Project:	Synthesis of systematic resources
Project acronym:	SYNTHESYS PLUS
Grant Agreement number:	823827
Work Package:	Workpackage 8
Deliverable number:	8.2
Deliverable title:	Tools and services for extracting, enhancing and annotating natural history specimen data
Deliverable author(s):	Laurence Livermore - Natural History Museum, London Jonathan Blettery - Muséum national d'Histoire naturelle, Paris Robert Cubey - Royal Botanic Garden Edinburgh Mathias Dillen - Meise Botanic Garden Carole Goble - The University of Manchester Helen Hardy - Natural History Museum, London Elspeth Haston - Royal Botanic Garden Edinburgh Christopher Kermorvant - TEKLIA Mario Lasseck - Museum für Naturkunde, Berlin Matthias Obst, University of Gothenburg Andreas Plank - Botanical Garden and Botanical Museum Berlin Ben Scott - Natural History Museum, London Stian Soiland-Reyes - The University of Manchester Oliver Woolland - The University of Manchester Zhengzhe Wu - Finnish Museum of Natural History, Helsinki
Date:	2023-08-08

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] and focuses on the tool and service development (Task 8.2). It is the formal report for the software demonstrator Deliverable 8.2. It describes the technology, development, and design approach of *tools, and summarises their functionality on an individual tool/service basis*.

1.1 Scope

This report primarily describes the tools and approaches used in 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].

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

Deliverable 8.4 [Livermore et al, 2023b] covers SDR integration with the wider DiSSCo architecture; documentation and evaluation; and dissemination and promotion of the SDR.

2. Design Description & Approach

The following sections have been adapted from [Hardisty et al (2022)], which describes the full workflow and FAIR Digital Object (FDO) approach for the SDR. Figure 1 gives an overview of the general structure of the SDR with a generic workflow within the Galaxy environment.

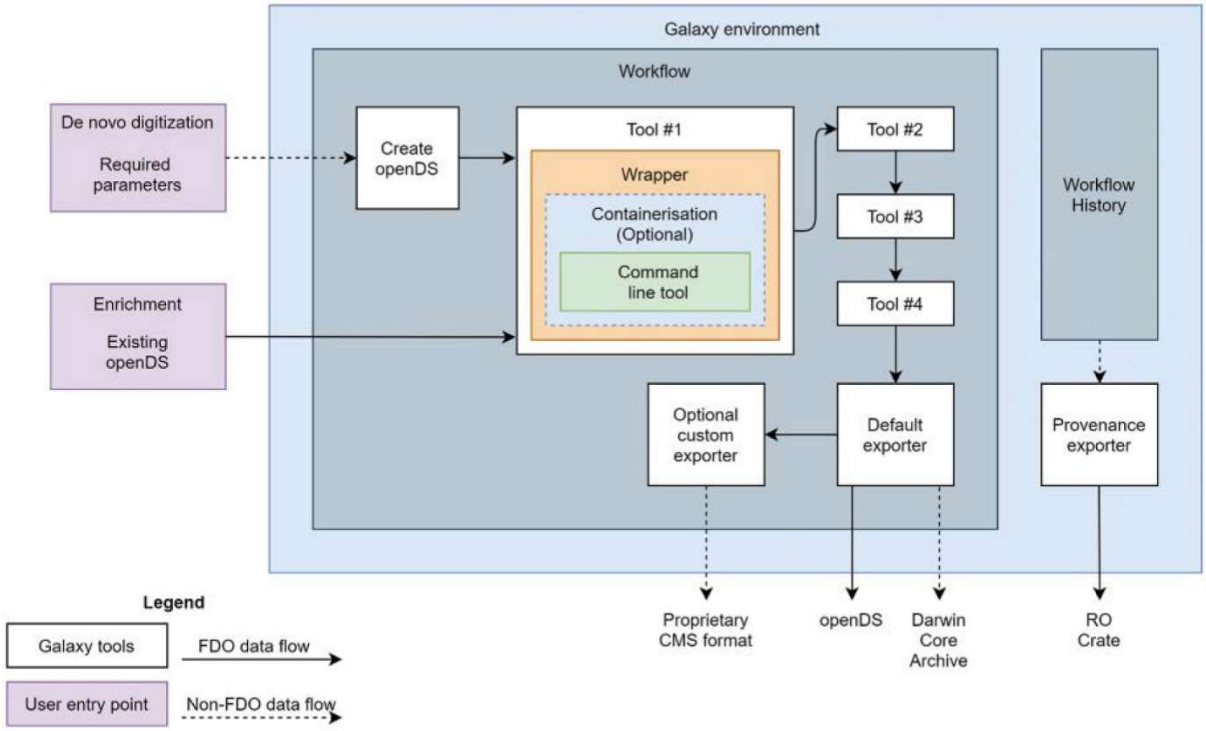

Figure 1 *- The general approach adopted for the SDR as a Galaxy workflow management system implementation [Hardisty et al (2022)].*

2.1 Need for automation

The costs of digitisation and "jury rigged" approach of digitisation workflows has been recognised for over a decade, as has the potential for automating parts of the digitisation process [Blagoderov et al 2012]. An analysis of transcription approaches by [Walton et al (2020b)] highlighted that efficient and cost-effective transcription of specimen label data [Figure 2] was still a major barrier in the mass digitisation of natural history collections. [Groom et al (2022)] more recently describe the need of a global infrastructure to support the extraction of data from biological collection images, as such images are currently often not very accessible, interoperable, and re-usable.

