Project:	Synthesis of systematic resources
Project acronym:	SYNTHESYS PLUS
Grant Agreement number:	823827
Work Package:	Workpackage 8
Deliverable number:	8.4
Deliverable title:	Development of cloud platform for data-processing services
Deliverable author(s):	Laurence Livermore - Natural History Museum, London Olaf Banki – Naturalis Biodiversity Centre, Leiden Robert Cubey - Royal Botanic Garden Edinburgh Carole Goble - The University of Manchester Helen Hardy - Natural History Museum, London Mario Lasseck - Museum für Naturkunde, Berlin Sam Leeflang - Naturalis Biodiversity Center, Leiden Ben Scott - Natural History Museum, London Stian Soiland-Reyes - The University of Manchester Oliver Woolland - The University of Manchester
Date:	2023-08-16

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) workpackage (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], the tool and service development report (Task 8.2/Deliverable 8.2), and the report on the development of the cloud platform (Task/Deliverable 8.3).

1.1 Scope

This report primarily describes how the SDR may integrate with the wider DiSSCo architecture; approaches to documentation and evaluation; and the activities undertaken to disseminate and promote the SDR.

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].

Details of Galaxy platform modification and deployment are summarised in [Livermore et al, 2023b] (the report for Task 8.3 “Development of cloud platform for data-processing services”).

2. Data exploitation and future integration

The SDR has the potential to generate vast quantities of structured information, associated with digitised specimen images and data - appropriate integrations will be critical to allow this information to be fully exploited by the wider bio/geoscience community. Any SDR platform(s) and tools must provide effective delivery of data integrated with DiSSCo services. These data will also be relevant to third party services such as Catalogue of Life+, as well as institutional collection management systems.

2.1 Integration of SDR with DiSSCo core infrastructure

What differentiates DiSSCo as infrastructure from data aggregators (such as GBIF and GeoCase) is that DiSSCo works with mutable data. The aim is to generate value by extending the information of the Digital Specimen. These extensions, which we call annotations, can be in the form of data linkages or generate new information derived from the existing data. Together these create an Extended Digital Specimen (DES), holding more information than was presented in the physical specimen [Hardisty 2022b].

While annotating the Digital Specimen can be done by both human and machine, we expect the latter to produce the bulk of the annotations - it is key for DiSSCo infrastructure to facilitate this machine annotation and data linkage. The DiSSCo core infrastructure provides a pluggable platform where adding new automated annotation services requires minimal effort [Leeflang 2022]. Automated annotation services can be externally developed and hosted, decoupling them from the core infrastructure.

The Specimen Data Refinery (SDR) is one of the automated annotation services which could connect to the DiSSCo infrastructure. Below we describe two setups on how the SDR could be integrated with DiSSCo core architecture.

2.1.1 Galaxy setup

Automated annotation services, such as the SDR, can be triggered by different actions, automatically or manually. As a service, it can be requested when a new dataset is ingested. This means that the SDR workflow is triggered for each new specimen DiSSCo receives in that dataset (fig. 1). In this scenario, the Digital Specimen Processing Service (where we evaluate if a specimen is new) will send an event to a dedicated SDR queue. This event indicates which digital specimen has been newly created and contains all data for the SDR workflow. As the SDR will be housed outside the DiSSCo infrastructure, we use an intermediate service (SDR wrapper) to read the event and call the Galaxy API. Galaxy is the web-based workflow management system in which the SDR has been developed.

The call to the Galaxy API will start the SDR workflow, where multiple SDR tools will be triggered in sequence. The result from these tools will be wrapped together in a Research Object Crate (RO-Crate). This Research Object will be returned to the SDR wrapper. Data in the RO-Crate will be converted to a DiSSCo annotation object which we will send to the Annotation Processing Services. This service will attach the annotation to the Digital Specimen.

We could also trigger this process when a digital specimen has been updated. When new information is received regarding the digital specimen this may require a check of the current annotation or provide information for new annotations. It is important that we trigger the SDR tools only when required to minimise resources wasted.
Besides automated activation for new or updated data, we also want users to be able to request an automated annotation service to run over a dataset. Users with sufficient rights will be able to select one or multiple Digital Specimens and request to run the SDR workflow over the items.

DiSSCo will keep provenance records on all actions on the Digital Specimen. This means that if any of the annotations from SDR is changed, the change will be traceable. We will know who triggered the action, at what time and what information was updated. Reverting to a previous version of the Digital Specimen or the annotation will be possible.

