatics, Cardiff University, UK. He is interested in bio/geodiversity informatics, the engineering of large-scale distributed information systems for data management and processing, virtual research environments and socio-technical issues of new technology adoption. Alex is a technical architect. Before his retirement he was leading DiSSCo technical work on open Digital Specimens (openDS), Minimum Information about Digital Specimens/Collections (MIDS/MICS) and exploiting machine-actionable FAIR Digital Objects. Alex was co-chairing the CWFR Working Group of the FDO Forum and was a member of the FDO Forum's Technical Specification and Implementation (TSIG) Working Group and the FDO Forum Steering Committee.
ORCID: 0000-0002-0767-4310



Paul Brack is a Research Associate in the eScience Lab at The University of Manchester where he works on the SYNTHESYS+ project with a focus on computational workflows. With an MPhil in proteomics, Paul has previously worked with genomics, cancer sciences, oncology, and health informatics.
ORCID: 0000-0002-5432-2748

Carole Goble is Professor of Computer Science at the University of Manchester and the founder of the eScience Group. Over 30 years she has applied technical advances in knowledge technologies, distributed computing, workflows, and social computing to solve information management problems for Life Scientists, Biodiversity and other scientific disciplines. She led the developments of scientific workflow systems (Taverna), methods for exchanging scientific outcomes (Research Objects) and the sharing of workflows (myExperiment, WorkflowHub) and project results (FAIRDOM). Her research interests are in reproducible research, asset curation and preservation, computational analytics, semantic interoperability, knowledge exchange between scientists and new models of scholarly communication. She currently coordinates the FAIRDOM initiative, serves as the Head of ELIXIR-UK (the national node of the ELIXIR European Research Infrastructure for life science data), manages the FAIR results of several other EU projects and is a co-founder of the UK's Software Sustainability Institute. She currently serves as the UK expert on the G7 Open Science Working Group and is director of FAIR Computational Workflows for the Workflow Community Initiative.
ORCID: 0000-0003-1219-2137

Working at the Natural History Museum, London, **Laurence Livermore** is a digital project manager with over a decade of experience in natural science collections. He specializes in digital innovation, mass digitization, biodiversity informatics and data. He enjoys working with data, promoting open access and collaborating to solve digital challenges faced by similar organizations.
ORCID: 0000-0002-7341-1842

Ben Scott is Technical Lead of Biodiversity Informatics at the Natural History Museum, London (NHM). He specializes in semantic open data and machine learning, building tools to explore, visualize and augment museum research and collections. Projects include the NHM's open data portal; automated trait extraction from botanical literature and herbarium sheets; using ships' logs to enhance the provenance of historical specimens; and using machine learning to automate collection digitization pipelines.
ORCID: 0000-0002-5590-7174

Quentin Groom is head of Biodiversity Informatics at Meise Botanic Garden. Quentin works at the interface of botany and information technology. His work has included projects such as the digitization of the herbarium and the digitization of the Flore d'Afrique Centrale. He often focuses on issues related to the interoperability of data, including biodiversity data standards and digitization quality. Throughout this time, he has also worked with invasion biologists and citizen scientists to increase and improve our data and knowledge on invasive species. When he has time, he likes to apply innovative data science and statistical techniques to biodiversity data to reveal what those data tell us.
ORCID: 0000-0002-0596-5376

Stuart Owen has been a Senior Research Software Engineer, working at The University of Manchester within the eScience Lab since 2006. Since 2008 Stuart has been the lead developer of the FAIRDOM-SEEK Commons and Cataloguing platform for supporting collaborative projects in the Life Sciences, and he leads the technical team for FAIRDOM initiative, including the FAIRDOMhub.org, an aggregation commons structuring and cataloguing data, models, etc. held in distributed repositories. More recently, Stuart has been working extending and adapting the FAIRDOM-SEEK to be a Workflow registry within EOSC-Life. He also leads the development of the RightField annotation platform, which is used for seamlessly embedding semantic Web terms within spreadsheets.
ORCID: 0000-0003-2130-0865

Stian Soiland-Reyes is a Technical Architect in the eScience Lab at The University of Manchester, and a Ph.D. Candidate in the INDE lab at University of Amsterdam. Since 2006 he has worked as a software engineer and researcher focusing on reproducibility, scientific workflows, interoperability, linked data, metadata, and open science. He is a persistent advocate of Open Scholarly Communication and is on the leadership teams of Common Workflow Language (CWL) and BioCompute Objects. He contributed to the W3C provenance model PROV-O and multiple Linked Data initiatives. He co-created the Research Object model and is co-chair of the Research Object Crate community.
ORCID: 0000-0001-9842-9718