Building on our experience in SYNTHESYS3 [Haston et al, 2015; Hudson et al 2015] we proposed to address this by testing repeatable workflows, composed of modular tools, to automate the transcription and contextualisation of label data, and other data from specimen images.

Proposed tools and functionality included:

1. **Optical character recognition data capture and analysis** - to extract both handwritten and typed text, and to extract and identify information from the text (e.g., classify text strings into people, places and dates).
2. **Data mining and linkage** - linking text created from the previous tools to specific entities (e.g., taxon names to the GBIF backbone, or people to Wikidata Q codes).

3. **Image analysis** - Automatically finding and annotating features of interest in standardised specimen images, including the specimen, barcodes, scale bar, labels, colour charts etc.
4. **Georeferencing** - Taking a locality string and determining coordinates.
5. **Human interaction and crowdsourcing** - A large focus on creating training datasets for the other tools but considering how “human in the loop” would work in workflows.
6. **A workflow and tools registry** - A place where we can store and reference tools and workflows.

2.2 Users, User Stories and Specimen Constraints

We identified two kinds of users that should be supported by the SDR: digitisers and collections managers/curators. Five high-level user stories describe and broadly encompass the functionality these users need:

1. As a digitiser, I want to construct a workflow from a set of predefined components, so I can use that workflow to digitise specimens to a predefined specification.

2. As a digitiser, 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 digitiser, I want to export the output of a digitization run, so I can consume the output of a digitisation run into my institution's collection management system.

To prove the SDR concept, three categories of preserved specimen types were selected to be supported initially: herbarium sheets, microscope slides and pinned insects (Figure 2).

Figure 2 *- The three specimen types tested in the SDR, all of which demonstrate the diversity of collection objects, which include handwritten, typed, and printed labels. (a) Microscope slide (NHMUK010671647), (b) Herbarium specimen (BM000546829), and (c) pinned insect (NHMUK013383979)*

2.3 Development Approach

We made the following design decisions on software development at the beginning of the project:

- **Containerisation** - Tools and services would be containerised using Docker. Containerisation is beneficial for a few reasons, including ensuring tools and services are portable and work consistently between systems, and compartmentalising functionality (e.g., our tools work) into discrete functional parts.
- **Python** - our main language would be Python to ensure interoperability between tools and services.
- **Open Source** (but not exclusively) - where possible we would develop and integrate open-source tools and services, but acknowledge the inclusion of private/licensed software where no suitable open source option was identified.
- **Workflow and platform** - use Galaxy [The Galaxy Community, 2022] as the workflow management platform (see [D8.3 report] for further details of this approach).

While working on the SDR we codified an approach for making software tools workflow-ready using the 'ten simple rules' approach [Brack et al, 2022]. These 'rules' aimed to provide guidance for writing software as dedicated tools within a workflow, or when converting an existing stand-alone research tool into a reusable, composable, well-behaved component within a larger workflow. The approach covered some general software development practices and generalised examples.

2.4 Data Standards

Specimen Data - We chose to use the open Digital Specimen (openDS) specification to group, manage and process fragments of information relating to specimens. At the time of tool implementation, the standard was still under development, but it will be a critical part of the FAIR infrastructure of DiSSCo [Hardisty et al., 2020]. Terms were initially mapped against the widely used Darwin Core standard[1].

Handwritten document processing results - We reviewed the different formats for describing handwritten document processing results. This was described in detail in a blog post [Bonhomme, 2021] and considered our need to store: 1) layout information (text lines); 2) optical character recognition and handwritten text recognition transcriptions; 3) named entities; and 4) metadata on source images. We considered ALTO, Page XML, XML TEI, but settled on our own JSON/JSON-LD as none of the other standards met our needs (see Appendix section 7.1). Both Microsoft and Google use JSON to output their OCR/NER results, and it is a commonly used notation format.

RO-Crate - One of the reasons for choosing Galaxy was the proposed support of RO-Crate[2], a lightweight approach to packaging research data with their metadata. In our case, we export data on the input data (specimen images) and all the resulting metadata, including the results of the workflow, the configuration of individual tools and the overall workflow used.