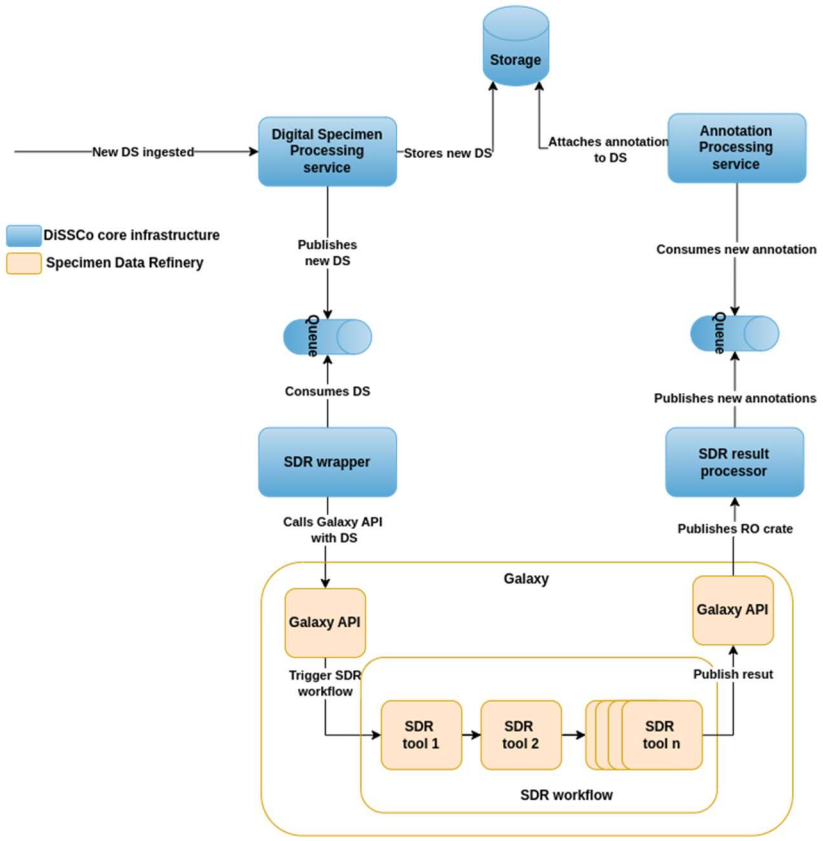

Figure 1 *- Architectural overview of SDR galaxy integration with DiSSCo.*

2.1.2 Standalone tools setup

While the above setup would work, there are some key considerations. The SDR and its platform (Galaxy) were mostly tested locally or via on-prem virtual machines. To scale and run these tools in parallel for many thousands of images we would likely want to make use of existing HPC infrastructure where possible (e.g., NHM has an HPC Galaxy instance for their bioinformatics laboratories). Within the DiSSCo core infrastructure we try to implement automated scaling as much as possible but would need to undertake additional work to benchmark and understand the compute cost requirements to scale the tools used within the SDR. We would also need to consider the costs of image storage and data transfer, aspects of which are discussed in the community paper by [Groom et al, 2022] but there are examples of using Galaxy at scale for imaging analysis (e.g., https://imaging.usegalaxy.eu/).

We can also consider running relevant tools outside of Galaxy. Depending on how they are used, some of these tools may be used directly in edge compute environments (e.g., on the desktop or laptops used by digitisation teams) but losing the benefits of the workflow management as described by [Livermore et al, 2023a]. If integrated with core DiSSCo infrastructure, Galaxy would likely be used via its API and further work would be required to understand the autoscaling and parallel processing required.

2.2 Future Development

With the help of SDR developers, DiSSCo has set up its own instance of SDR and has begun experimenting with the different tools. We will move forward with this pilot during the DiSSCo Transition phase. This will provide us valuable feedback on whether a workflow management system (e.g., using Galaxy or an alternative) approach will meet our needs and be able to be efficiently integrated with broader DiSSCo core services. Further work is also needed to explore the potential integration of the SDR with third party services.

3. Documentation and evaluation

3.1 Documentation

During the entire SDR research and development process we did as much as we could in an open way, either in public project documents, or in GitHub.

As part of the initial scoping work, we wrote and reviewed a minimum viable product document to understand the requirements, and to clarify any ambiguities in approach or functionality – this covered both Task 8.2 and Task 8.3. Once complete, a large proportion of this document was moved into a DiSSCo-managed GitHub repository (see https://github.com/DiSSCo/SDR/wiki/Minimum-Viable-Product-Review) to ensure community access and sustainability of technical documentation and design decisions. All significant project outputs were recorded here (https://github.com/DiSSCo/SDR/projects/2).

As we reached the end of the project we agreed to use the Diátaxis documentation approach [Procide, 2023] which separates the needs of documentation users into four modes:

1. Tutorials: learning-orientated and practical
2. How-to Guides: task-orientated and practical
3. Explanation: understanding-orientated and theoretical
4. Reference: Information-orientated and theoretical

The documentation is signposted from the SDR repository README.md: https://github.com/DiSSCo/SDR

Documentation includes:

Tutorial

- SDR tutorial

How-to

- How to: create a new input file
- How to: deploy a new instance of the SDR
- How to: invoke the SDR workflow using the Galaxy API
- How to: configure the SDR job submission engine
- How to: add new tools for the SDR
- How to: customise landing page

Explanation

- Explanation: SSL certificates

Reference

- Reference: SDR tools
- Reference: Deployment Variables

As the SDR and many of its tools are still at a prototype stage, we prioritised documenting the initial setup and the components that we thought we less likely to change over a year or so after the project concludes. This should be enough support for the components that will be used in future DiSSCo work.

3.2 Evaluation

All the individual tools were tested, and tested in larger workflows where tools would rely on the inputs of other tools. While we got promising results from all the tools, we realised that we needed multiple iterations of model refinement (e.g., collect the poor results, manually annotate them correctly, retrain the model, test again) to improve the tools. We had limited time to do this, so many of the models incorporated in the tools were not improved during the project, or only retrained once. Tools that were dependent on the outputs of other tools had less opportunity for testing as much of the platform and tool development work happened in parallel.

Part of the challenge was using multiple systems: many of the tools used their own platforms to annotate and create training data, and the outputs of the SDR could not be easily imported for subsequent annotation. As we were using Galaxy in a novel way, there were no pre-existing Galaxy tools to support the kinds of annotation or review of results that other dedicated machine learning/image processing platforms have (e.g., Pixel Accuracy or Intersection over Union).

We provide some qualitative results for segmentation (fig 3) and text line detection (fig 4) as these were both fundamental tools upon which others were dependent on the outputs. HTR/OCR is discussed in Section 3.2.1.

