Project:	Synthesis of systematic resources
Project acronym:	SYNTHESYS PLUS
Grant Agreement number:	823827
Work Package:	Workpackage 8
Deliverable number:	8.3
Deliverable title:	Development of cloud platform for data-processing services
Deliverable author(s):	Laurence Livermore - Natural History Museum, London Carole Goble - The University of Manchester Helen Hardy - Natural History Museum, London Ben Scott - Natural History Museum, London Stian Soiland-Reyes - The University of Manchester Oliver Woolland - The University of Manchester
Date:	2023-08-12

1. Background

A key limiting factor in organising and using information from global natural history specimens is making that information computable. More than 95% of available information currently resides on labels attached to specimens or in physical registers and is not in a digital format at all. The scale of the task to digitise all the specimens held in natural history collections has required a staged process of digitisation, prioritising images, and basic catalogue records rather than capturing computable data about them (e.g., transcribing and linking data from labels, or creating descriptive morphological descriptions).

In the SYNTHESYS+ project, the Specimen Data Refinery (SDR) work package (WP8) had the objective of building a prototype cloud-based platform with tools and services to automate the extraction, enhancement, and annotation of specimen images. We envisaged building a modular system that could be used in different digitisation workflows and collections and could be used by a range of staff involved in digitisation or digital curation of collections. We chose to adopt a user-configurable approach because we assumed prospective users would want to customise their own workflows, and that the trade-off between configurability and complexity would be worthwhile.

This report follows on from the landscape analysis report [Walton et al, 2020a] and the tool and service development report [Livermore et al, 2023a]. It is the formal report for the software demonstrator Deliverable 8.3. It describes the technology, development, and design approach of a *cloud platform for data-processing services*.

A key concept behind our approach is FAIR workflows, where a workflow is a chain of analysis or tool steps. Many current approaches to creating digitised specimens, including those used by the authors, can be hard to reproduce or share with others in the absence of formalised workflows. We explore a more formal approach to workflows and discuss the pros and cons around the platform technology and (semi) automated software deployment from the perspective of the SDR use case. Deployment in this sense covers all the software dependencies, steps, and processes to make the workflow platform available to its users - in this case making the installation of the Galaxy workflow platform and the SDR components as easy and automated as possible for other system administrators.

1.1 Scope

This report primarily describes the customisation and development on the Galaxy cloud platform and deployment of the Specimen Data Refinery.

An initial landscape and gap analysis of platforms and training datasets was undertaken in Deliverable 8.1 - see [Walton et al, 2020a] (the report for Task 8.1).

Tools and services development are summarised in [Livermore et al, 2023a] (the report for Task 8.2).

[Livermore et al, 2023b] covers SDR integration with the wider DiSSCo architecture; documentation and evaluation; and dissemination and promotion of the SDR.

2. Technology Choices

2.1 Workflow Platform

A workflow platform (or workflow management system) is needed to define and execute workflows in a reusable way and keep a record of the processing. Hardisty et al (2022a) note that these systems may offer a variety of features, e.g., 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; and licensing.

In the original description of work we had yet to settle on which of the many workflow platforms we would use for the Specimen Data Refinery, but we eventually decided upon Galaxy [Galaxy Community, 2022], in conjunction with Common Workflow Language (CWL) [Crusoe et al, 2022], an open standard for describing how to run command line tools and connect them to create interoperable workflows.

While Galaxy was originally designed for computational biology it now has many available tool components and supports multiple domains. It is becoming widely used across many different domains, including biodiversity informatics [Royaux et al 2022].

One of the key features of Galaxy is the abstraction of complexity, both computational infrastructure and tool complexity (Figure 1).

In Galaxy, 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 [from Hardisty et al (2022)]. This means that all the underlying data, tool/workflow configuration parameters are preserved using Galaxy's histories - an inbuilt workspace that acts both as a record of provenance, and to show a user's analysis over time. These histories can be annotated and shared with others.

Galaxy functionality offers features specified in the task description for SDR, including a common entry webpage that directs users to tools or workflows, and an API endpoint.

Galaxy scales over multiple cloud and cluster compute systems[1] such as PULSAR network[2], and current work in the Horizon Europe EuroScienceGateway project[3] extends Galaxy execution capability to other compute infrastructures including HPC.

[1] Galaxy Project "Connecting to a Cluster": https://docs.galaxyproject.org/en/latest/admin/cluster.html
[2] Pulsar Network's documentation: https://pulsar-network.readthedocs.io/en/latest/
[3] EuroScienceGateway: https://esciencelab.org.uk/projects/eurosciencegateway/

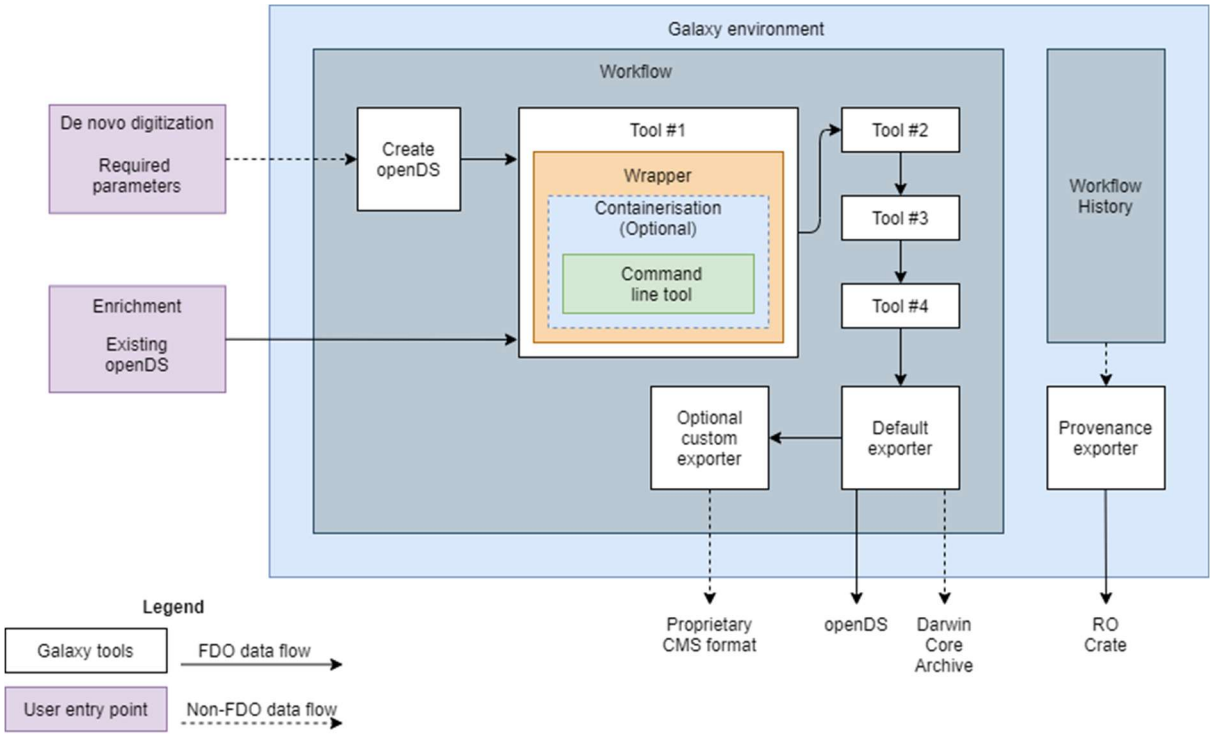

Figure 1: *Overview of the general structure of the SDR with a generic workflow within the Galaxy environment. Reproduced from [Hardisty et al, 2022] under the CC-BY 4.0 licence.*