[1] https://dwc.tdwg.org
[2] https://www.researchobject.org/ro-crate/

3. Overview of Tools and Training Data

This section describes the tools developed or integrated into the SDR. For some of the more complicated tools, a short description of how they work is given. The original conceptual diagram of the SDR is given in Figure 3 which shows how the different potential tools would pass information between one another.

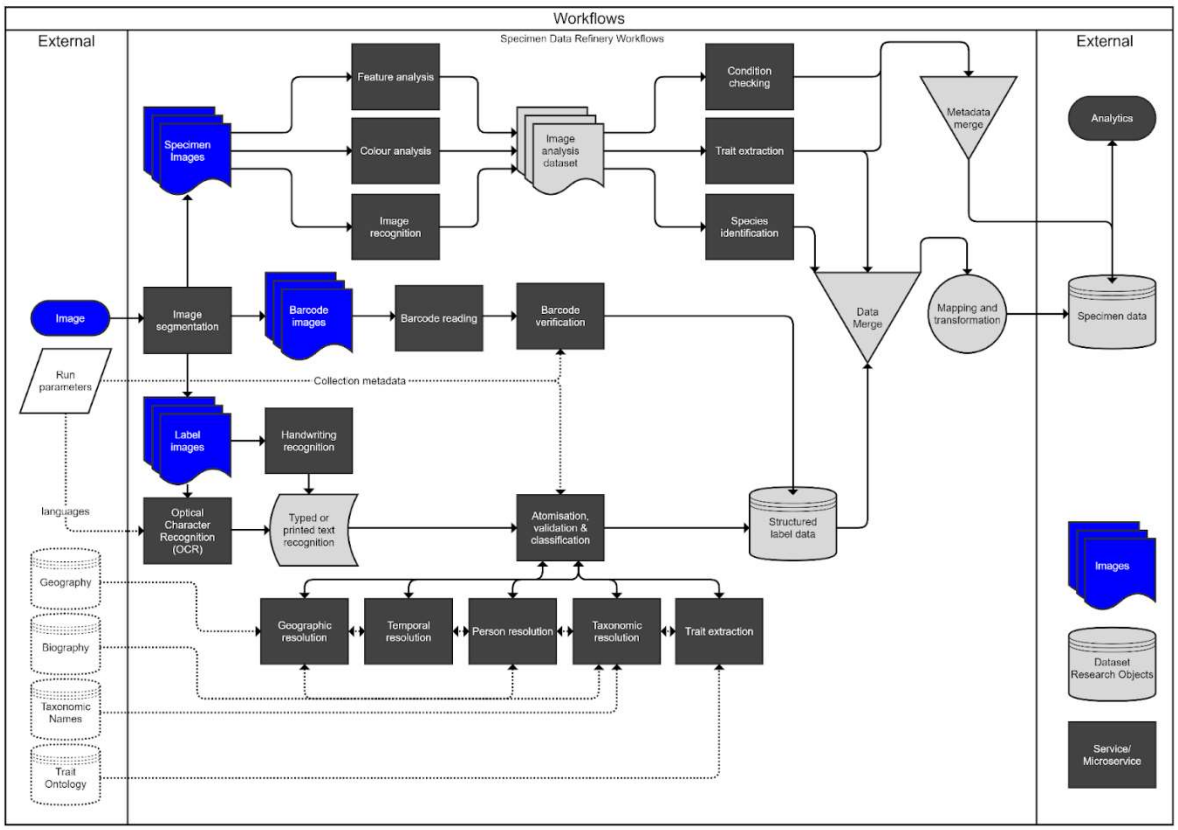

Figure 3 *- An overview of potential Specimen Data Refinery workflows based on image inputs and their derivatives, datasets, and services.*

3.1 OCR data capture and analysis

Handwritten Text Recognition (HTR)

Extracting handwritten text from labels is a key function that enables many other downstream tools (see Figure 3).

The HTR tool is based on the kaldi library[3]. Kaldi is a library for automatic speech recognition that can also be used for handwritten text recognition. The model consists of two parts - the optical model and the language model. For input, the tool gets an image of a single text line. The optical model processes the text line image and outputs probabilities of different character sequences. The language model then modifies these probabilities based on how much sense they make. For example, the model has seen “I can do it” probably

[3] https://github.com/kaldi-asr/kaldi

much more often than “I cam do it”, so it will increase the probability of “can”, even though looking at the image it’s not clear whether it’s “n” or “m”. A decoder chooses the best sequence of characters based on the probabilities and produces a transcription for the text line. This transcription is then added to the openDS object as the output of the HTR tool.

How does the HTR tool work in practice?
The HTR tool takes single text lines from the Document Layout Analysis (DLA) tool (see section 3.3 below), processes the image, and outputs character sequences, before assessing which is the most probable sequence. Figure 4 shows the JSON outputs of the labels with their confidence intervals (0 being no confidence, 1 being high confidence).

