SYNTHESYS+
Synthesis of Systematic Resources a DiSSCo project

LANDSCAPE ANALYSIS FOR THE SPECIMEN DATA REFINERY

Authors: Stephanie Walton, Laurence Livermore & Carole Goble

Data Contributors: Olaf Banki, Robert Cubey, Mathias Dillen, Robyn Drinkwater, Markus Englund, Quentin Groom, Elspeth Haston, Mattias Obst, Mario Lasseck, Laurence Livermore, Sarah Phillips Isabel Rey, Dominik Roepert, Celia Santos, Stephanie Walton, Alan Williams

Grant Agreement Number | **823827**
Acronym | **SYNTHESYS PLUS**
Call | **H2020-INFRAIA-2018-2020**
Start date | **01/02/2019**
Duration | **48 months**
Work Package | **8**
Work Package Lead | **Laurence Livermore**
Delivery date | **21-02-2020**

Contents

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+
Synthesis of Systematic Resources a DiSSCo project

Introduction

A key limiting factor in organising and using information from global NH specimens is making that information computable. More than 95% of available information currently resides on labels attached to specimens or in physical registers. Institutional digitisation pipelines have tended to focus more on the specimens themselves than on efficiently capturing computable data about them. SYNTHESYS+ will address this gap using technologies developed to harvest, organise, analyse and enhance information from other sources (such as books, photographs and maps), offering the prospect of greatly accelerated data capture.

The objective of the Specimen Data Refinery (SDR) is to combine these technologies into a cloud-based platform for processing specimen images and their labels *en masse* in order to extract essential data efficiently and according to standard best practices.

A workflow was developed to illustrate the various steps required to fully automate the process from image capture to a full specimen dataset *(Image 1)*. There are two core components of building a workflow that must be considered. First, the tools available to complete the individual tasks required, such as tools that can execute image segmentation or tools that can conduct automated text extraction. Research and development has been conducted to varying degrees on tools and methods for executing these steps. Most of this research and development has been conducted in isolation, addressing one step in the process but not the workflow in its entirety. In developing a Specimen Data Refinery, there are opportunities to take advantage of pre-existing research and development on some tools, but it will also encounter significant gaps in others.

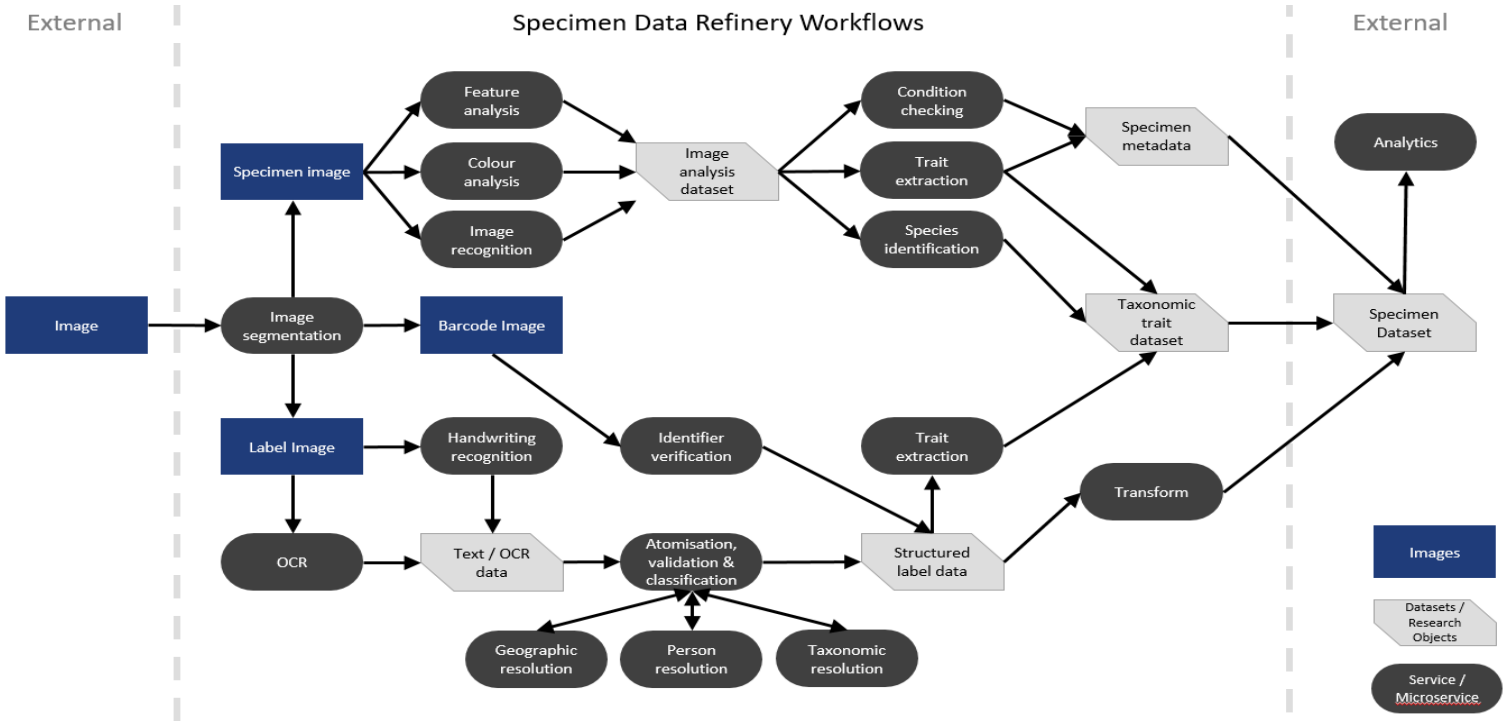

Image 1: Specimen Data Refinery Workflow

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

The second component of building an automated workflow is developing the links between each tool - the environment in which the entire process is executed and the technology that executes the process. This is a different set of platforms and services that will connect what are currently various disparate pieces into a whole working system. It requires a technology stack that is reliable, sustainable and cost-effective.