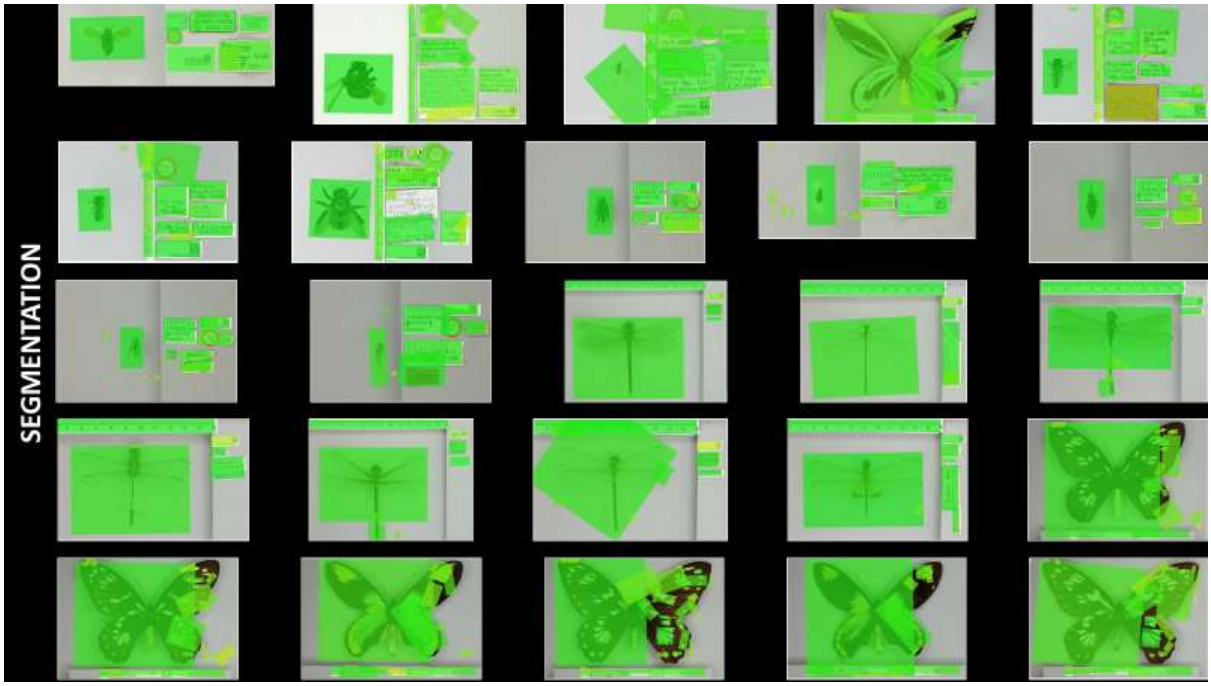

Figure 2 *- Segmentation results for 25 pinned insect specimen images outside of the initial training and evaluation set.*

In [fig. 2] for general image segmentation we can see underfitting on specimen types that were poorly represented in the initial training dataset – the large butterflies and Odonata, it's notably worse on the large Lepidoptera with false positives on the right-hand side of the image (where labels are typically placed). Segmentation works relatively well on more typical images in row 2.

Figure 3 *- Text line detection results for 25 pinned insect specimen images outside of the initial training and evaluation set.*

In [fig. 3] for line detection, prior to OCR/HTR we also see underfitting. You can also see false positives on high contrast regions of large lepidopterans and on scale bar variants that were poorly represented in the training dataset.

In addition to providing more training data, we could combine some of the previous regions to increase accuracy. It's worth noting we have also seen false positives in commercial barcode reading software on some Lepidoptera wing patterns.

3.2.1 Cloud services comparison - text clustering

The core components in the SDR workflow - semantic segmentation to detect labels, optical character recognition, and natural language processing - are available through many cloud AI services. These services can be chained together, providing workflow functionality like the SDR model, with tools for tracking the data through the service layer. The efficacy of the SDR and leading AI service providers (Google; Amazon; Aegir) was evaluated through a simple experiment: manually transcribe the key informational elements (taxonomic name; locality, collector, collection date) from 400 specimen labels, and compare these with the predicted results from the four services. The accuracy metric was calculated from the Levenshtein distance between the actual and predicted values.

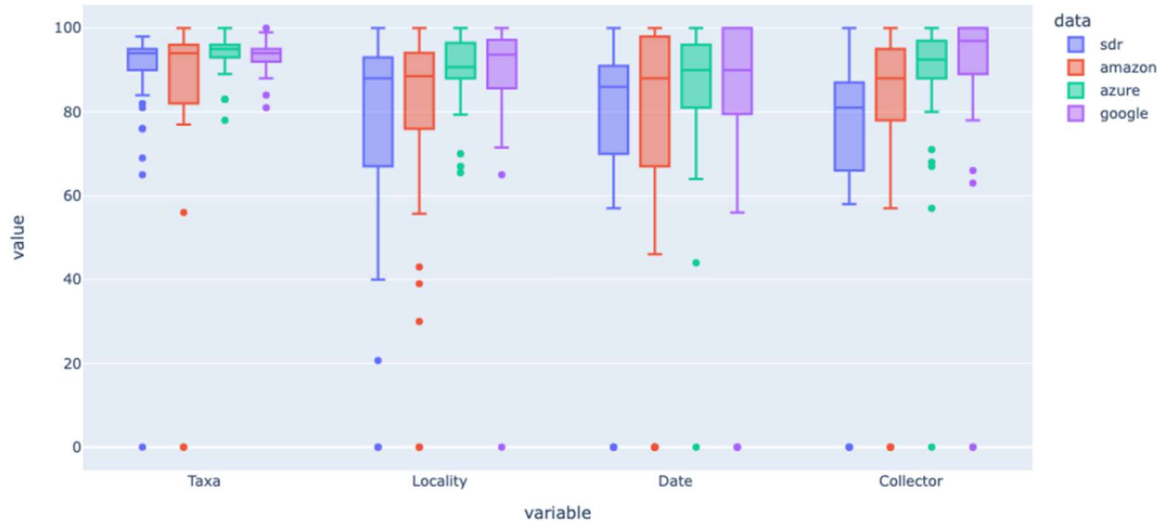

Figure 4 *– Comparison of accuracy between the SDR, Amazon, Azure and Google's named entity recognition on four different entity types (taxa, localities, dates and collectors).*

The results demonstrate that many cloud AI models outperformed our own developed for the SDR (fig. 4). What is perhaps surprising is that these results were achieved "out of the box". The cloud models were not trained on a custom corpus of literature or specimen images. The task of reading information from specimen labels is analogous to many other machine learning tasks the models have been trained for, and they were able to transfer that learning to our domain. Further improvements in the cloud AI accuracy might be achievable if they were pretrained on our datasets. This work was presented at TDWG 2022 - [Scott, 2022].