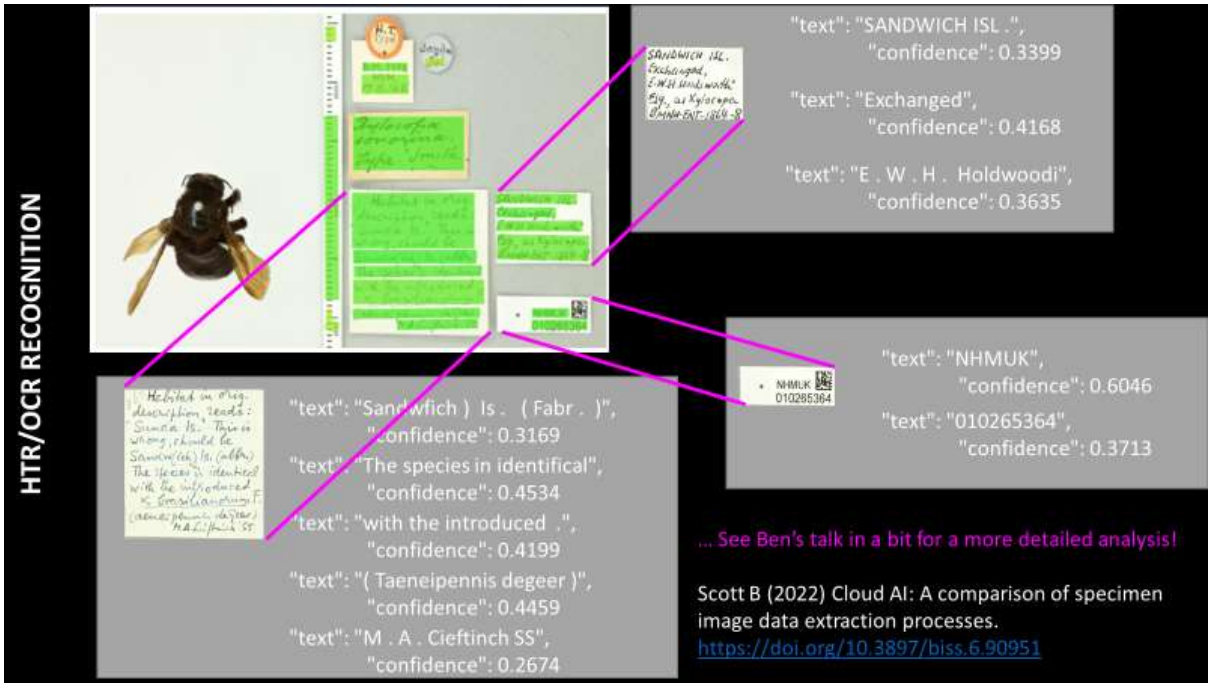

Figure 4 - HTR outputs (grey boxes, white text - top right, bottom left, bottom right) with lines indicating where the lines were processed from in the original image (top left).

Named Entity Recognition (NER)

Named entity recognition categorises the text from the HTR tool for further processing. This allows it to be mapped more easily to an existing data standard, run through human quality control, or imported into a collections management system. We settled on 11 classes for processing HTR data from pinned insect data[4]:

1. Collection/Donation
2. Date
3. Determination
4. Expedition
5. Identifier
6. Location name
7. Person name
8. Sampling Protocol
9. Sex
10. Taxon name
11. Type status

[4] https://github.com/DiSSCo/SDR/issues/4

The NER tool is based on spaCy[5]. It is a library for advanced NLP (natural language processing). It is used to find different types of entities in labels, for example specimen name, collector, date, location. The model is based on a transformer, which is the state-of-the-art model architecture for NLP. It is an encoder decoder model that uses a method called attention to have a better understanding of the context and the meaning of the words.

How does the NER tool work in practice?

For each token in the text, the model will predict which entity type it belongs to (if any). Since an entity can consist of multiple tokens the model will also predict whether it is the beginning, inside or last token of an entity. For example, the phrase “John Smith and Jane Smith have lived in New York since 1999.” could have entities like those in Table 1. The entities are concatenated and added to the corresponding openDS object.

Token	Entity Type
John	B-Person
Smith	L-Person
and	O
Jane	B-Person
Smith	L-Person
lived	O
in	O
New	B-Location
York	L-Location
since	O
May	B-Date
1999	L-Date
.	O

Table 1 *- Example tokens and entity types for named entity recognition.*

3.2 Data mining and linkage

See ‘3.8 Descoped Tools’

3.3 Image analysis tools

[5] https://github.com/explosion/spaCy

Barcode Reader

The barcode reading tool is a Python wrapper around Softek Barcode Reader Toolkit for Linux[6] version 9.1.4, a commercial library for extracting barcode values from images. It reads 2D barcodes, including Databar, Datamatrix and QR-code types, using JPEG or TIFF images as input.