In order to assess the state of each phase in the workflow, a gap analysis on available tools for each step was conducted to discover where a Specimen Data Refinery might readily move forward and where considerable hurdles can be expected (Section 3). Then an initial assessment was conducted on the technology stack required to assemble these tools together into an automated workflow (Section 4).

1.1 Scope

The scope of this initial landscape analysis (Task 8.1) is to evaluate existing platforms based on their approach and service offering and to identify sources of data including reference/ground truth/training datasets. It will also identify any missing tools/service and datasets.

This report does not include: evaluation of existing tools, service registries and platform-based approaches; assessment for the potential to use pan-European Collaborative Data Infrastructure; creation of reference/ground truth/training datasets. This is planned in the subsequent follow-on Task 8.2.

1.1.1 Machine Learning and Training Data Sets

The tools in this landscape analysis include both unsupervised and supervised math-based and machine learning resources. For example, the image segmentation tools are unsupervised math-based whereby their methods for identifying parts of an image include thresholding, contouring, clustering, etc and how it segments an image does not change no matter the number of images processed. In comparison, Google Vision uses supervised learning, requiring a 'training period' for image recognition when it is 'taught' to identify specific items in an image based on a ground-truth set of images. The more images it processes, the more accurate its recognition capabilities should become.

Many of the machine learning tools included in this study are specific to natural history collections and, in many cases, are designed for specific taxons. Thus these tools have been trained and tested with species datasets. In order to gain a comprehensive picture of the tools available, items have also been included that were not designed specifically for scientific collections and have undergone limited or no testing on natural history collections.

It was not within the scope of this deliverable to develop new training data sets but, in identifying tools that have not been tested in a natural history context, to identify where the development of new training datasets needs to be prioritized.

1.1.2 Prior Research on Automation

A collection of research has already been conducted in the SYNTHESYS3 and ICEDIG projects on the capabilities of automation tools in digitisation. Haston et al. (2015) conducted a series of tests on image segmentation, OCR and handwriting recognition and natural language processing (NLP) for automatic metadata capture. Further research has also been conducted in ICEDIG on label and transcription automation capabilities. Tests were conducted on methods for automated text

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

digitisation for ICEDIG with recommendations on specific workflows and OCR tools (Owen et al., 2018).

1.1.3 Crowdsourcing and Human-in-the-Loop

As the Specimen Data Refinery is intended to integrate both artificial intelligence (AI) and human-in-the-loop (HitL) approaches to extraction and annotation, citizen science platforms such as plant identification apps and volunteer transcription services were included in the initial research. However, the primary focus of this landscape analysis is on AI platforms as these hold the greatest potential for mass efficiency gains and centralised workflows.

Methodology

In order to collect an aggregated list, the SYNTHESYS+ partners from partner institutions were invited to contribute known tools, methods, resources and pilot projects [see supplementary file]. Over the course of six months, various people added to the list, made updates, cited sources and contributed new tools. Each tool was categorised based on their place in the data refinery workflow.

Where available, the data added for each tool included:

- Brief service description
- Delivery platform (eg. web application, software library, R package, etc.)
- Associated academic papers
- Known test pilots
- Cost (where applicable)
- Input/Output formats
- License

In total, 76 tools, methods and resources were collected (Appendix).

After the aggregation phase was complete, the list was reviewed in its entirety. Each tool and resource was mapped onto the data refinery workflow in order to assess where reusable resources are available and where there are major gaps or potential risks. Each step in the workflow was graded according to a traffic-light system - green for the existence of a variety of resources that could be repurposed, amber for the existence of resources with limited reuse potential, and red for a major gap where either no resource exists or there is no reuse potential. A number of steps in the workflow (identifier verification, trait extraction, transform and analytics) had no associated tools submitted and were marked as grey in the workflow map. The workflow map was then distributed to the contributing partners to identify any further gaps or missing areas.

Upon completion of the gap analysis, an initial assessment was conducted on the technology stack available to compile each of the tools together into a workflow. A high-level consultation was conducted with a computer science team at a partner institution with prior experience developing similar complex human-in-the-loop workflows. Their recommendations have been documented for further study and research in the next phase.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Gap Analysis on Tools & Services

This analysis revealed that there are some areas where considerable efforts have been put towards developing a toolkit while others have received less efforts *(Image 2)*.

3.1 Image Segmentation (green)

Image segmentation involves dividing an image into its component parts, such as separating the specimen itself from the barcode and the label. This crucial step allows for deeper analysis into each component piece. In addition to a large suite of tools available for batch photo editing (cropping, resizing, rotating, etc.), there were three reported tools that could segment an image. scikit-image (Pandey, 2019) is a Python package with a suite of methods for segmenting an image including thresholding, active contouring, random walkers, etc. ImageSURF is a JAVA API and FIJI plugin that segments based on nearest-neighbour colour annotations. OpenCV, an open source computer vision and machine learning software library, provides algorithms to segment images for different programming languages, which is also a useful tool for image recognition, can segment images as well.