The SDR has been able to considerably leverage developments in the EOSC-Life Cluster project and its development of the Workflow Collaboratory [Goble, et al 2021]. The Workflow Collaboratory offers an ecosystem of interoperating services for researchers and workflow specialists to find, use and reuse workflows, and deploy them using European Open Science Cloud (EOSC) infrastructure. A web-friendly metadata framework supports the description and exchange of workflows across the services. SDR used the following interoperating elements and standards:

- *The WorkflowHub registry*[4] supports the finding and sharing of workflows and supports workflow FAIRness through rich metadata. Five SDR workflows are registered in the WorkflowHub team *Specimen Data Refinery*[5] within the WorkflowHub space *DISSCo*[6].
- *The Bio.tools registry*[7] [Ison et al, 2019] supports the finding of tools. The SDR bio.tools collection[8] aims to collect the tools used in SDR.
- *Bioschemas*[9] [Gray et al, 2023] schema.org profiles for Computational Tool, Computational Workflow and Formal Parameter provide metadata about a workflow and its tools that are discipline independent, despite the "bio" prefix. The *EDAM Ontology*[10] [Ison et al, 2013] adds

[4] https://workflowhub.eu
[5] https://workflowhub.eu/projects/72
[6] https://workflowhub.eu/programmes/15
[7] https://bio.tools
[8] https://bio.tools/t?collectionID=%22Specimen%20data%20refinery%22
[9] https://bioschemas.org
[10] https://edamontology.org

informatics-specific metadata, such as strong typing of inputs and outputs and describes the overall workflow topics and operations to help find workflows.

- *RO-Crate*[11] [Soiland-Reyes et al, 2022a], a community-developed standardised approach for FAIR Digital Objects [Soiland-Reyes et al, 2022c], packages executable workflows, their components, such as example and test data, abstract CWL, diagrams and their metadata, making workflows more readily re-usable. RO-Crate is the unit of currency of exchange between the services, archiving workflows in public repositories such as Zenodo, and recording the provenance of workflow runs [Leo et al, 2023]. RO-Crates use Bioschemas profiles for describing workflows. [Soiland-Reyes 2023c].
- The *GA4GH Tools Registry Service API*[12] supports the exchange of scientific tools and workflows and enables users to search for and retrieve metadata about registered tools, so that workflow execution platforms can search and import workflows from WorkflowHub and WorkflowHub can directly launch workflows.
- The *Common Workflow Language*[13] [Crusoe et al, 2022] is encouraged as a canonical workflow description to accompany native workflow definitions. CWL represents the structure and steps of workflows in an interoperable way across workflow languages. The CWL representation may be *abstract* (detailing the steps and their connections, but now how each step is executed) or concrete (each step is executable, e.g., declaring Docker container image and command line arguments). Executable CWL can be run on a range of workflow engines[14], which again support many different compute backends on cloud and local clusters.
- Galaxy workflows and their entries on WorkflowHub can be annotated with EDAM ontologies and bioschemas
- Galaxy workflows can be converted to "Abstract CWL" for documentation purposes, though they are still executed using their native language on a native Galaxy instance.
- RO-Crate is used to deposit Galaxy workflows in the WorkflowHub registry from resources such as the Galaxy Intergalactic Workflow Commission using the Workflow RO-Crate profile.
- Galaxy has added RO-Crate support to preserve a workflow and its execution history (provenance) using the Workflow Run-RO-Crate profile [De Geest et al, 2022].
- Galaxy and WorkflowHub support the GA4GH TRS API; consequently, Galaxy workflows can be executed on a public Galaxy instance directly from WorkflowHub and a Galaxy instance can directly find workflows in WorkflowHub. Other Galaxy installations such as the SDR instance can also retrieve workflows from WorkflowHub using this API.
- Other facilities available to SDR from the Collaboratory (but not explored in this pilot) include workflow testing and benchmarking. *LifeMonitor* monitors and triggers automated workflow tests and automated checks on metadata and adherence to best practices on the workflow's source code Git repository, with dedicated support for Galaxy using the Planemo testing framework. The *OpenEBench* service benchmarks tools, and monitors software quality as well as scientific benchmarking to help determine the precision, recall and other metrics of bioinformatics resources in unbiased scenarios.

The Workflow Collaboratory supports multiple workflow platforms, however additional support has been developed for the *Galaxy workflow platform and execution instances*:

[11] https://www.researchobject.org/ro-crate
[12] https://www.ga4gh.org/news_item/tool-registry-service-api-enabling-an-interoperable-library-of-genomics-analysis-tools
[13] https://www.commonwl.org
[14] https://www.commonwl.org/implementations

- Tools installed in a Galaxy server, and Galaxy workflow entries on WorkflowHub can be annotated with tool identifiers to link through to bio.tools entries[15].
- Galaxy workflows and their entries on WorkflowHub can be annotated with EDAM ontologies and bioschemas
- Galaxy workflows can be converted to "Abstract CWL" for documentation purposes, though they are still executed using their native language on a native Galaxy instance.
- RO-Crate is used to deposit Galaxy workflows in the WorkflowHub registry from resources such as the Galaxy Intergalactic Workflow Commission[16] using the Workflow RO-Crate profile[17].
- Galaxy has added RO-Crate support[18] to preserve a workflow and its execution history (provenance) using the Workflow Run-RO-Crate profile[19] [De Geest et al, 2022].
- Galaxy and WorkflowHub support the GA4GH TRS API; consequently, Galaxy workflows can be executed on a public Galaxy instance directly from WorkflowHub and a Galaxy instance can directly find workflows in WorkflowHub[20]. Other Galaxy installations such as the SDR instance can also retrieve workflows from WorkflowHub using this API.

Other facilities available to SDR from the Collaboratory (but not explored in this pilot) include workflow testing and benchmarking. *LifeMonitor*[21] monitors and triggers automated workflow tests and automated checks on metadata and adherence to best practices on the workflow's source code Git repository, with dedicated support for Galaxy using the Planemo testing framework[22]. The *OpenEBench*[23] service benchmarks tools, and monitors software quality as well as scientific benchmarking to help determine the precision, recall and other metrics of bioinformatics resources in unbiased scenarios.