DLA (Document Layout Analysis)

The DLA tool applies segmentation models to perform two tasks - text line detection (to be given to the HTR model) and segmentation of objects (color bar, bar code, label, specimen, scale bar). Both models use the same DLA tool, but with different configuration and models trained on different data.

The DLA tool is based on Doc-UFCN[7] which uses PyTorch[8]. It is a U-Net model, which predicts for each pixel in the downscaled image which class it belongs to (for example text line vs background). Then the pixels are grouped together to create regions in the openDS object.

How does the DLA tool work in practice?

The DLA tool has two different functions. The first is the location of text line regions. It takes a specimen image and locates potential text lines, outputting rectangular bounding coordinates (see Figure 5). These coordinates can then be passed on to the HTR tool.

The second function is similar but locates other objects or regions of interest like the specimen, scale bar, entire labels, and barcodes (see Figure 6). These coordinates can also be passed on to other tools, like a barcode reader.

[6] https://www.bardecode.com/en1/
[7] https://pypi.org/project/doc-ufcn/
[8] https://pytorch.org/

Figure 5 *- visualisation of the DLA tool text line detection.*

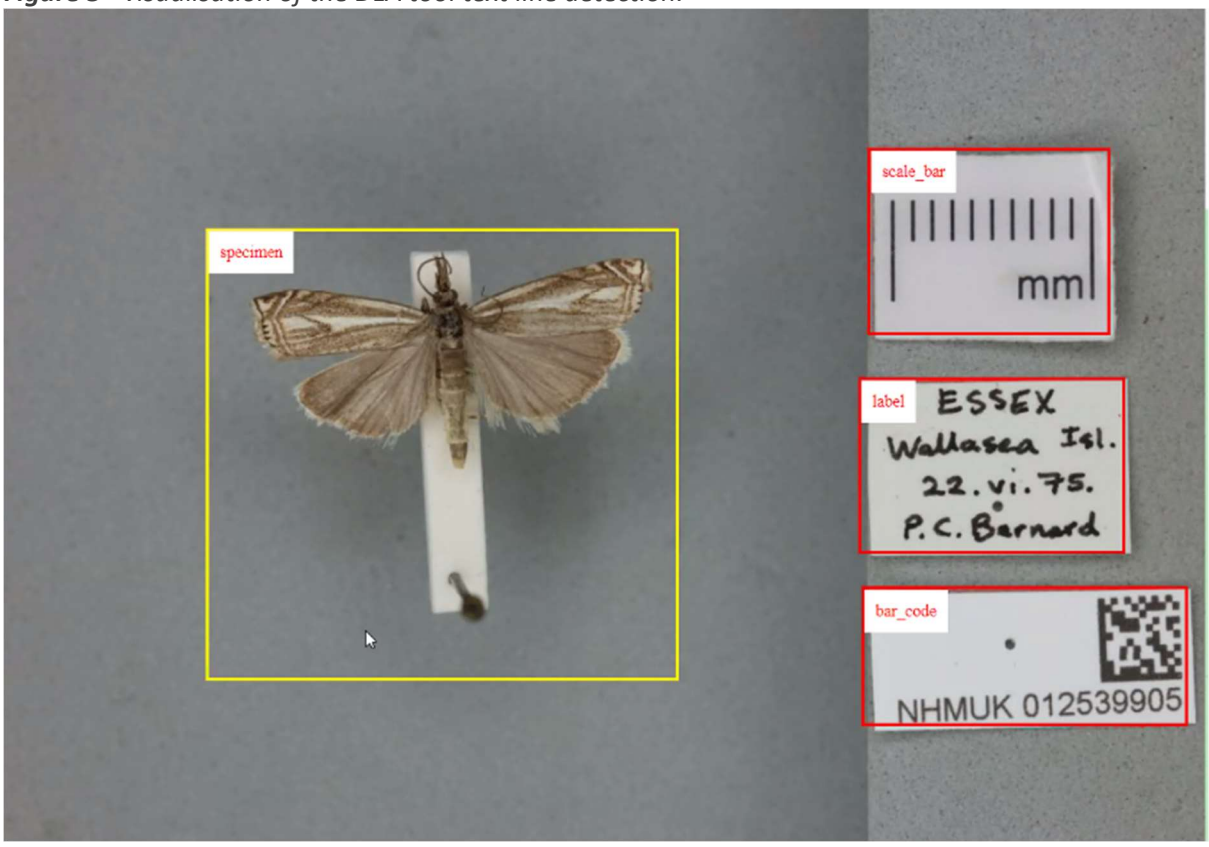

Figure 6 *- visualisation of the DLA tool object detection.*

Mothra

Mothra is an externally developed tool that analyses images of Lepidoptera (butterflies and moths) to segment the specimens, find ruler ticks to get physical pixel measurements, and measure the wings of Lepidoptera.