Semantic segmentation is another method of image segmentation that is currently being developed and tested on herbarium collections. YOLO V3 has been tested for identifying between the different items that are commonly found on a herbarium sheets - the plant specimen, scale bar, stamp, color pallet, specimen label, envelope and bar-code (Triki et al., forthcoming). Semantic segmentation was also used in another study based on a dataset of 400 images of ferns to train a deep learning algorithm to segment the image of the specimen from the image background (White et al., 2019). These approaches could be adapted and reused for general herbarium sheets and generalised for use with other specimen images.

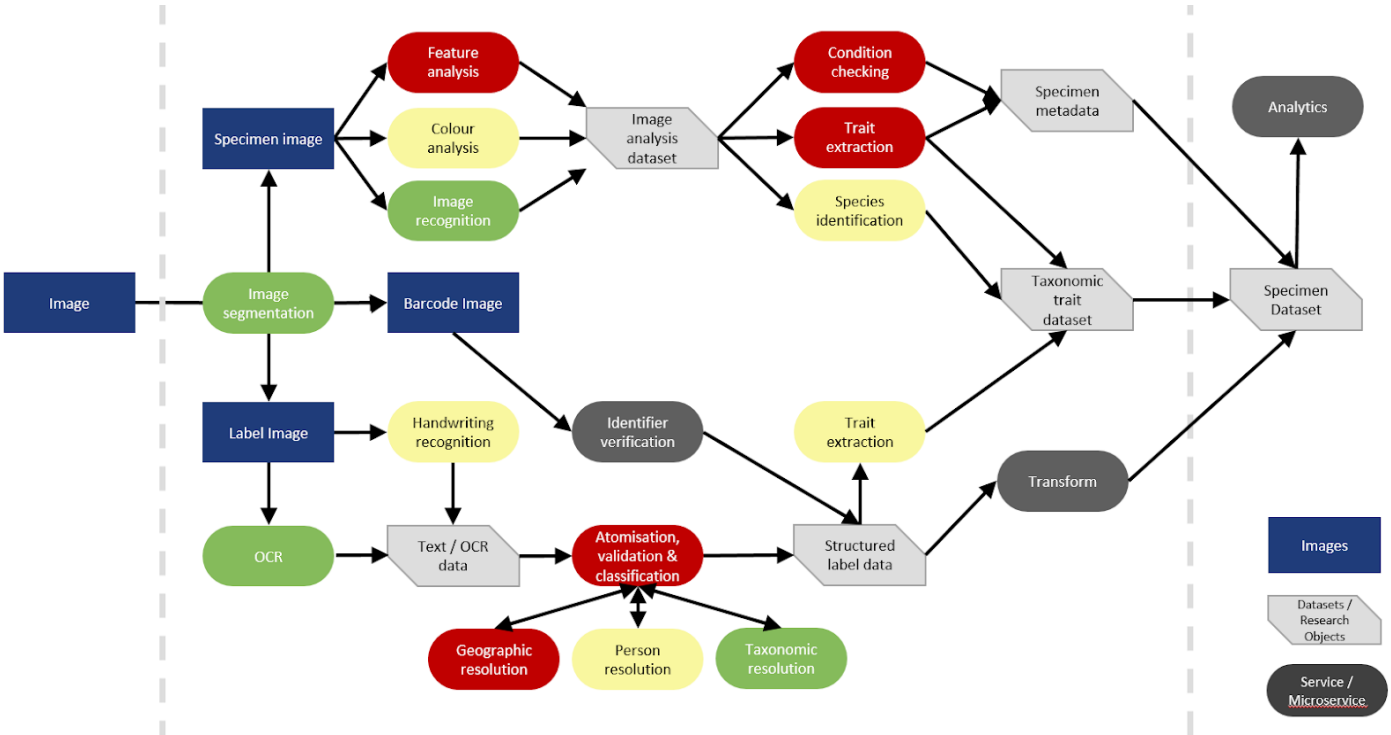

Image 2: Traffic-light results of gap analysis

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

The step was marked as green because there is more than one tool available and each tool provides a different method for going about image segmentation, thus offering a variety of options that could be tested based on the needs of the collection's images. More importantly, scikit-image underwent significant testing by the Natural History Museum as part of the SYNTHESYS3 work package (Haston et al., 2015) and YOLO V3 has been trained on natural history collections by a group of universities and returned accurate results (Triki et al., forthcoming).

3.2 Feature Analysis (red), Colour Analysis (amber) and Image Recognition (green)

In the aggregation process, many tools were listed as feature analysis resources but were ultimately categorised as species identification tools because they used some level of feature analysis to identify a specimen (primarily plants). The only tool used exclusively for feature analysis was Computer Vision which segments the specimen from the background of the image by identifying its edges and then, for butterflies and moths only, takes measurements of the wings. Feature analysis was marked red because only one tool was available and it is used for primarily only one type of specimen.

Colour analysis was categorised as amber because there is only one tool available. Image Quality Assessment, in addition to predicting the technical quality of the image, can group sets of images together based on similar colours. However, other tools used for image segmentation and recognition, like scikit-image, may be used for this.