3.2.2 Learning from ChecklistBank training and evaluation

One of the tools originally planned for SDR was taxon name reconciliation - this was descoped due to time constraints; the need to improve the outputs of the tools upon which they would depend (e.g., Handwritten Text Recognition and Named Entity Recognition); and the fact that there are existing services to reconcile taxon names [Livermore et al, 2023a].

In recent years the Catalogue of Life/Species 2000 (COL), with its secretariat based at the Naturalis Biodiversity Center in the Netherlands, has been working together with the Global Biodiversity information Facility (GBIF, https://gbif.org) to develop a shared infrastructure. The new Catalogue of Life infrastructure was launched in December 2020. It consists of a public portal (https://catalogueoflife.org), giving access to the most recent Catalogue of Life Checklist; the authoritative listing of all the world's known species. In addition, it involves the ChecklistBank infrastructure (https://checklistbank.org) and API (https://api.checklistbank.org) that is jointly developed by GBIF and COL/Species 2000. ChecklistBank is an open data publishing platform focused on taxonomic and nomenclatural checklists. The infrastructure also supports (custom) taxonomic data services for biodiversity data infrastructures (such as GBIF and DiSSCo), but also policy initiatives such as the Convention on Biological Diversity and the European Environment Agency.

Catalogue of Life and ChecklistBank are expected to deliver taxonomic services to Natural History Collections directly as well as to DiSSCo. This includes the delivery of taxonomic services from Catalogue of Life and ChecklistBank to the SDR, which will be operational in DiSSCo.

During the SYNTHESYS+ project, Naturalis set up user helpdesk facilities and communication messages for the new Catalogue of Life infrastructure, including the ChecklistBank infrastructure. User feedback guided the development of training material for ChecklistBank, resulting in the 1st ChecklistBank tutorial: https://docs.gbif-uat.org/course-checklistbank-tutorial/en/index.en.html. This tutorial describes functionality of ChecklistBank that has been developed by GBIF and COL partially in the framework of the EU funded H2020 project BiCIKL (https://bicikl-project.eu/, grant agreement number: 101007492).

This tutorial formed the basis - with presentations and worked examples - for 6 ChecklistBank workshops, organised between June-December 2022 for a total of 72 participants.

The ChecklistBank approach and feedback (further details below) offers insights for future evaluation, training and documentation on SDR tools.

SYNTHESYS+ ChecklistBank training workshops - further details

These workshops provided a hands-on demonstration of four ChecklistBank tools:

- Explore the ChecklistBank repository to search, inspect and download checklists;
- Cross dataset search tool to look up the name usage of a particular scientific name in all data sources available in ChecklistBank;
- Name match tool, which enables a comparison of the COL Checklist with one or two other datasets in ChecklistBank in terms of taxon name matching

- Dataset comparison tool, which allows a comparison of two taxonomic datasets in ChecklistBank on a scientific name by scientific name basis.

Event	Date	Location	Number of participants
SPNHC conference	9 June 2022	Edinburgh, Scotland	15
COL Global Team meeting	29 June 2022	Leiden, Netherlands + online	12
CSIRO / Atlas of Living Australia workshop	6 Sep. 2022	Online	15
Naturalis workshop	16 Sep. 2022	Leiden, Netherlands	7
ENVRI-FAIR workshop	10 Nov. 2022	Online	11
Naturalis workshop	28 Nov. 2022	Leiden,Netherlands	12
		Total number of participants	72

Table 1 *ChecklistBank workshop events and number of participants.*

The majority of participants were working for a Natural History Collection, although there was also representation of other organisation types, such as biodiversity infrastructures (Fig. 5).