We implemented Mothra in the Galaxy instance of the SDR, but we were unable to integrate it elegantly as a containerised tool due to issues with containerisation and passing filenames. This was a useful exercise in integrating other tools and understanding some of the issues we would encounter using a Galaxy-based containerised tool approach.

How does Mothra work in practice?
Mothra takes an image of a lepidopteran set out in a standardised format and analyses it, before creating visual (Figure 7) and textual outputs (e.g., a result .csv with wing measurements). More information is available in the Mothra repository (https://github.com/machine-shop/mothra).

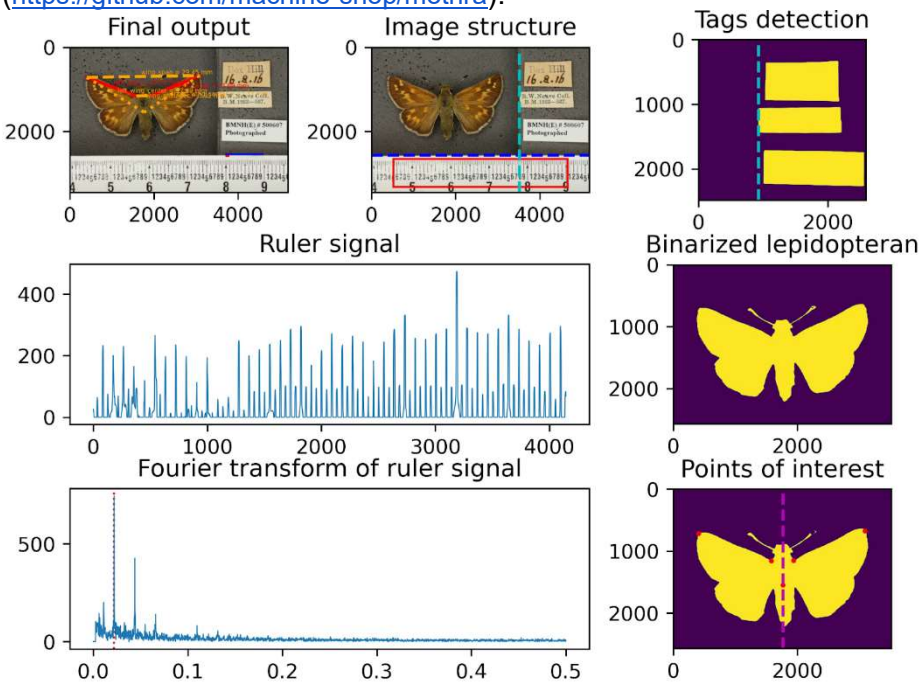

***Figure 7** - Visual output of Mothra.*

3.4 Georeferencing

Georg

Georg is an external web application designed to support georeferencing (the process of obtaining geographic coordinates from a locality description) for natural history collections data. It allows a user to select among suggested matches to an entered text string, or to choose a map point and then select a named place based on geographic proximity to the marker. Georg currently supports single-locality queries as well as limited batch processing for uploaded lists of localities. It is built on top of Pelias geocoder[9] and the data processing is carried out with the workflow management system Snakemake[10] and Python scripts. Georg was originally developed for searching for places in the Nordic countries, and during this

[9] https://github.com/pelias/pelias
[10] https://snakemake.readthedocs.io/en/stable/

project it was expanded to include the UK. It is possible to increase coverage to potentially any country in the world based on Who's on first[11] and OpenStreetMaps[12] data. More information is available in the Georg repository (https://github.com/Naturhistoriska/Georg)

How does Georg work in practice?

- Locality information is gathered for georeferenced locations from different data sources.
- Information is processed to make the locations searchable before they are imported into Georg.
- When you search, Georg compiles and ranks the result so that you get the most relevant search hits.

3.5 Training datasets and human interaction tools

Training Data

Label Studio - Label Studio is a tool for different annotation tasks. In S8+ Label studio is used for annotating named entities. The texts to be annotated are imported from Arkindex. When a user has finished annotating the annotations are exported to Arkindex.
Label Studio is open source, anybody can host it themselves. However, there is an instance of Label Studio running that is configured to communicate with Arkindex.

Arkindex - Arkindex is a platform for document processing. DLA and HTR annotations are created on arkindex. Models that have already been trained can be used to help with the annotation, and can then be iteratively improved by using more annotated data.
Arkindex is not open source - to set up your own instance you need to contact Teklia.