2.2 Deployment

When we started SDR development we considered a centralised instance, but it became clear in the project that different partners desired to install their own instances of SDR. This for instance allows installation of additional or customised tools or using private compute resources and data storage. Future work in the EuroScienceGateway project, as previously mentioned, will reduce the difference between such private instances and public instances through the common Pulsar network; these advantages could later be of consideration for SDR as well.

For reproducible deployment of Galaxy and the SDR we chose to use Ansible. Ansible is a system administrator tool that can automate installation of software and configuration on servers, using a declarative configuration in the form of *playbooks* that specifies the desired state, e.g., presence of certain software packages and services. A large library of such playbooks are available[24], including for base installation of Galaxy[25].

---

[15] https://workflowhub.eu/workflows/518
[16] https://gallantries.github.io/video-library/modules/ro-crate
[17] https://www.researchobject.org/ro-crate/1.1/workflows.html
[18] https://galaxyproject.org/news/2023-02-23-structured-data-exports-ro-bco
[19] https://www.researchobject.org/workflow-run-crate
[20] https://usegalaxy-eu.github.io/posts/2021/03/25/wfh-video
[21] https://lifemonitor.eu
[22] https://planemo.readthedocs.io/en/stable/best_practices_workflows.html
[23] https://openebench.bsc.es
[24] https://galaxy.ansible.com – note that Ansible Galaxy is not related to the Galaxy workflow system.
[25] https://training.galaxyproject.org/training-material/topics/admin/tutorials/ansible-galaxy/tutorial.html

In addition to deploying Galaxy itself, deploying SDR also means adding the SDR tools that will be invoked by the workflow (D8.2) along with their supporting configuration. In addition, we found it useful for Ansible to further configure the Galaxy server, for instance adding users and changing the landing page, as well as pre-installing the SDR workflows[26]. For this we extended Galaxy's Ansible Roles with additional recipes gathered in a SDR playbook[27]. In addition to the base install of Galaxy, the SDR playbook installs the SDR GitHub repository[28] onto the Galaxy server (Figure 2).

For installation of the SDR tools we rely on Docker[29] containers, mainly provided by Teklia[30]. Docker containers are ways to distribute and execute an installed software stack without the complexity of virtual machines. An advantage of using Docker containers here that we found was that it hides any incompatibility with other software on the system, for instance both Galaxy and some of the tools are implemented using Python, but with different versions of Python and Python libraries. The use of containers also ensures that different deployments of SDR get the same SDR tool infrastructure independent of its base OS distribution, which may vary across sites (e.g., CentOS vs Debian), and that these tools can be coherently updated.

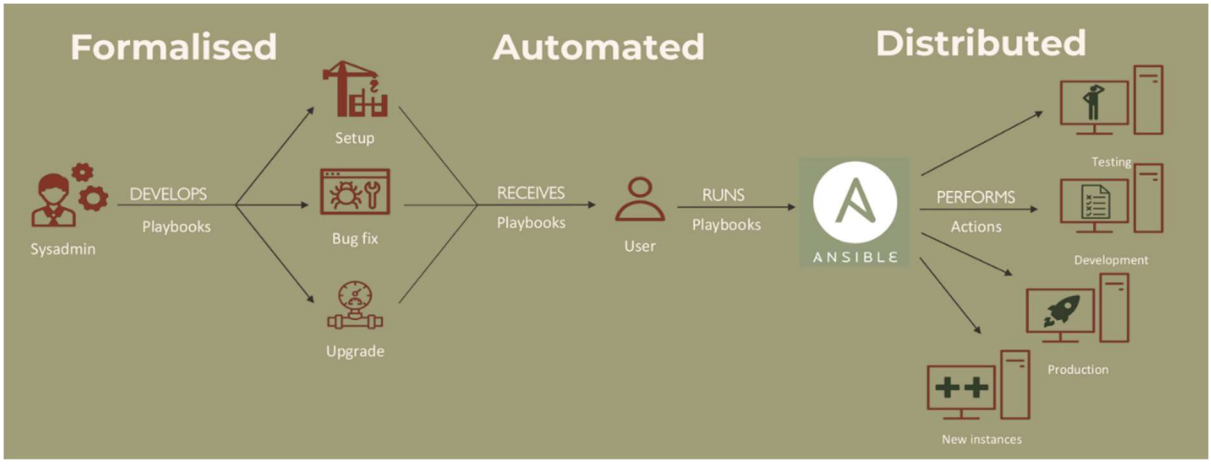

Figure 2: *Lifecycle of Ansible notebooks and their deployment on multiple instances. Extracted from [Woolland et al, 2022] under a CC-BY 4.0 licence.*

The result of running the SDR Ansible playbook then is a fully configured Galaxy instance with the SDR tools and workflows enabled. In addition, general Galaxy tools (e.g., CSV import) are enabled; system administrators can enable additional tools from Galaxy's extensive toolshed[31] (e.g., for genomics), or user-provided[32], by modifying the Ansible configuration.

The Ansible approach is equally applicable to physical servers, virtual machines, and cloud deployments. In our testing we have used both the Microsoft Azure cloud and VMware VMs.

3. Overview

[26] https://workflowhub.eu/projects/72#workflows
[27] https://github.com/DiSSCo/SDR/blob/main/ansible/deploy-sdr.yml
[28] https://github.com/DiSSCo/SDR
[29] https://www.docker.com/
[30] https://hub.docker.com/u/teklia
[31] https://toolshed.g2.bx.psu.edu/
[32] https://github.com/DiSSCo/SDR/blob/main/docs/how-to/add-new-tool.md

3.1 Using Galaxy for SDR workflows

Galaxy has an extensive set of documentation and training materials[33]. The Galaxy web interface is mostly intuitive, but it can take some time for researchers to get used to *workflow thinking* [Crusoe et al, 2023]. We produced a demonstration video[34] to show how the SDR tools are used in Galaxy, in addition to several DiSSCo webinars and external presentations [Livermore et al, 2022; Woolland et al, 2022].

A typical SDR workflow (Figure 3) begins with creating a skeleton openDS object [openDS]. This takes some initial metadata which can be provided as a CSV file (see example[35]) that includes:

- **Catalog number**, as indexed by the specimen owner's collection
- **Image license**, e.g., to indicate Creative Commons licence
- **Image URI**, typically to a JPEG file
- **Object type**, e.g., "Pinned insect"
- **Rights holder**, typically the organisation holding the specimen
- **Institution URL**, an identifier for the specimen owner
- **Higher classification** of the specimen.
- **Person name**

Person identifier, e.g., https://orcid.org/0000-0001-9842-9718

Figure 4: *Galaxy view of openDS JSON, showing the initial specimen metadata, extracted image information and detected regions.*

The final openDS represents the completed digital specimen. Longer term this will then be deposited in the DiSSCo infrastructure as a FAIR Digital Object. However, it is likely during workflow execution

[33] https://training.galaxyproject.org/
[34] https://www.youtube.com/watch?v=Nryz7qmyQpo
[35] https://github.com/DiSSCo/SDR/blob/main/galaxy-workflow/samples/single.csv

that some workflow parameters may need to be tweaked after manual inspection of the SDR outputs.