Image recognition was categorised as green because there are two well-developed and heavily-supported resources available. Google Vision comes with enterprise-level support and longevity and offers toolkits for both non-coders and programmers. OpenCV has a strong open-source development infrastructure underneath it. Both can be trained to recognise items in an image and organise them into pre-set categories.

3.3 Condition Checking (red), Image Trait Extraction (red) and Species Identification (amber)

No tools or resources were submitted for condition checking and this appears to be a major gap in the workflow.

A majority of the image trait extraction tools and resources developed have been for biomedical/epidemiological purposes. Trait extraction was marked as red because only two tools were submitted and both are applicable only to plants. Plant Trait Extraction is capable of phenotypic trait extraction but only for a subset of collections (Jin et al., 2018) and traitEx can take measurements but only of leaves (Gaikwad, 2019).

Species identification, in contrast, has received a tremendous amount of concerted effort, research and R&D. As a result, numerous tools and methods have been developed spanning the range of neural network machine learning tools (Wu et al., 2007) to citizen science photo apps. However, it was still marked as amber because a majority of those submitted are either methods that have only been discussed in research papers (Novotny, 2013; Jamil et al., 2015; Munisami, 2015; Şekeroğlu, 2016; Lasseck, 2017; Xi et al., 2019) or apps for which data and machine learning quality require further analysis. Three out of the remaining four existing tools are related only to plant

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

identification. So while there is a strong foundation of methodologies from which to build on, species identification will still require considerable input.

3.4 Handwriting Recognition (amber) and OCR (green)

These two areas have also been the recipients of considerable research and development. While Transkribus is the only listed tool available for handwritten text transcription, it is supported by EU funding and has been successfully deployed on a collection of specimens from the Natural History Museum Edinburgh. Transkribus also offers a host of web and cloud services.

OCR was marked green as there are multiple tools available, although ABBY is an enterprise-level software that will likely have cost associated. The Natural History Museum has tested Tesseract OCR against the Biodiversity Heritage Library (BHL) corpus and achieved comparable results to BHL's OCR engine (powered by ABBYY FineReader). While Tesseract is also capable of handwritten text recognition, accuracy with serif and cursive text was poor. Tesseract OCR has been tested in large scale on the herbarium sheet images in EUDAT pilot Herbadrop project. Google Vision also provides OCR services. Several other tools, including langid.py and Stanford NER were tested as part of the ICEDIG work package (Owen et al., 2018).

3.5 Atomization, validation and classification (amber)

Many OCR tools are capable of named entity recognition or the ability to extract strings of text and thereby break a label into its component parts such as designate between place names, person names or taxon names. The main tools - NLTK, spaCy and Stanford NER - are capable of deep learning so can be trained to recognise specific strings and categories by being trained with a ground truth dataset. These tools have been used to derive structured data from taxonomic publications (e.g. traits) but still require further research in the context of natural history collection labels. There are also a couple of tools available for extracting ecologically-relevant terms from a label. ClearEarth and Taxon Concepts are both capable of identifying such terms and categorising.

There are also several language detection tools available. However, there is still considerable work to be done on a more efficient method for segmenting out the different identifiers on a label for further classification.

3.6 Geographic Resolution (red), Person Resolution (amber) and Taxonomic Resolution (green)

Geographic resolution is a task natural history collections have struggled to automate. There are numerous tools available for general geocoding - MapQuest Geocoding, Google Geocoding, CartoDB, Pelias. However, these tools require a known address, city, country or region name in order to identify an associated latitude and longitude. They are not designed for historical place names and cannot accommodate changing boundaries over time or vague or general place descriptions. GEOLocate is the only tool listed that is designed specifically to assist in the geographic resolution of natural history collections that is currently still active. Several other tools like BioGeomancer (Guralnick, 2006) and R packages like R BIOgeo and R GeoNames have also been developed but are outdated and no longer unavailable. In the case of BioGeomancer, the code has not been active since 2012. Further research and resources will be necessary to develop this part of the workflow.

Person resolution was marked as amber because Bloodhound is currently the only tool designed specifically to match a collector with the specimens they collected. Numerous efforts are also underway to assign unique person identifiers to researchers, present-day and historical. ORCID, ISNI

and ResearcherID have databases of person identification numbers and VIAF combines the person with numerous countries' national libraries into an aggregated database. In relation to published academic papers, Elsevier assigns a researcher ID for all authors in its database through Scopus and there are a number of sites to which researchers can upload their publishing profile.

MNHN is currently developing a Person Refinery, expected to be completed by April 2020 which has revealed a number of challenges in efficiently developing data structures and alignments for person resolution, chief of which is how the various researcher ID systems can help disambiguate people collections and whether there is particular people identifier system which will prove to be most relevant for all types of collections

Taxonomic resolution is the most developed. The Catalogue of Life is an authoritative global species checklist for all life on earth, that is built on 172 global taxonomic resources (Catalogue of Life, 2019 Annual Checklist). GBIF has aggregated the taxonomic databases of numerous sources, taking the Catalogue of Life as a starting point, for a single entry-point for taxonomic synonym identification and name resolution for living specimens. The NHM has developed a java-based ETL process that utilises the GBIF taxonomic backbone to resolve names, while still allowing colleagues with taxonomic expertise to validate results and adjust certain query parameters. The Netherlands Biodiversity Data Services developed and maintained by Naturalis Biodiversity Center is making use of the Catalogue of Life in addition to the Netherlands Species Register to validate names of biological collections. GBIF and the Catalogue of Life are constructing a joint infrastructure for names and taxonomy, that will include an extended Catalogue of Life as the replacement of the GBIF Backbone Taxonomy. In addition to this resource, Fossilworks is available as a taxonomic database for historic specimens and there are numerous other databases available specifically for plants, mammals or other taxons for further identification. In addition to these databases, there are also a number of out-of-the-box tools for synonym identification and resolution. Taxize is an R package developed specifically for this purpose (Chamberlain & Szöcs, 2013) as well as Taxosaurus, a thesaurus for taxonomy names, along with several other resources.