Model creation - For training a model we have different tools to generate training data in the correct format from Arkindex. The models are trained on the training data, then validation data is used to choose the best parameters. In the end the model is evaluated on test data to see how the model performs on unseen data. Currently the training is done locally - not related to Arkindex or Label Studio.

Human Interaction Tools

See '3.8 Descoped Tools'

3.6 Workflow and tools registry

The prototype workflows created in Task 8.2 are registered in WorkflowHub. This allows them to be publicly published, registered and more easily shared. WorkflowHub is mutually coupled with Galaxy so that workflows can be discovered in the Hub and potentially used by other public-use Galaxy instances.

SDR project in WorkflowHub: https://workflowhub.eu/projects/72

[11] https://whosonfirst.org/
[12] https://www.openstreetmap.org/#map=4/62.99/17.64

3.7 Workflow Utility Tools

We developed a set of tools that were necessary in the Galaxy workflow approach. These include:

Wrapper Tool - The team implemented a generic wrapper tool, which received an openDS object input, python entry points for acting on the object, and then validating and output of the modified openDS object. This tool was then cloned and entry points modified for the development of each custom tool.

OpenDS validation tool - Validation was performed by validating the openDS JSON object, against the development version of the OpenDS JSON schema [ref: https://github.com/DiSSCo/openDS]. The JSON schema file object to validate against was set as an SDR environment variable, to allow us to update to the latest schema as the development of the standard progressed.

Split file to collection - this tool processed a CSV file containing separate rows of URIs for specimen images and their metadata, and created a single Galaxy 'Collection'. In Galaxy, a 'Collection' of multiple datasets (in this case specimen images) can be processed and displayed as a single unit, making the user interface less cluttered.

Image download - A simple tool to download images from the URIs provided in an input specimen CSV file.

3.8 Descoped Tools

Data mining and linkage tools - We had planned to build tools for reconciling people names[13] and taxon names[14]. These tools were eventually descoped due to time constraints, and the need to improve the outputs of the tools upon which they would depend (e.g., HTR and NER). There are existing services that can resolve both people and taxon names e.g., Bionomia's people name reconciliation tool (https://bionomia.net/reconcile) and GBIFs Species API which includes name matching functionality (https://www.gbif.org/developer/species).

Human Interaction Tools - We tested Galaxy's interactive tools to evaluate whether they could be included in any of our workflows. One of the potential uses was to have a user give immediate feedback through annotations on the results, either of the OCR or image analysis. This would have potentially made it much easier for our developers to iterate and improve the various models behind the tools. Unfortunately we were unable to get the interactive tools working effectively with our workflows. This meant that any annotations and subsequent corrections on results needed to be made outside of Galaxy, which was a slow and inefficient process.

[13] "Create a Galaxy tool for reconciling taxon names" https://github.com/DiSSCo/SDR/issues/90
[14] "Create a Galaxy tool for reconciling people names" https://github.com/DiSSCo/SDR/issues/89

4. Dissemination of Results

We have included a high-level summary of presentations and publications associated with Task 8.2. Much of the dissemination work done for Task 8.2 includes work done as part of Task 8.3 ”Development of cloud platform for data-processing services” and Task 8.4 "Data Delivery and Exploitation". A comprehensive list is given in the D8.4 report [Livermore et al, 2023b].

- [Peer reviewed paper] Brack P, Crowther P, Soiland-Reyes S, Owen S, Lowe D, Williams AR, et al. (2022-03-24) Ten simple rules for making a software tool workflow-ready. PLoS Comput Biol 18(3): e1009823. https://doi.org/10.1371/journal.pcbi.1009823
- [Presentation] Livermore, Laurence; Scott, Ben; Dillen, Mathias (2021-07-22): Contemporary and Established Provenance Issues in Natural History Collections. figshare. Presentation. https://doi.org/10.6084/m9.figshare.15035370.v1
- [Blog post] https://teklia.com/blog/202104-export-formats (2021-04-01)

5. Discussion and Future Development

While we built and tested several tools, it was challenging to package them in a user-friendly way that made them amenable for use in configurable workflows.

The containerisation approach, combined with the execution of the tools during each workflow run, meant that there were substantial overheads, including excessive RAM usage and processing overheads. We recommend that most of the tools, except those with the lowest computer requirements, are run as standalone services. This can still support a FAIR workflow-based approach, but would be much more efficient.

One of the limiting factors for tool development was the limited amount of training data. A key recommendation from discussions in SDR closure meetings was to use commercial AI services to assist in the creation of training datasets e.g., identifying regions of interest for DLA, and textlines; or creating verbatim text for manual named entity labelling, especially for entities which are uncommon in existing training corpora.

The SDR was impacted by a common drawback of data refinery platforms: the need to implement each task as a standalone tool requiring substantial development overhead. Some tools, for example the Barcode tool, could be used in a few lines of code, but our implementation required five files in XML and Python to integrate with our workflow model. The required overengineering added substantial development time to the project.