Figure 5: *Galaxy view of detected specimen and text labels. On the right are shown the SDR tool executions of the workflow, which can then be individually re-executed if modifying parameters. Adapted from https://www.youtube.com/watch?v=Nryz7qmyQpo*

Galaxy provides powerful visualisation mechanisms for many file formats, and we added Galaxy integration to show the detected text regions overlaid on the specimen photo (Figure 5).

However, a visualisation tool of openDS JSON structure did not exist, so we created an additional Galaxy tool that converts the openDS to a flattened CSV, which can be inspected in regular spreadsheet applications or even within Galaxy. Although this representation is not ideal for digital specimens (e.g., a single specimen row will show multiple text regions and transcribed texts, but it may be hard to match these up), the DiSSCo users found this as a useful intermediary tool as they are not all familiar with JSON.

As openDS is specified using JSON Schema, we also added a validation step to ensure the output JSON is conforming to the schema's requirement. Facing co-development of openDS and SDR, the supported version of the openDS schema is part of the Ansible installation of SDR.

3.2 Bulk operations

A core motivation for using a workflow system in SDR was to be able to scale up digitization. Galaxy supports this with a minor adjustment to the workflow that splits the initial CSV file into multiple data entries for each row. The rest of the workflow will then produce a collection of openDS objects, each processed separately and (if configured) in parallel.

As the final openDS includes all the information from the previous stages, the approach developed by SDR has a major advantage compared to traditional Galaxy workflows, which would typically use a series of tools with various intermediate data outputs which would then need to be "lined up" (e.g., by index) for correlation to the corresponding input (e.g., specimen).

We performed performance testing with a moderate number of specimens (200) which revealed minor errors on some SDR tools for some of the specimens, e.g., inability to detect text regions which then prevents subsequent SDR tools from executing. Galaxy supports such partial failures and completes the rest of the workflow, but the errors can cause further problems when merging back items in bulk collections e.g., to a single CSV output. Further work here could be to propagate the partial openDS with an embedding of error messages, rather than the tool itself failing workflow-wise.

As Galaxy provides an API for executing workflows, such bulk operations can then be triggered by the DiSSCo core infrastructure.

3.3 Provenance

Execution of SDR workflows don't necessarily lead to digital specimen registrations, pending researcher validation for instance. To allow SDR workflow execution to remain flexible and be applicable both for machines and users, we did not add openDS publishing as part of the workflow, but rather this would be done by the DiSSCo Annotation Processing Services after calling the Galaxy APIs.

However, in both cases we found it important to include not just the openDS, but also details of its generation, the workflow and image references. For this, we utilise RO-Crate [Soiland-Reyes et al, 2022a] and developed an additional step to the workflow that generates the crate based on the openDS objects. This final output can then be used to make a single output for bulk operations of multiple digital specimens, as an RO-Crate is commonly transmitted as a ZIP archive with embedded metadata, which here contains each of the openDS objects.

Concurrently with SDR's development, in collaboration with Horizon Europe projects BY-COVID and EuroScienceGateway, Galaxy developed official support for exporting workflow provenance[36] as an RO-Crate [De Geest et al, 2023]. This follows the newly developed Workflow Run Crate[37] profile which details the provenance of the workflow execution and its file outputs, as well as individual step executions. The workflow definition is also embedded in the crate both as Galaxy and abstract CWL. UNIMAN has had a key role in the development of this profile, incorporating experiences from the SDR, and this is now seen as a workflow-system-independent successor to CWLProv [Zaib Khan et al, 2023], with Workflow Run Crate already implemented by at least 6 workflow systems.

3.4 Administration

We ran four instances:

1. Local instance for lead UNIMAN developer
2. UNIMAN development instance (hosted virtual machine, firewalled)
3. NHM development instance (hosted virtual machine, firewalled)
4. NHM production (hosted virtual machine) for user tested and public facing

[36] https://galaxyproject.org/news/2023-02-23-structured-data-exports-ro-bco/
[37] https://www.researchobject.org/workflow-run-crate/

At the end of the project, we will keep the NHM production instance running for up to a year.

In addition to the four instances, we ran during the SYNTHESYS+ project, the development team at Naturalis also ran a test instance which verified the Ansible approach.

All these instances used our Ansible deployment approach, although initial prototyping was done with manually installed Galaxy.

3.3 Descoped Features

Authentication and Authorisation Infrastructure (AAI)

Work on AAI for broader DiSSCo services was explored and developed by other partners in SYNTHESYS+. The approach will use GRNET's AAI infrastructure implemented across DiSSCo services using Keycloak. It should be noted that the EOSC-Life Collaboratory (including WorkflowHub, Galaxy etc) uses the LS-Login AAI system[38], which supports multiple authentication sources, including academic institutions and ORCID.

Ledger

Being implemented and tested outside of the SDR work package as core DiSSCo infrastructure for managing and tracking annotations across specimens using the openDS standard.

Common Workflow Language

Our earlier landscape analysis [Walton et al, 2020a] put stronger emphasis on Common Workflow Language [Crusoe et al, 2022] to specify workflows and execute them on multiple platforms. In early phases of SDR development we used development builds of Galaxy that included support for executing tools described in CWL. In this scenario, the individual SDR tools were also composable in other workflow systems that support CWL. In this co-evolution with CWL and Galaxy developers, we helped at several ELIXIR Biohackathons[39], yet ultimately SDR was progressing ahead of the projection for when CWL support would be included in official Galaxy releases.

With the desire for stable Galaxy deployment on multiple sites (See section 2.2), as well as customizable input parameters to better inform on their usage, we transitioned to using native Galaxy tool configurations and wrappers, which were more easily integrated with Ansible deployment of regular Galaxy releases. It is worth noting that, with the majority of the SDR tools exchanging openDS objects, it is relatively trivial to map each of them in both CWL[40] and Galaxy[41].

Partial FDOs

In [Hardisty et al, 2022] we proposed a strong integration with openDS, FAIR Digital Object (FDO) and SDR. FAIR Digital Objects [Ivonne 2022; De Smedt et al, 2020] is an emerging concept for publishing data in a structured machine-actionable form, with strong emphasis on persistent identifiers, types, validation and resolution. There are multiple possible implementations of FDO, several of which are being tested by the biodiversity community [Islam et al, 2022; Plale 2022].