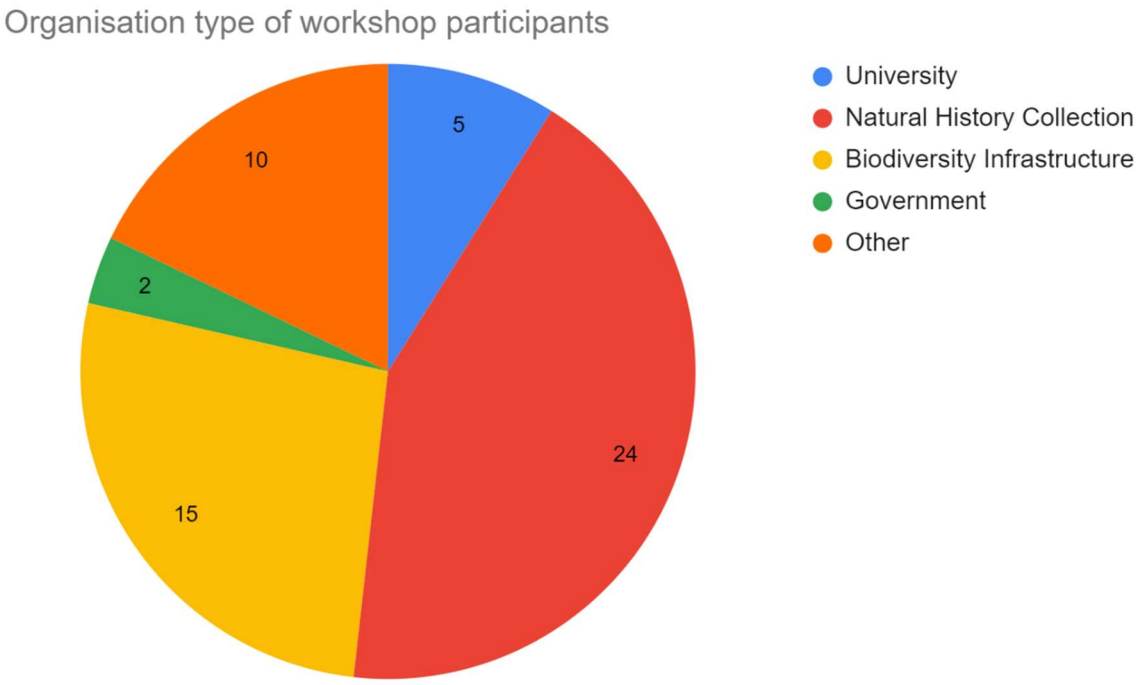

Figure 5 *- Pie chart showing the representation of participants across types of organisations.*

The main role of participants within their organisation was quite varied ranging from data curators and managers to programmers and taxonomists (Fig. 6).

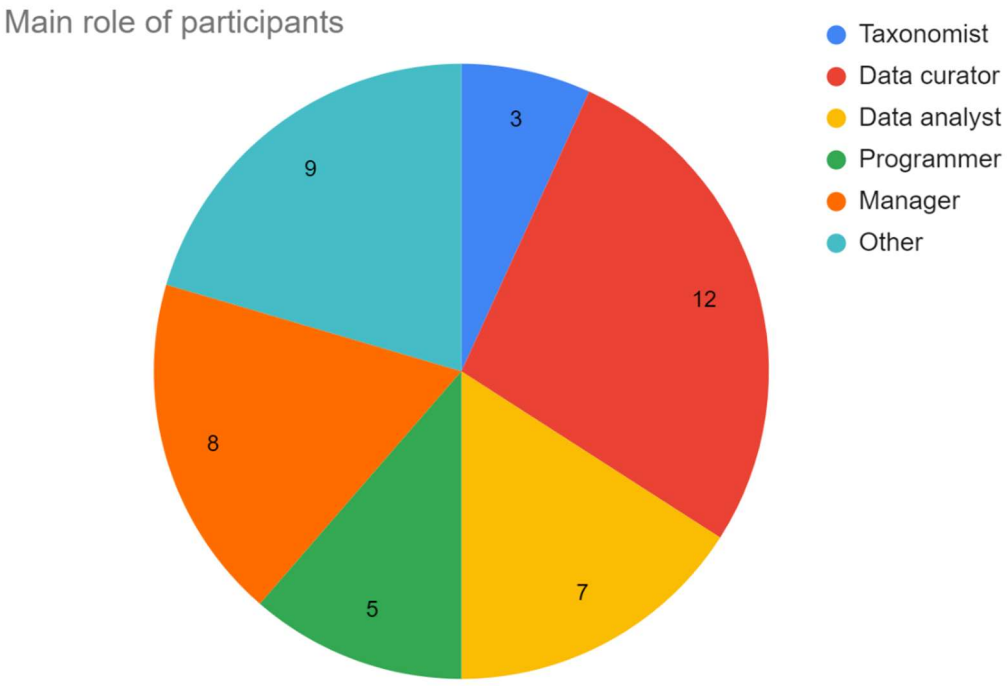

Figure 6 *- Pie chart showing the main work role of the participants in most workshops*

At the start of each workshop, most participants indicated to be familiar with what taxonomic checklists are (average score of 3.2 out of 5), about half of the audience had been using checklists

before (2.4 out of 5), but most people had not much experience with developing checklists themselves (1.8 out of 5).

Participants of the workshops were asked what their biggest pain points currently are with checklists. Aspects that were often mentioned are:

- Lack of transparency and metadata for taxonomic decisions (i.e. what is the taxon concept and who developed it)
- Incompleteness of synonymy
- Uncertainty to currency (is a checklist up to date) and accuracy
- Disagreement between lists on taxonomic concepts, which makes it difficult to align different checklists
- Limited integration between systems
- No links between the scientific names and the treatments in the literature
- No persistent identifiers for scientific and taxonomic names

Participants indicated that the ChecklistBank tools were relatively easy to use, with the dataset comparison (diff viewer) the most difficult, but still only 2.7 out of 5 (Fig. 7).