Taxonomic resolution is marked green as there are several tools and resources available. However many of the data sources such as NCBI and ITS offer different classifications and levels of scientific name resolution resulting in a scattered and blurry landscape for users. When these services are combined in platforms like GBIF, these differences are not currently resolved and it is up to users to do so. Work is underway to develop a joint infrastructure, but these discrepancies should be kept in mind in the meantime.

3.7 Label (Biological) Trait Extraction (amber)

Biological trait extraction has been largely confined to literature (Endara et al., 2018; Gaikwad et al., 2019; Jin et al., 2018; Thessen et al., 2018). However, while infrequent, a small number of specimen labels may include trait descriptions. There has been a considerable amount of research and development on semantic machine-learning software for extracting trait descriptions for large sources of text, some of which may be applied to label text. This category has been marked amber because of three tools specific to ecological/biodiversity terms that may be utilised or repurposed for specimen labels. ClearEarth (Thessen et al., 2018) and Explorer of Taxon Concepts (Endara et al., 2018) can extract ecologically-relevant terms from text for further study. Phenoscrape and the associated SCATE project connect trait analysis tools to semantic reasoning tools.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

Building a Workflow

The Specimen Data Refinery (SDR) aims to take the tools identified above and package them into a cohesive workflow for processing and analysis. This requires a technology stack that will create the links between different tools and the operating environment in which the workflow is executed and managed. While there are many different technology services available for workflow development, the priority for the SDR will be identifying a technology stack that contains all of the required functionality while being reliable, sustainable and cost-effective.

4.1 Selecting a Human-in-the-Loop Workflow Management Systems (WfMS)

There are many examples in bioinformatics of automated workflows that string together a collection of tools and execute a series of steps with no intervention required by a user (Perkel, 2019). A Workflow Management System (WfMS) is the software that strings the tools together. It designs, executes and monitors a workflow while shielding users from underlying executional complexities. It manages code and data access and movement, logging, errors, parameter configurations and data provenance (where, when and with what parameters and inputs a task was run) among other tasks (Cohen-Boulakia et al., 2017; Deelman et al., 2018).

There are currently over 266 Workflow Management Systems, each with its own strengths and weaknesses. Typically, they vary on whether they are focused on either linking tools or linking infrastructure layers, whether they are domain-specific of general and who they target as their user-base and the level of expertise required. It is not necessary, however, to only choose one. Workflow management systems can be combined to develop custom solutions.

In addition to these considerations, the SDR has an added layer of complexity because it will require human interaction and decision-making at various steps in the process. For example, the workflow could execute the steps to get an image to the point where it is ready to be georeferenced, but a user may need to select which type of georeferencing algorithm is most appropriate for the label based on the locality information within it. This is called a human-in-the-loop (HitL) workflow.

Therefore, the environment within which the workflow is executed must be interactive, providing a space in which a user can give commands that then dictate the next steps of the workflow. Similar HitL workflows have been developed for other biodiversity projects (Mathew et al., 2014) and there are technology services to facilitate this type of interaction, such as OpenRefine which includes functionality for recording human interactions so they can be repeated in future runs of the workflow.

Galaxy is another WfMS designed specifically for bioinformatics that offers HitL functionality. It has been adopted by EOSCLife, a cluster of 13 research infrastructures, to develop the SEEK Platform. Galaxy is also used by the IBISBA1.0 project, part of IBISBA-EU.

4.2 Implementing a standardised workflow language for interoperability

The steps of a workflow (scripts, tools, command-line tools and workflows themselves) are linked together and executed by the workflow engine within the WfMS. Linking all of these disparate interfaces, scripts, methods and datasets together requires each step to be in the same language so that they can communicate consistently with each other.

Different WfMS typically have different language requirements and protocols and limit interoperability. Several attempts have been made to standardise workflow descriptions and enable

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

workflow interoperability between different systems in order to support the long-term preservation of workflows that may outlive any specific WfMS. The Workflow Description Language and the Common Workflow Language (CWL) (Amstutz et al., 2016; Khan et al., 2019) are recent community efforts to implement a standard language. OpenAPI and the use of APIs for task execution (e.g. GA4GH Task Execution API and GA4GH Workflow Execution API) is contributing to standardised communication between interfaces. The EDAM ontology is another step towards standardising descriptions of the inputs and outputs between bioinformatics tools.

Of these contributions, the Common Workflow Language is the best open standard for compiling workflows and describing how to run the command line tools inside them in a way that makes them portable and scalable. It is a WfMS-agnostic common language that developers can use to better document workflows and assist with workflow portability and interoperability when working between different systems. The current CWL Standard (v1.1) provides authoritative documentation of the execution of CWL documents.