[38] https://lifescience-ri.eu/ls-login/
[39] https://elixir-europe.org/events/biohackathon-europe
[40] https://www.commonwl.org/user_guide/topics/command-line-tool.html
[41] https://github.com/DiSSCo/SDR/blob/main/docs/how-to/add-new-tool.md

Our strategy for SDR was for a dual approach of traditional Handle-based FDOs for openDS publication, along with a Web-based RO-Crate FDOs [Soiland-Reyes et al, 2022c; Soiland-Reyes et al, 2023a] of the corresponding workflow execution history.

As assignment of persistent identifiers for FDOs are typically strongly linked with the repository it will be stored in, we found SDR workflows would get a dependency on DiSSCO infrastructure which was still under development. The openDS schemas (the FDO type) was also under development at the same time as SDR, however the workflows needed to exchange openDS objects even if these schemas were not yet available as FDO types. We also took advantage of being able to incrementally build openDS objects, as detailed in section 3.1; clearly there would not be much purpose in depositing an FDO remotely which a second later would be augmented with additional information. This is the main reason why SDR decided on a *partial FDO* approach which is neutral as to where the openDS digital specimens are to be published (or not), yet contain all the metadata in order to make such FAIR publication possible.

Likewise we decided to be neutral from Galaxy as to what would be the final destination of the RO-Crate of the workflow execution provenance (section 3.3). The Galaxy instance itself is a suitable home until the workflow run could be *promoted* by DiSSCO infrastructure (e.g. because its openDS outputs become registered as FDOs), in which case the RO-Crate would be stored as an additional FDO in the same storage infrastructure or in external repositories like Zenodo. This however adds a complication in that the initial RO-Crate output can only reference the openDS by value (files in the ZIP), as it cannot know in advance which PIDs they will be assigned.

By using FAIR Signposting a lightweight approach to FDO can be achieved [Soiland-Reyes 2022b] with minimal changes to the Web architecture. This is how Workflow RO-Crates are resolvable from DOI handles on WorkflowHub [Goble & Soiland-Reyes, 2023; Soiland-Reyes et al, 2022c], e.g. https://doi.org/10.48546/workflowhub.workflow.375.1 – future work would be to also add such signposting to Galaxy itself for workflow runs. This will be partially addressed by the EuroScienceGateway project which is strengthening the FDO and RO-Crate support in Galaxy. The FAIR-IMPACT project is also addressing the link between Signposting and RO-Crate, with funded support given[42] to implementation in institutional repository software like DataVerse.

4. Dissemination of Results

We have included a high-level summary of presentations and publications associated with Task 8.3. Much of the dissemination work done for Task 8.3 includes work done as part of Task 8.2 ”Tools and Services for Extracting, Enhancing and Annotating Natural History Specimen Data" in [Livermore et al, 2023a]. A comprehensive list of dissemination materials is given in the report by [Livermore et al, 2023b].

Five SDR workflows are registered in WorkflowHub Team Specimen Data Refinery[43] in the WorkflowHub DiSSCo Space[44] (Figures 6 and 7). The SDR bio.tools collection[45] aims to collect the tools used in SDR.

[42] https://fair-impact.eu/1st-open-call-support-closed
[43] https://workflowhub.eu/projects/72
[44] https://workflowhub.eu/programmes/15
[45] https://bio.tools/t?collectionID=%22Specimen%20data%20refinery%22