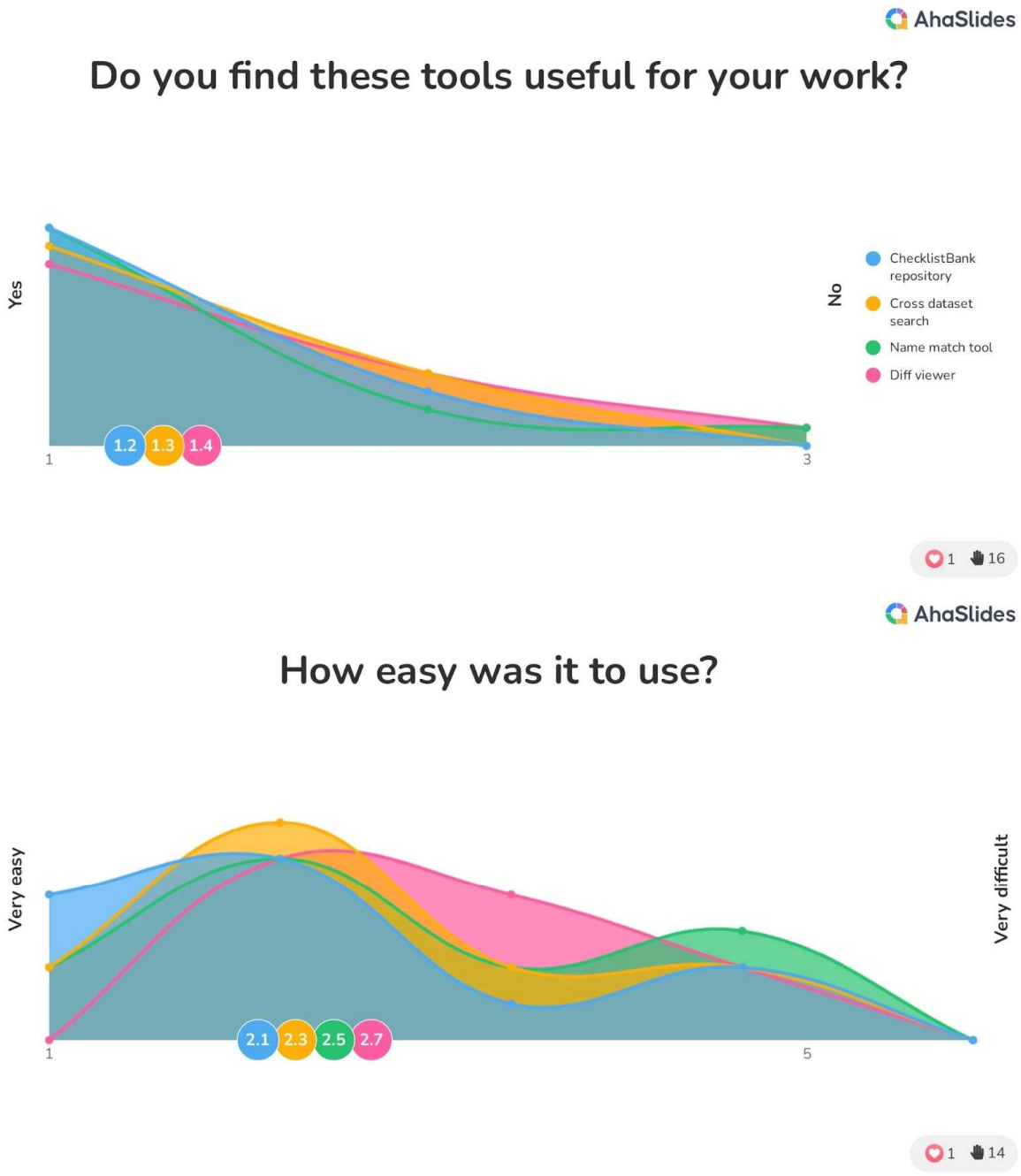

Figure 7 *Graph showing on a scale from very easy (1) to very difficult (5) the ease of the use of each of the four functionalities of ChecklistBank. The results of the graph shown above originate from the CSIRO-ALA workshop.*

Workshop participants also indicated that the tools in ChecklistBank are very relevant for their work. Overall the dataset comparison (Diff viewer) and the name match tool were seen as the most interesting features (Fig. 8). That means that most participants are looking for support in the matching of taxon names and comparisons between checklists to investigate and understand taxonomic differences.

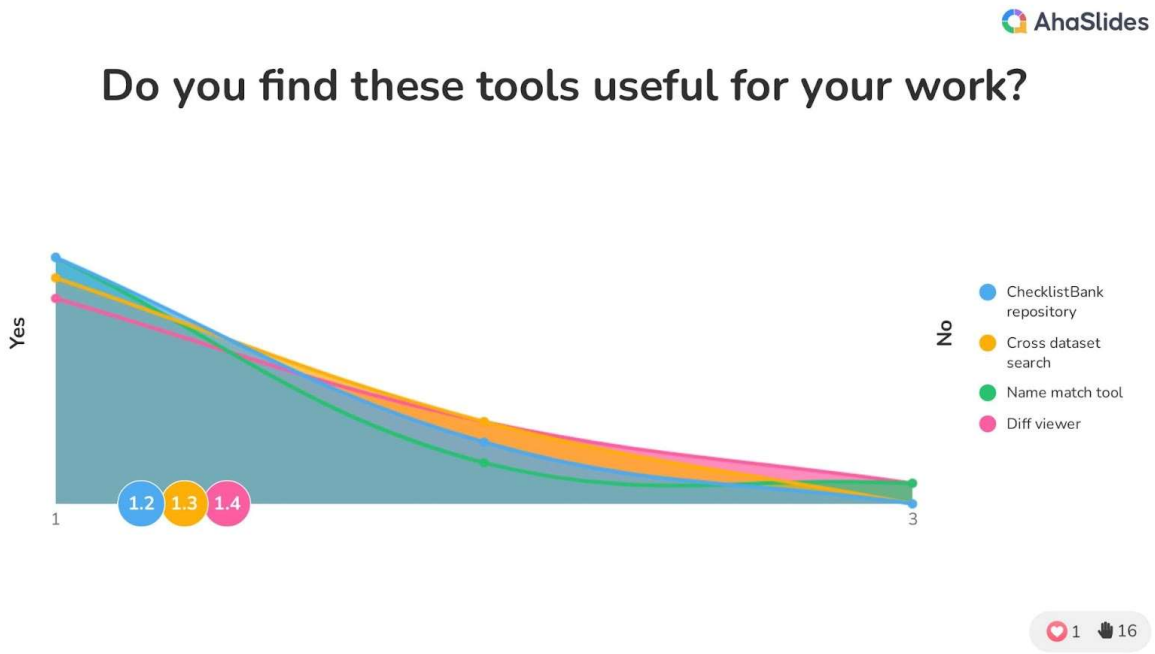

Figure 8 *- Graph showing to what extent participants felt the presented ChecklistBank functionalities were useful for their work. The response per tool was collected at a scale from yes useful (1) to no not useful (2). The results of the graph shown above originate from the CSIRO-ALA workshop.*

The following aspects were mentioned as possible improvements for future work on Catalogue of Life and ChecklistBank:

- Make taxonomic gaps in the Catalogue of Life Checklist apparent so groups know they need to fill them and to contribute to these gaps;
- Include the possibility to build and rank your own taxonomic backbone based on the Catalogue of Life Checklist for national portals;
- DOIs for all datasets in ChecklistBank, not only for the COL Checklist and underlying data sources;
- Show and share how ChecklistBank is built to be transparent about what is or is not included from a given checklist, and especially in relation to the COL Checklist.