CWL is recommended for the SDR. ELIXIR, a sister ESFRI to DISSCo, has invested in the support of CWL and it is used by the EU's BioExcel2 Centre of Excellence for Biomolecular modelling, by the IBISBA ESFRI for Industrial Biotechnology and by the EOSCLife Cluster project. This strong community and financial support for the development of CWL is indicative of its longevity and anticipated sustainability for the SDR.

4.3 Assembling the workflow

Workflows are made up of a collection of metadata and files - test data, example data, validation data, design documents, parameter files, parameter setting files, result files, provenance logs, etc (Khan 2019).

While the Common Workflow Language is the language in which a workflow is written and described, a research object (RO) is a service for packaging the metadata of disparate objects along certain standards and conventions so that the packages can be exported and exchanged between WfMSs with the necessary detail to be reused and reproduced (Belhajjame 2015). RO-Crate is a recently-developed research object that organizes file-based data with its associated metadata in both human and machine readable formats along with the ability to include additional WfMS-specific metadata. The RO-Crate Metadata File contains information about the dataset as a whole and, optionally, about some or all of its files. This provides a simple way to, for example, assert the authors (e.g. people, organizations) of the workflow or one its files or to capture more complex provenance for files such as how they were created.

Along with the CWL and Galaxy, RO-Crate has been adopted by EOSCLife and IBISBA as the service for describing and packaging workflows and their related files. Based on this community and financial support for these capabilities, a number of WfMSs, including Galaxy, will support CWL and ROCrate.

4.4 The Specimen Data Refinery tech stack

Executing the SDR workflow will require a foundational tech stack and infrastructure for two core pieces - a registry and a run platform.

A registry is a library of workflows. All of the tools and steps in the workflow will be comprised of smaller sub-steps and sub-workflows that make up the building blocks of the entire engine. These building blocks will be housed in a registry built for the SDR. WorkflowHub is a workflow library currently under development for EOSCLife and IBISBAHub for IBISBA workflows and is the underlying

platform for both of these. Hubs can also be utilized for the SDR. It will describe and store the SDR tools and steps in such a way that they satisfy FAIR principles and so that end users understand the workflows data provenance and quality (Goble et al., 2020).

A run platform is the technology stack that will pull these tools, services and processes together. Along with a variety of other services, the recommended SDR run platform (Image 3) can utilise services like Galaxy, CWL, RO-Crate and Workflow Hub that are currently supported by other ESFRI initiatives like EOSCLife and IBISBA. Further research is required to identify the best partners for other components like data storage (e.g. AWS).

Image 3: The recommended workflow technology stack for the SDR

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Conclusion

This gap analysis has made apparent which categories of tools and resources have been specifically developed for specimen images or can be readily generalised and potentially used. Image segmentation, OCR and taxonomic resolution have a broad range of existing and well-tested approaches. Other areas such as visual trait extraction or text processing tools to convert “strings to things” are lacking. There are some general tools and commercial services which deal with contemporary languages but Latin and Greek are commonly encountered in scientific names, in diagnostic descriptions (especially botanical descriptions) and as abbreviations on labels such as “cf.” (*confer*). Other potential issues that are yet to be tested or understood in scope are: the frequency of co-occurring languages on labels; the frequency of differing co-occurring hands on labels; and how challenging the abbreviated technical writing style of labels is compared to natural language documents.

Many of the tools and services will require initial or further testing and analysis with training datasets that are domain-specific to natural history collections in order to assess their quality and accuracy. For example, BGM recently undertook an image recognition pilot with Google Vision to extract label information but the results have yet to be analysed for accuracy (ICEDIG D4.4, forthcoming). Transkribus, a handwriting recognition tool capable of deep learning on new handwriting, has undergone one test with herbarium sheets but would need to undergo more rigorous testing (Haston et al., 2015). Named entity recognition tools like spaCy will need to be tested specifically with natural history collection labels.

While there are a broad selection of taxonomic name resolution tools and services, many of which are incorporated into GBIF’s name backbone (GBIF Secretariat, 2019), there are still conflicts and ambiguities that make it hard for end users. This includes limited adoption of synonyms recorded in Catalogue of Life being distinguished by other services

We expect to develop training datasets for the following components of the SDR workflow:

- Image segmentation
- Image recognition
- Feature analysis
- Trait extraction
- Condition checking
- Species identification
- Atomisation, validation and classification
- Person and geographic resolution

However, the development of ground-truth training data sets requires considerable time and resources (Dillen, 2019). GBIF could serve as a general source for training datasets, particularly for geographic resolution, but there are many Darwin Core terms that lack consistent community use of identifiers. This includes, but is not limited to, the terms covering: people (recordedBy, identifiedBy, georeferencedBy), protocols (georeferenceProtocol, measurementMethod, measurementUnit) and location data (higherGeographyID, waterBody, island, locality) which make it harder to develop tools to resolve strings, fix ambiguities and link data. While the biodiversity and natural history data community are discussing how to better implement identifiers, they have yet to reach a consensus. There are also verbatim terms in Darwin Core standards, making it difficult for the machine to interpret the data. A lack of identifier adoption also causes problems for tracking data provenance, an aspect that we have not addressed in this report but is crucial to technical implementation and for the required metadata about digital specimens - this includes information about hardware used

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

and people involved in the process of creating digital specimens. Inconsistent recording and use of image metadata by institutes will also be a challenge - the implementation of image metadata in DarwinCore is minimal, however there is an extension (Audubon Media Description) but we have yet to assess its usage or suitability.