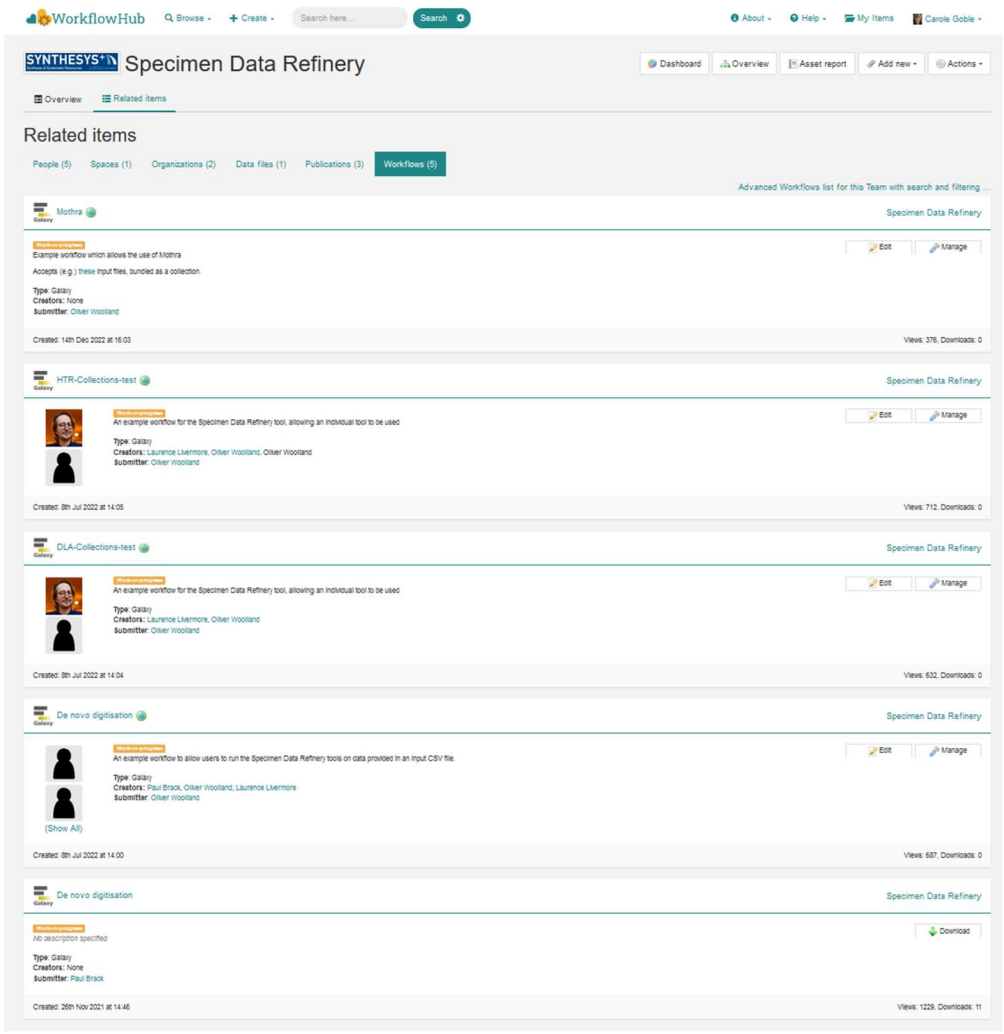

Figure 6: *SDR Team workflows registered on WorkflowHub.*

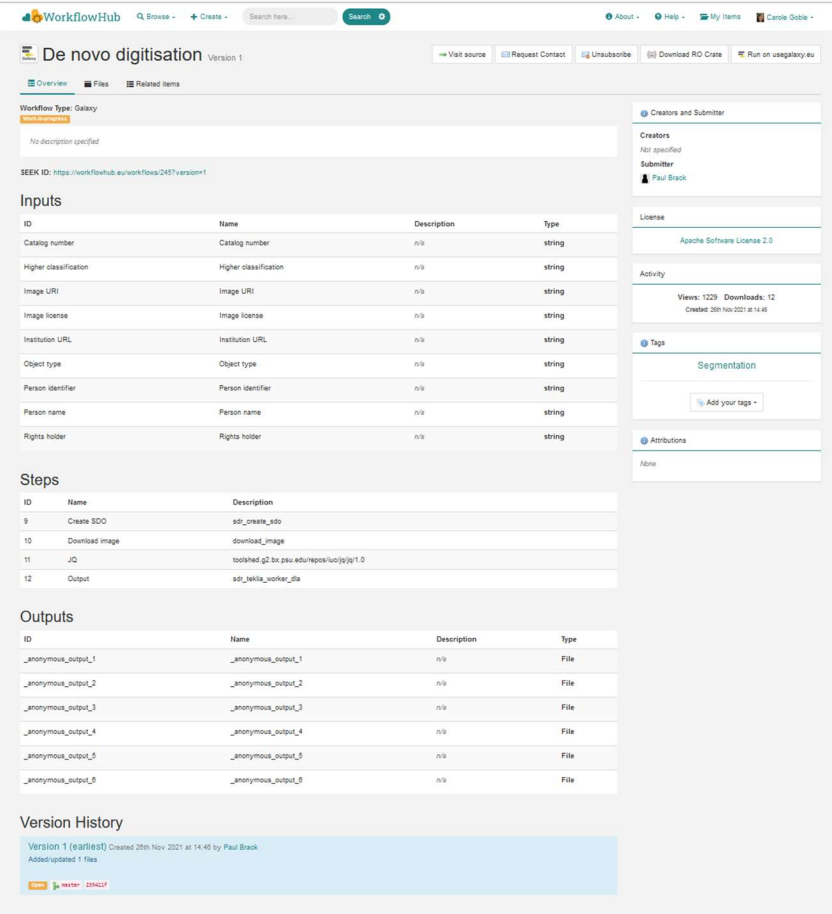

Figure 7: *SDR De Novo digitisation workflow registered entry WorkflowHub*

5. Discussion and Future Development

5.1 Incremental building of digital specimens using workflows

As pointed out in [Woolland et al, 2022], a possible disadvantage of the incremental openDS approach is that it can make debugging more difficult for workflow developers. For instance, the text extraction tool cannot be tested without first preparing a partial openDS with the text regions. As an alternative, we modified some SDR Galaxy tools to also permit “flat” input parameters, e.g., directly providing the specimen image. Such changes allow “dual use” of the SDR tools outside an openDS ecosystem.

Likewise, if users want to utilise existing Galaxy tools, they will need *shim* steps to “decompose” the required parts from the openDS, as these typically expect direct file inputs or string values. We found the openDS-to-CSV output tool could be utilised for this purpose; future work could add more custom, schema-driven user interface to simplify such openDS extraction.

Users will need to combine SDR tools in a particular order so that the openDS prerequisites of that tool are fulfilled. While Galaxy is type-aware and typically restricts possible inputs to only show compatible (possibly convertible) inputs from previous steps, with most values in SDR workflows being openDS objects it may inadvertently indicate all of them are possible inputs. Future work would consider marking sub-*profiles* of openDS, e.g., *openDS-with-region*. Some SDR operations also

only make sense for certain specimen types, so all aspects of the openDS object should be taken into consideration.

As pointed out in Section 3.2, by moving slightly from the traditional *workflow thinking* in Galaxy of processing files and values with intermediate, to passing monolithic openDS objects, then error handling can become more cumbersome and may have to be handled more by openDS-aware tools than the workflow engine. Error handling will have to be revisited for larger scale specimen digitization.

One of the key advantages of using a workflow to extract data from specimen records in incremental steps is that it allows users to customise the workflow and select different data extraction components. However, as evidenced in prototype user testing, users wanted out-of-the-box workflow for end-to-end extraction of specimen data. Building a chained pipeline of tools is complex and requires understanding each tool and its role in a machine-learning pipeline. For example, semantic segmentation to detect labels is a prerequisite of all downstream tasks. Developing machine learning models as atomic tasks has significant development and computational impact. All tasks' data inputs and outputs must be aligned; each task writes output to disk requiring high IO; for Dockerised tools, each task initiates a standalone image. Without the requirement for GUI customisation, an API service-based approach offers a more straightforward mechanism for interfacing with specimen data extraction mechanisms. Internally, a workflow model could still be used, to track object provenance through the process.

Future maturing of SDR towards TRL-8 will of course depend on other integration requirements within DiSSCo, as well as the maturing of the tools (D8.2) and the SDR workflows, but we should be assured by successful large-scale installations of Galaxy such as the EOSC service UseGalaxy.eu (TRL-9) having more than 50.000 users across Europe.

One challenge we found when using Ansible approach, which perhaps is comparable to any software release mechanism, is that supporting concurrent development of SDR tools is more formal than in early prototype stages, for instance requiring the tool's corresponding Docker image to have been built and published. At the same time, Ansible allowed such upgrades (and any changes to workflows) to be reliably tested on separate SDR instances.

5.2 Relying on an established workflow system

Galaxy is an active open-source project, and several releases came out during this project with desirable improvements, for instance CWL and RO-Crate support. However, by relying on community contributions and collaboration, such features did not necessarily come at the time that would have suited SDR development.

For the future it would be beneficial to further embed into the open development processes of the underlying software platforms, and directly contribute desired features and documentation. Recently we have for instance contributed RO-Crate training to the Galaxy Training Network[46].

5.3 Stable deployment using Ansible

The ability for SDR to keep up with Galaxy releases was another reason for using the Ansible deployment approach, simplifying upgrade testing separate from production instances. In this

[46] https://training.galaxyproject.org/training-material/topics/fair/tutorials/ro-crate-intro/tutorial.html

project we moved from developing a prototype of SDR (TRL-3 to TRL-4) to a production-ready demonstrator tested at multiple sites (TRL-6 to TRL-7).

Future maturing of SDR towards TRL-8 will of course depend on other integration requirements within DiSSCO, as well as the maturing of the tools (D8.2) and the SDR workflows, but we should be assured by successful large-scale installations of Galaxy such as the EOSC service UseGalaxy.eu (TRL-9) having more than 50,000 users across Europe[47].