4. Dissemination and promotion of the SDR

Part of the purpose of this task was to promote the understanding and usage of the SDR across the DiSSCo consortium. This dissemination and promotion has taken place across a wide range of meetings, workshops and conferences representing the whole community, as set out below.

4.1 Presentations

2023-06-01 Results and outcomes of the SYNTHESYS+ JRA3 (Joint Research Activity # 3) - Specimen Data Refinery (SDR). Presented by Robert Cubey - written by JRA3 SDR Research team in Narrowing the Gaps: The role of digital infrastructure in shortening the distance between physical collections and their derivative research products pt. 1 at SPNHC 38th Annual Meeting: Taking the Long View

(San Francisco) DOI (not yet available) Recording of prsenedtation https://www.youtube.com/live/wEfR1G0Yu8c?feature=share&t=3916

2023-02-08 Livermore, Laurence; Scott, Ben; Woolland, Oliver; Soiland-Reyes, Stian (2023): Specimen Data Refinery Showcase. figshare. Presented at DiSSCo Futures. https://doi.org/10.6084/m9.figshare.22040348.v2

2023-02-07 Livermore, Laurence; Scott, Ben; Woolland, Oliver; Soiland-Reyes, Stian (2023): Transforming Digitisation Using Automation. figshare. Presented at DiSSCo Futures https://doi.org/10.6084/m9.figshare.22027988.v1

2022-10-18 Livermore, Laurence; Brack, Paul; Scott, Ben; Soiland-Reyes, Stian; Woolland, Oliver (2022): The Specimen Data Refinery: Using a scientific workflow approach for information extraction. figshare. Presented at TDWG 2022. https://doi.org/10.6084/m9.figshare.21312345.v1

2022-06-07 Livermore, Laurence; Brack, Paul; Scott, Ben; Woolland, Oliver (2022): Specimen Data Refinery: A novel approach to automating digitisation. figshare. Presented at SPNHC 2022. https://doi.org/10.6084/m9.figshare.19947845.v2

2022-04-07 Livermore, Laurence (2022): Specimen Data Refinery. figshare. Presented at DiSSCo Prepare All Hands Meeting 2. https://doi.org/10.6084/m9.figshare.19529572.v2

2021-12-08 Livermore, Laurence; Scott, Ben; Gu, Qianqian; Carole Goble; Brack, Paul (2021): Work package 8 JRA3: Specimen Data Refinery. figshare. Presented at SYNTHESYS+ AGM . https://doi.org/10.6084/m9.figshare.17124377.v1

2021-07-22 Livermore, Laurence; Scott, Ben; Dillen, Mathias (2021): Contemporary and Established Provenance Issues in Natural History Collections. figshare. Presented at ProvenanceWeek 2021. https://doi.org/10.6084/m9.figshare.15035370.v1

2021-04-26 Livermore, Laurence (2021): Specimen Data Refinery: Project Overview and Update. figshare. Presented at CETAF ISTC/DWG Meeting. https://doi.org/10.6084/m9.figshare.14472570.v1

4.2 Poster

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 (FDO2022), Leiden, The Netherlands. Zenodo. https://doi.org/10.5281/zenodo.7233688

4.3 Peer-reviewed papers, preprints and abstracts

Brack P, Crowther P, Soiland-Reyes S, Owen S, Lowe D, Williams AR, et al. (2022) Ten simple rules for making a software tool workflow-ready. PLoS Comput Biol 18(3): e1009823. https://doi.org/10.1371/journal.pcbi.1009823

2022-10-12 Woolland O, Brack P, Soiland-Reyes S, Scott B, Livermore L (2022) Incrementally building FAIR Digital Objects with Specimen Data Refinery workflows. Research Ideas and Outcomes 8: e94349. https://doi.org/10.3897/rio.8.e94349

Groom, Q., Dillen, M., Addink, W., et al. Envisaging a global infrastructure to exploit the potential of digitised collections. Authorea. October 31, 2022. https://doi.org/10.22541/au.166678848.82362633/v2

2022 Hardisty, A., Brack, P., Goble, C., Livermore, L., Scott, B., Groom, Q., Owen, A., Soiland-Reyes, S.; 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

5. Conclusion and future work

Any practical implementation of machine learning and software-based automation requires a low barrier to entry to get wide adoption in a workforce with mixed technical skills. The bioinformatics community has been developing and using Galaxy for over 17 years, with a broader adoption of workflow and virtual research environment-based approaches to work. Galaxy is now used outside the biosciences for Climate change modelling, Bioimage processing, astronomy, public health, materials sciences etc. The EuroScienceGateway project expands the Galaxy platform and its Pulsar Network (a wide job execution system distributed to scale computing power over heterogeneous resources) to become a production-ready interface for European computing resources, including EGI and EuroHPC as Pulsar providers. FAIR workflow services pioneered in the EOSC-Life cluster (WorkflowHub, Workflow RO-Crate, metadata standards like schema.org and FAIR workflow best practices are being further developed in a basket of Horizon Europe EOSC projects, including: EuroScienceGateway, BY-COVID, FAIR-IMPACT, EOSC4Cancer, AgroServ, BioIndustry4.0, and, in Biodiversity, BioDT and BGE. They are also being developed in Australian (Galaxy Australia, Australian BioCommons) and the USA.

While parts of the natural science collections community have some experience with workflow management systems, this is typically for genomic and molecular work, rather than digitisation and digital curation of specimens. We had assumed that our users would be familiar with the concept of programmatic thinking, and would find workflows intuitive.