Previous projects to develop toolsets or platforms, like BioGeomancer, have suffered from sustainability issues after project funding ceased. Tools and datasets developed in the next phase of SDR work should prioritise software sustainability. Considerations for sustainability include making use of existing standards, service/tool documentation, and having a maintenance plan - these are summarised in detail by the Software Sustainability Institute. In terms of workflow platform sustainability, we should use a pre-vetted platform, ideally with hosting support, that makes use of existing European investment and prior efforts in training, notably in the ESFRI Cluster EOSCLife and the ESFRI IBISBA.

While these complexities and hurdles need to be taken into consideration in developing the SDR, this analysis also revealed there is a considerable about of open-source technology available and research that has already been conducted into automating these processes. There is significant opportunity to take advantage of this research by combining it into a workflow that will greatly improve the efficiency and scalability of NH digitisation efforts.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Glossary

Active Contouring: a method of image segmentation that identifies object contours in an image in order to detect outlines

AWS: Amazon Website Service

ETL: Extract, transform, load

GBIF: Global Biodiversity Information Facility (https://www.gbif.org/)

Google Vision: a machine learning tool for automated image recognition and categorisation (https://cloud.google.com/vision)

Ground truth data: a dataset comprised of information acquired through direct observation rather than through inference or automation

Hands: handwritten script attributable to an individual/individuals

HitL: Human-in-the-loop

ICEDIG: Innovation and consolidation for large scale digitisation of natural heritage (https://www.icedig.eu/)

Metadata: a set of data that describes and gives information about other data, such as the file format of timestamp of an image or the provenance and processing inputs of a data run

Neural network: a set of algorithms that are designed to recognize patterns and connections through training on a dataset (see training dataset)

NLP: natural language processing

OCR: optical character recognition

Reference datasets: data that sets standards to which the fields in other datasets adhere

RO: research object

SDR: Specimen Data Refinery

SEEK: a digital object management and cataloguing platform that underpins the Workflow Hub and IBISBAHub.

Thresholding: a method for segmenting an image by converting a colour image to grayscale and then filtering out pixels that are above a certain setting on the grayscale - a threshold - and maintaining pixels that fall below it

Training datasets: datasets that are used to train a machine learning platform in a particular set of capabilities, for example to identify something in an image

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

WfMS: Workflow Management System

YOLO V3: the third release of "You only look once", an tool for detecting images in an object and segmenting them

References

Amstutz, P., Crusoe, M., Tijanić, N. (editors), Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., Scales, M., Soiland-Reyes, S., Stojanovic, L. (2016): Common Workflow Language, v1.0. Specification, Common Workflow Language working group. https://w3id.org/cwl/v1.0/; http://doi.org/10.6084/m9.figshare.3115156.v2

Belhajjame K, Zhao J, Garijo J, Gamble M, Hettne KM, Palma R, Mina E, Corcho O, Gómez-Pérez JM, Bechhofer S, Klyne G, Goble CA: **Using a suite of ontologies for preserving workflow-centric research objects**. *J. Web Sem*. 32: 16-42 (2015), https://doi.org/10.1016/j.websem.2015.01.003

Chamberlain SA and Szöcs E. (2013). taxize: taxonomic search and retrieval in R. *F1000Research*. 2:191 https://doi.org/10.12688/f1000research.2-191.v2

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Lemoine, Y., Mareuil, F., Ménager, H., Pradal, C., Blanche, C. (2017). Workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*. 75: 284-298. https://doi.org/10.1016/j.future.2017.01.012

Dillen, M., Groom, Q., Chagnoux, S., Güntsch, A., Hardisty, A., Haston, E., Livermore, L., Runnel, V., Schulman, L., Willemse, L., Wu, Zhengzhe, Phillips, S. (2019). A benchmark dataset of herbarium specimen images with label data. *Biodiversity Data Journal*. 7:e31817. https://doi.org/10.3897/BDJ.7.e31817

Deelman, E., Peterka, T., Altintas, I., Carothers, C.D., Van Dam, K.K., Moreland, K., Parashar, M., Ramakrishnan, L., Taufer, M., Vetter, J.S. (2017). The Future of Scientific Workflows. *The International Journal of High Performance Computing Applications*. 32(1), 159-175. https://doi.org/10.1177/1094342017704893

Gaikwad, J., Triki, A., Bouaziz, B. (2019). Measuring Morphological Functional Leaf Traits from Digitized Herbarium Specimens Using TraitEx Software. *Biodiversity Information Science Standards.* 3:e37091. https://doi.org/10.3897/biss.3.37091

Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M., Peters, K., Schober, D. (2020). FAIR Computational Workflows. *Data Intelligence*. https://doi.org/10.1162/dint_a_00033

Endara, L., Cui, H., Burleigh, G. (2018). Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. *Applications in Plant Sciences*. 6:3. https://doi.org/10.1002/aps3.1035

Gaikwad, J., Triki, A., Bouaziz, B. (2019). Measuring Morphological Functional Leaf Traits from Digitized Herbarium Specimens Using TraitEx Software. *Biodiversity Information Science and Standards*. 3:e37091. https://doi.org/10.3897/biss.3.37091

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

GBIF Secretariat (2019). GBIF Backbone Taxonomy. Checklist dataset https://doi.org/10.15468/39omei accessed via GBIF.org on 2020-02-19.