Furthermore, user testing of our prototype platform demonstrated the limited need for user-customisable workflows. None of our beta-testing user groups tried to implement a custom workflow; they all just utilised the out-of-the-box, end-to-end specimen extraction workflow. Providing user-customisable workflows does not just add back-end engineering complexity, it invalidates the key principle of front-end design: don't make the user think. An alternative paradigm is to use a simple service to extract all data from a specimen image, with any non-core requirements implemented as auxiliary external processes. For example, the use-case of revalidating existing datasets. Focussing on the one core need allows us to greatly simplify the implementation. Subsequent development as part of DiSSCo will focus on extracting core tools from the SDR and implementing them as streamlined services, embedded in the DiSSCo ecosystem.

6. Code Repository & Related Issues

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

Summary of development work that contributed to Deliverable 8.2:
https://github.com/DiSSCo/SDR/issues/77

7. Appendix

7.1 Proposed JSON for representing textual label data

See [Bonhomme, 2022] for more information/


```
{
 "@id": "some_page_id",
 "name": "PAGE_NAME",
 "image": {
 "@id": "some_image_id",
 "url": "https://iiif.teklia.com/main/iiif/2/some_image.jpg",
 "width": 1500,
 "height": 2318
 },
 "metadata": [],
 "subelements": [
 {
 "@id": "some_text_line_id",
 "name": "text_line_name",
 "type": "text_line",
 "zone": {
 "polygon": [[12,9],[12,42],[226,42],[226,9],[12,9]],
 "image": {
 "@id": "some_image_id",
 "url": "https://iiif.teklia.com/main/iiif/2/some_image.jpg"
 }
 },
 "transcriptions": [
 {
 "@id": "some_transcription_id",
 "text": "THE NATURAL HISTORY MUSEUM, LONDON",
 "confidence": 1,
 "source": {
 "worker_slug": "manual",
 "worker_version": null
 },
 "entities": [
 {
 "text": "THE NATURAL HISTORY MUSEUM, LONDON",
 "@id": "http://dbpedia.org/resource/Natural_History_Museum,_London",
 "type": "organization",
```

"offset": 0,
"length": 34,
"source": {
"worker_slug": "manual",
"worker_version": null
}
}
]
}
],
"metadata": [
{
"type": "text",
"name": "script_type",
"value": "typewritten"
}
]
}
]
}

8. References

[Brack et al, 2022] 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 Computational Biology 18(3): e1009823. https://doi.org/10.1371/journal.pcbi.1009823

[Blagoderov et al 2012] Blagoderov V, Kitching I, Livermore L, Simonsen T, Smith V (2012) No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209: 133-146. https://doi.org/10.3897/zookeys.209.3178

[Bonhomme, 2022] Bonhomme, M.-L. What is the best export format for handwritten document processing results? (2022-04-01) https://teklia.com/blog/202104-export-formats/

[de la Hidalga et al, 2022] de la Hidalga, A.N., Rosin, P.L., Sun, X. et al. Cross-validation of a semantic segmentation network for natural history collection specimens. Machine Vision and Applications 33, 39 (2022). https://doi.org/10.1007/s00138-022-01276-z

[Groom et al, 2022] 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

[Hardisty et al, 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

[Hardisty et al, 2020] Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul DL, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure -

DELIVERABLE D8.1. Research Ideas and Outcomes 6: e54280. https://doi.org/10.3897/rio.6.e54280

[Haston et al, 2015] Haston, E, Albenga, L, Chagnoux, S, et al. Automating data capture from natural history specimens. September 18, 2015. https://synthesys3.myspecies.info/node/695

[Hudson et al, 2015] Hudson LN, Blagoderov V, Heaton A, Holtzhausen P, Livermore L, et al. (2015) Inselect: Automating the Digitization of Natural History Collections. PLOS ONE 10(11): e0143402. https://doi.org/10.1371/journal.pone.0143402

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

[The 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

[Livermore et al, 2023a] Livermore, L., Goble, C., Hardy, H., Scott, B., Soiland-Reyes, S., Woolland, O. Deliverable 8.3 - Specimen Data Refinery: Development of cloud platform for data-processing services. 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. Deliverable 8.4 - Specimen Data Refinery: Usage - data exploitation, evaluation and dissemination. August 2023.

[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 (2020a) Landscape Analysis for the Specimen Data Refinery. Research Ideas and Outcomes 6: e57602. https://doi.org/10.3897/rio.6.e57602

[Walton et al (2020b)] Walton S, Livermore L, Dillen M, De Smedt S, Groom Q, Koivunen A, Phillips S (2020b) A cost analysis of transcription systems. Research Ideas and Outcomes 6: e56211. https://doi.org/10.3897/rio.6.e56211