One challenge we found when using Ansible approach, which perhaps is comparable to any software release mechanism, is that supporting concurrent development of SDR tools is more formal than in early prototype stages, for instance requiring the tool's corresponding Docker image to have been built and published. At the same time, Ansible allowed such upgrades (and any changes to workflows) to be reliably tested on separate SDR instances.

5.4 Integrate with FDOs

The openDS and RO-Crates returned from SDR are *partial FDOs*, they need to be integrated into the DiSSCo infrastructure for publishing digital specimens.

As this infrastructure matures, as well as the FDO specifications [Ivonne 2022], it will become equally important to consider FDOs as starting points for SDR workflows, e.g., re-analysing a previous openDS with an improved digitization workflow, or for mass processing of newly imaged specimens published as FDOs.

In this case the provenance of the output FDO as well will need to be propagated – this would grow the importance of tracking the workflow runs as RO-Crate and cross-relate them and with other FDOs.

5.5 Utilising the Workflow Collaboratory fully, more openness

While the SDR currently have workflows published in WorkflowHub, it does not fully utilise the Workflow Collaboratory as mentioned in section 2.1. Part of the reason for this is that the SDR tools (D8.2), while under development and testing, were not eligible to be published in the official Galaxy Toolshed. Their licensing also had to be clarified for inclusion in the Ansible installation; in practice some of the tools, while open source, needed an encryption key for a trained machine learning model which was closed source.

For SDR workflows published in WorkflowHub, there is therefore (at time of writing) a requirement to execute them on a Galaxy server which has the SDR tools pre-installed, e.g., using the Ansible playbook. To take full advantage of the Workflow Collaboratory, e.g., automatic testing and integration with public Galaxy instances, the SDR tool suite (or a subset) could be further curated so they can be promoted to the Galaxy Toolshed. Combined with a "dual use" approach (section 5.1) this would open for specimen data refinery workflows to be developed beyond the purposes envisioned by this project.

Resurrecting the CWL integration of the SDR tools does not require any Galaxy Toolshed registration (just public Docker images) and could likewise be a route for SDR tools to be used by additional communities that use other workflow systems than Galaxy.

[47] https://usegalaxy-eu.github.io/posts/2022/06/23/reached-50000-users/

6. Code Repository & Related Issues

GitHub repository for overall SDR project: https://github.com/DiSSCo/SDR
Workflows: https://workflowhub.eu/projects/72#workflows
Summary of development work that contributed to Deliverable 8.3:
https://github.com/DiSSCo/SDR/issues/78

7. Acknowledgements

We acknowledge the SYNTHESYS+ and DiSSCo project members who have been invaluable in early evaluation and feedback on the development of SDR. We particularly want to thank Paul Brack who prototyped the SDR in Galaxy and gave us many ideas for further development.

In addition to SYNTHESYS Plus (823827), this work has been developed in collaboration with effort in EU projects: DiSSCo Prepare (871043), BioExcel-2 (823830), EOSC-Life (824087), BY-COVID (101046203), EuroScienceGateway (101057388; UKRI 10038963).

8. References

[Clark-Casey 2022] Justin Clark-Casey, Stian Soiland-Reyes (2022):
Making EOSC Research Objects FAIR with RO-Crate: A common metadata overlay for EOSC repositories.
EOSC Symposium 2022
https://doi.org/10.5281/zenodo.7323480

[Crusoe et al, 2022] Michael R. Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, Stian Soiland-Reyes, Bogdan Gavrilović, Carole Goble, The CWL Community (2022):
Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language.
Communications of the ACM **65**(6)
https://doi.org/10.1145/3486897

[De Geest et al, 2022] Paul De Geest, Frederik Coppens, Stian Soiland-Reyes, Ignacio Eguinoa, Simone Leo (2022):
Enhancing RDM in Galaxy by integrating RO-Crate.
1st International Conference on FAIR Digital Objects (FDO 2022) (poster)
Research Ideas and Outcomes **8**:e95164
https://doi.org/10.3897/rio.8.e95164

[De Smedt et al, 2020] Koenraad De Smedt, Dimitris Koureas, Peter Wittenburg (2020):
FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units.
Publications **8**(2):21
https://doi.org/10.3390/publications8020021

[Galaxy Community, 2022] The Galaxy Community (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, https://doi.org/10.1093/nar/gkac247

[Goble et al, 2021] Carole Goble, Stian Soiland-Reyes, Finn Bacall, Stuart Owen, Alan Williams, Ignacio Eguinoa, Bert Droesbeke, Simone Leo, Luca Pireddu, Laura Rodriguez-Navas, José Mª Fernández, Salvador Capella-Gutierrez, Hervé Ménager, Björn Grüning, Beatriz Serrano-Solano, Philip Ewels, Frederik Coppens (2021):
Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory.
Zenodo
https://doi.org/10.5281/zenodo.4605654

[Goble & Soiland-Reyes, 2023] Carole Goble, Stian Soiland-Reyes (2023):
Sharing research artefacts as FAIR Digital Objects using RO-Crate.
Brookhaven National Laboratory, 2023-01-23.
https://doi.org/10.5281/zenodo.7559338

[Gray et al, 2023] Alasdair Gray, Leyla J. Castro, Nick Juty, Carole Goble (2023):
Schema.org for Scientific Data.
Artificial Intelligence for Science (Chapter 27) pp 495-51.
https://doi.org/10.1142/9789811265679_0027 [no OA preprint available]

[Hardisty et al, 2022] Alex Hardisty, Paul Brack, Carole Goble, Laurence Livermore, Ben Scott, Quentin Groom, Stuart Owen, Stian Soiland-Reyes (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 2022; 4 (2): 320–341. https://doi.org/10.1162/dint_a_00134

[Ison et al, 2013] Jon Ison, Matúš Kalaš, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer, Peter Rice (2013):
EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats.
Bioinformatics **29**(10) pp 1325–1332.
https://doi.org/10.1093/bioinformatics/btt113

[Islam et al, 2022] Islam S, Weber A, Tóth-Czifra E (2022):
From Green Deal to Cultural Heritage: FAIR Digital Objects and European Common Data Spaces .
Research Ideas and Outcomes **8**:e93815.
https://doi.org/10.3897/rio.8.e93815

[Ison et al, 2019] Jon Ison, Hans Ienasescu, Piotr Chmura, Emil Rydza, Hervé Ménager, Matúš Kalaš, Veit Schwämmle, Björn Grüning, Niall Beard, Rodrigo Lopez, Severine Duvaud, Heinz Stockinger, Bengt Persson, Radka Svobodová Vařeková, Tomáš Raček, Jiří Vondrášek, Hedi Peterson, Ahto Salumets, Inge Jonassen, Rob Hooft, Tommi Nyrönen, Alfonso Valencia, Salvador Capella, Josep Gelpí, Federico Zambelli, Babis Savakis, Brane Leskošek, Kristoffer Rapacki, Christophe Blanchet, Rafael Jimenez, Arlindo Oliveira, Gert Vriend, Olivier Collin, Jacques van Helden, Peter Løngreen, Søren Brunak (2019):
The *bio.tools* registry of software tools and data resources for the life sciences.
Genome Biology **20**(1):164
https://doi.org/10.1186/s13059-019-1772-6

[Ivonne et al, 2023] Anders Ivonne, Christophe Blanchi, Daan Broder, Maggie Hellström, Sharif Islam, Thomas Jejkal, Larry Lannom Larry, Karsten Peters-von Gehlen, Robert Quick, Alexander Schlemmer, Ulrich Schwardmann, Stian Soiland-Reyes, George Strawn, Dieter van Uytvanck, Claus Weiland, Peter Wittenburg, Carlo Zwölf (2023):
FAIR digital object technical overview. Version PEN 2.0.