Guralnick, R., Wieczorek, J., Beaman, R., Hijmans, R., BioGeomancer Working Group. (2006). BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data. *Biodiversity Data*. PLoS Bio 4(11): e381. https://doi.org/10.1371/journal.pbio.0040381

Haston, E., Albenga, L., Chagnoux, S., Drinkwater, R., Durrant, J., Gilbert, E., Glöckler, F., Green, L., Harris, D., Holetschek, J., Hudson, L., Kahle, P., King, S., Kirchhoff, A., Kroupa, A., Kvacek, J., Le Bras, G., Livermore, L., Mühlenberger, G., Paul, D., Phillips, S., Smirnova, L., Vacek, F., Walker, S. (2015). Optimal automated metadata capture. *SYNTHESYS3: Synthesis of systematic resources*.

Jamil, N., Hussin, N.A.C., Nordin, S., Awang, K. (2015). Automatic Plant Identification: Is Shape the Key Feature? *Procedia Computer Science*. 76, 4376-442. https://doi.org/10.1016/j.procs.2015.12.287

Jin, Shichao & Su, Yanjun & Wu, Fangfang & Pang, Shuxin & Gao, Shang & Hu, Tianyu & Liu, Jin & Guo, Qinghua. (2018). Stem-Leaf Segmentation and Phenotypic Trait Extraction of Individual Maize Using Terrestrial LiDAR Data. *IEEE Transactions on Geoscience and Remote Sensing*. PP. 1-11. https://doi.org/10.1109/TGRS.2018.2866056

Khan, F. A., Soiland-Reyes, S., Sinnott, R., Lonie, A., Goble, C., Crusoe, M. (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
Lasseck, M. (2017). Image-based Plant Species Identification with Deep Convolutional Neural Networks.

Mathew, C., Guntsch, A., Obst, M., Vicario, S., Haines, R., Williams, A., de Jong, Y., Goble, C. (2014). A semi-automated workflow for biodiversity data retrieval, cleaning and quality control. *Biodiversity Data Journal*. 2:e4221. https://doi.org/10.3897/BDJ.2.e3221

Owen, D., Groom, Q., Hardisty, A., Leegwater, T., van Walsum, M., Wijkamp, N., Spasic, I. (2018) Methods for Automated Text Digitisation. *Zenodo*. https://doi.org/10.5281/zenodo.3364502

Pandey, P. (2019). "Image segmentation using Python's scikit-image module: An overview of scikit-image library's image segmentation methods." *Towards Data Science*. [Online]. https://towardsdatascience.com/image-segmentation-using-pythons-scikit-image-module-533a61ecc980

Perkal, J. (2019). "Workflow systems turn raw data into scientific knowledge." *Nature*. [Online]. https://www.nature.com/articles/d41586-019-02619-z

Munisami, T., Ramsurn, M., Kishnah, S., Pudaruth, S. (2015). Plant Leaf Recognition Using Shape Features and Colour Histogram with K-nearest Neighbour Classifiers. *Procedia Computer Science*. 58, 740-747. https://doi.org/10.1016/j.procs.2015.08.095

Novotný, P., Suk, T. (2013). Leaf recognition of woody species in Central Europe. *Biosystems Engineering*. 115:4, 444-452. https://doi.org/10.1016/j.biosystemseng.2013.04.007

Şekeroğlu, B., İnan, Y. (2016). Leaves Recognition System Using a Neural Network. *Procedia Computer Science.*. 102, 578-582. https://doi.org/10.1016/j.procs.2016.09.445

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

Thessen, Anne & Preciado, Jenette & Jain, Payoj & Martin, James & Palmer, Martha & Bhat, Riyaz. (2018). Automated Trait Extraction using ClearEarth, a Natural Language Processing System for Text Mining in Natural Sciences. *Biodiversity Information Science and Standards*. 2:e26080. https://doi.org/10.3897/biss.2.26080

Triki, A., Bouaziz, B., Mahdi, W., Gaikwad, J. (2020). Objects Detection from digitized herbarium specimen based on improved YOLO V3. *Forthcoming*.

White, A., Dikow, R., Baugh, M., Jenkins, A., Frandsen, P. (2019). Generating Masks for Image SEgmentation in Digitized Herbarium Specimens. *Biodiversity Information Science and Standards*. 3:e37479. https://doi.org/10.3897/biss.3.37479

Wu, S., Bao, F.S., Xu, E.Y., Wang, Y., Chang, Y., Xiang, Q. (2007). A Leaf Recognition Algorithm for Plant Classification using Probabilistic Neural Network. *2007 IEEE International Symposium on Signal Processing and Information Technology*. https://doi.org/10.1109/ISSPIT.2007.4458016

Xi, T., Wang, J., Han, Y., Wang, T., Ji, L. (2019). The Effect of Background On a Deep Learning Model in Identifying Images of Butterfly Species. *Electrical and Electronics Engineering: An International Journal.* 8:1. https://doi.org/10.1481/elelij.2019.8101

Appendix

Aggregated tools, methods and services spreadsheet

https://docs.google.com/spreadsheets/d/1t_E0PG2kuJ_l4ikWdRFEw6uiOsazKuSES4aZlf_1qCc/edit?usp=sharing

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project