Following discussions with prospective users, and considering other practical considerations like the stage at which to implement the SDR in digitisation workflows, it became clear that using Galaxy and workflows has integration challenges for digitisation teams. Much existing processing of images and files is done locally. In a separate project to process images from a specialised digitisation workstation, we are planning to use Luigi to manage workflows locally on the workstation computer. This also has the advantage of processing images without the need to transfer them across networks. Using workflow management systems like Galaxy has some clear advantages, especially when scaling up enrichment services.

Currently over 330 workflow management systems (https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems) are in circulation in the scientific community. Each has its pros and cons and the best selection of a system depends on the parallelization and execution capabilities, the "plugged-in" support of data types & specific codes, the skills level of the workflow developers, and its popularity & sustainability by communities. Currently in the Biosciences the top three systems with significant communities are: Galaxy, Nextflow and Snakemake. Best practice in workflow design that is sympathetic to the tasks, user base and computational setup is also essential (see Taylor Reiter and others, Streamlining data-intensive biology with workflow systems, *GigaScience*, Volume 10, Issue 1, January 2021, giaa140,

https://doi.org/10.1093/gigascience/giaa140). No workflow system can wholly compensate for poor tool implementation.

We recommend:

- A review of the experiences of the SDR pilot to systematically identify the capabilities needed of an entire workflow infrastructure, including compute locality and availability, in order to pick the most appropriate framework and platforms or mix of platforms.
- A focus on building effective, parallelisable and workflow enabled tools [Brack et al, 2022] that can be incorporated in a range of workflow systems.
- A focus on the holistic sets of services needed. The workflow output FDOs would benefit from a repository and to date there is none.
- A continued partnership with ELIXIR, the sponsor of the EOSC-Life Workflow Collaboratory, for co-development of services. The Collaboratory is designed to be workflow agnostic and to support any kind of workflow platform and even combinations. The WorkflowHub currently registers workflows from 14+ different systems and provides sharing space for workflow teams; the LifeMonitor workflow testing monitor supports a range of testing frameworks for different systems
- Continue to use general interoperability standards. RO-Crate is a general framework for any kind of workflow and data and Bioschemas profiles are equally general. CWL may be used as an implementation or as a documentation sister for native platforms (Abstract CWL).
- Opening up collaborations with sister Research Infrastructures working in related fields - notably EuroBioimaging-ERIC and INSTRUCT-ERIC who do a great deal of large scale image processing using workflows.
- Developing the capacity of the SDR community with regard to workflow best practices and programmatic thinking.

Our very broad ecosystem of digital asset management systems and collections management systems means there is limited consensus on data import and exchange between systems. This means any generic tool or workflow will require additional mapping and conversion before data can be imported. Many data types are not supported at all by these systems (e.g., image annotation of text lines or other features, provenance) meaning any outputs for SDR-like tools cannot be easily archived, searched, or reused internally (e.g., for creating new training datasets). This lack of interoperability will create challenges for future DiSSCo services as it reduces the ease that collection holding institutes will be able to benefit from community tools, and improved or enriched data. In addition, the historically slow pace of change of these systems to support new data standards will stymy efforts.

Since we wrote the original proposal in late 2017/early 2018 (the call deadline for SYNTHESYS+ was 22nd March 2018) the landscape for AI and ML tools has changed dramatically. All the major cloud service providers provide a range of competitively priced API-based AI services (e.g., Google's Cloud Vision API, Amazon's Rekognition, Microsoft's Azure Cognitive Service for Vision) and the application of generative AI (e.g., ChatGPT which was publicly released on 30th November 2022) has received widespread news coverage and public usage. The Galaxy Machine Learning community has incorporated an extensive range of ML tools, and AlphaFold is just one example of an AI tool available to Galaxy using GPU clusters. However, our SDR pilot did not exploit these tools.

Were we to start the SDR work in 2023, we would almost certainly make use of commercial and open-source services to assist with the creation of training datasets and the ecosystem of tools they provide to support AI/ML work. While there are concerns about provenance and reusability with commercially available services, we need to evaluate the costs and benefits of these services versus

the costs and challenges associated with building and deploying our own AI/ML models for digitisation.

The investigative and pilot work done through the SDR will inform the data management plan for DiSSCo, and we intend to re-use tools and approaches in the core DiSSCo infrastructure upon completion of the SYNTHESYS+ project. Follow-up work has been planned in subsequent phases of DiSSCo development, e.g., in DiSSCo Transition. Other DiSSCo-linked projects will continue to explore FAIR Computational Workflows e.g., BioDT and BGE Biodiversity Genomics Europe.

6. Code Repository & Related Issues

GitHub repository for overall SDR project: https://github.com/DiSSCo/SDR

7. References

[Hardisty 2022b] Hardisty, A., Brack, P., Goble, C., Livermore, L., Scott, B., Groom, Q., Owen, A., Soiland-Reyes, S.; **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

[Leeflang 2022] Sam Leeflang, Claus Weiland, Jonas Grieb, Mathias Dillen, Sharif Islam, David Fichtmüller, Wouter Addink, Elspeth Haston (2022):
DiSSCo Prepare WP D6.2 Implementation and construction plan of the DiSSCo core architecture DiSSCo Prepare Deliverable 6.2
https://doi.org/10.34960/50b9-kj05

[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., Goble, C., Hardy, H., Scott, B., Soiland-Reyes, S., Woolland, O. (2023) **Deliverable 8.3 - Specimen Data Refinery: Development of cloud platform for data-processing services**. August 2023.

[Procida, 2023] Procida, D (2023) **Diátaxis documentation framework**. https://diataxis.fr/

[Scott, 2022] Scott B (2022) Cloud AI: A comparison of specimen image data extraction processes. Biodiversity Information Science and Standards 6: e90951. https://doi.org/10.3897/biss.6.90951

[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