FDO Specification Documents Full FDO Overview PEN-2.0-v2
FAIR Digital Objects Forum
https://doi.org/10.5281/zenodo.7824714

[Leo et al, 2023] Simone Leo, Laura Rodríguez-Navas, José M. Fernández, Paul De Geest, Luca Pireddu, Michael R. Crusoe, Daniel Garijo, Iacopo Colonnelli, Raül Sirvent, Stian Soiland-Reyes (2023):
Making workflow provenance FAIR across workflow systems with Workflow Run RO-Crate.
ELIXIR All Hands meeting 2023 (poster), Dublin, Ireland, 2023-06-05/–08
https://doi.org/10.5281/zenodo.8004793

[Livermore et al, 2022] Laurence Livermore, Paul Brack, Ben Scott, Stian Soiland-Reyes, Oliver Woolland (2022):
The Specimen Data Refinery: Using a scientific workflow approach for information extraction
Biodiversity Information Standards (TDWG 2022).
Biodiversity Information Science and Standards **6**:e93500
https://doi.org/10.3897/biss.6.93500
Slides: https://doi.org/10.6084/m9.figshare.21312345.v1

[Livermore et al, 2023a] Livermore, L., Blettery, J., Cubey, R., Goble, C., Hardy, H., Haston, E., Kermovant, C., Lasseck, M., Obst, M., Plank, A., Scott, B., Soiland-Reyes, S., Woolland, O., and Wu, Z. (2023)
Deliverable 8.2 - Specimen Data Refinery: Tools and services for extracting, enhancing and annotating natural history specimen data. August 2023.

[Livermore et al, 2023b] Livermore, L., Banki, O., Cubey, R., Goble, C., Hardy, H., Lasseck, M., Leeflang, S., Scott, B., Soiland-Reyes, S., Woolland, O. (2023)
Deliverable 8.4 - Specimen Data Refinery: Usage - data exploitation, evaluation and dissemination. August 2023.

[openDS] openDS: Draft specification for open Digital Specimens (openDS). Available at: https://github.com/DiSSCo/openDS . Accessed 10 August 2021

[Plale 2022] Plale, Beth (2022): **Achieving low barriers to entry in the FAIR Digital Objects (FDO) data space: a Use Case in Biodiversity Extended Specimen Networks**
The Data To Insight Center
https://hdl.handle.net/2022/27837

[Royaux et al, 2022] Royaux C, Arnaud E, Sananikone J, Jossé M, Madelin M, Pelletier D, Norvez O, Le Bras Y (2022) Open Science for Better FAIRness: A biodiversity virtual research environment point of view. Biodiversity Information Science and Standards 6: e95110.
https://doi.org/10.3897/biss.6.95110

[Soiland-Reyes et al, 2022a] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022):
Packaging research artefacts with RO-Crate.
Data Science 5(2)
https://doi.org/10.3233/DS-210053

[Soiland-Reyes 2022b] Stian Soiland-Reyes, Leyla Jael Castro, Daniel Garijo, Marc Portier, Carole Goble, Paul Groth (2022):
Updating Linked Data practices for FAIR Digital Object principles.
1st International Conference on FAIR Digital Objects (FDO 2022) (presentation).
Research Ideas and Outcomes 8:e94501
https://doi.org/10.3897/rio.8.e94501

[Soiland-Reyes et al, 2022c] Stian Soiland-Reyes, Peter Sefton, Leyla Jael Castro, Frederik Coppens, Daniel Garijo, Simone Leo, Marc Portier, Paul Groth (2022):
Creating lightweight FAIR Digital Objects with RO-Crate.
1st International Conference on FAIR Digital Objects (FDO 2022) (poster)
Research Ideas and Outcomes **8**:e93937
https://doi.org/10.3897/rio.8.e93937

[Soiland-Reyes 2023a] Stian Soiland-Reyes, Carole Goble (2023):
Building diverse FDO Collections using RO-Crate.
FAIR Digital Object Forum workshop "Defining FDO Collections", 2023-04-14.
https://doi.org/10.5281/zenodo.7828632

[Soiland-Reyes 2023b] Stian Soiland-Reyes, Carole Goble, Paul Groth (2023):
Evaluating FAIR Digital Object and Linked Data as distributed object systems.
arXiv 2306.07436 [cs.DC]
https://doi.org/10.48550/arXiv.2306.07436

[Soiland-Reyes 2023c] Stian Soiland-Reyes, Leyla Jael Garcia (2023):
Overview of FAIR data publishing with Bioschemas & RO-Crate.
ELIXIR All Hands meeting 2023, workshop "Building lightweight FAIR data packages with Bioschemas and RO-Crate", Dublin, Ireland, 2023-06-05/–08
https://doi.org/10.7490/f1000research.1119459.1

[Walton et al, 2020a] Walton S, Livermore L, Bánki O, Cubey RWN, Drinkwater R, Englund M, Goble C, Groom Q, Kermorvant C, Rey I, Santos CM, Scott B, Williams AR, Wu Z (2020) **Landscape Analysis for the Specimen Data Refinery**. Research Ideas and Outcomes 6: e57602.
https://doi.org/10.3897/rio.6.e57602

[Woolland et al, 2022] Oliver Woolland, Paul Brack, Stian Soiland-Reyes, Ben Scott, Laurence Livermore (2022):
Incrementally building FAIR Digital Objects with Specimen Data Refinery workflows.
1st International Conference on FAIR Digital Objects (FDO 2022) (poster)
Research Ideas and Outcomes 8: e94349
https://doi.org/10.3897/rio.8.e94349
Poster: https://doi.org/10.5281/zenodo.7233688

[Zaib Khan et al, 2023] Farah Zaib Khan, Stian Soiland-Reyes, Richard O. Sinnott, Andrew Lonie, Carole Goble, Michael R. Crusoe (2019):
Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv.
GigaScience **8**(11)
https://doi.org/10.1093/gigascience/giz095