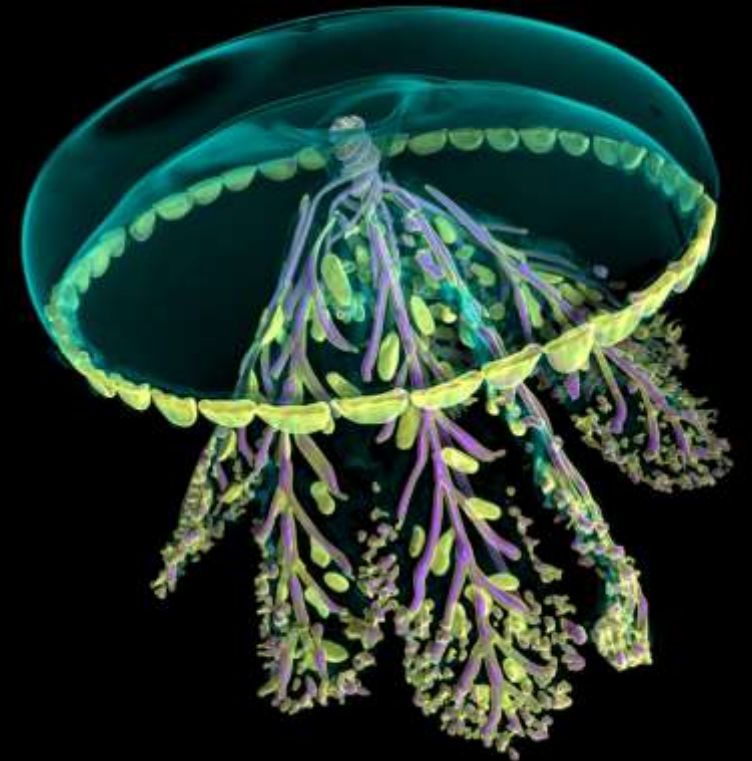


## Work package 8

### JRA3: Specimen Data Refinery



WP Lead: Laurence Livermore (NHM)

Task Leaders: RBGE, NHM, UNIMAN, NBC



[synthesys@nhm.ac.uk](mailto:synthesys@nhm.ac.uk)



<https://on.fb.me/1KrD2Ko>



[@SYNTHESYSEU](https://twitter.com/SYNTHESYSEU)

[www.synthesys.info](http://www.synthesys.info)

Develop a platform that integrates **artificial intelligence** and human-in-the-loop approaches to **extract**, **enhance** and **annotate data** from digital images and records at scale.

SDR service  
development



Packaging of workflows  
and tools

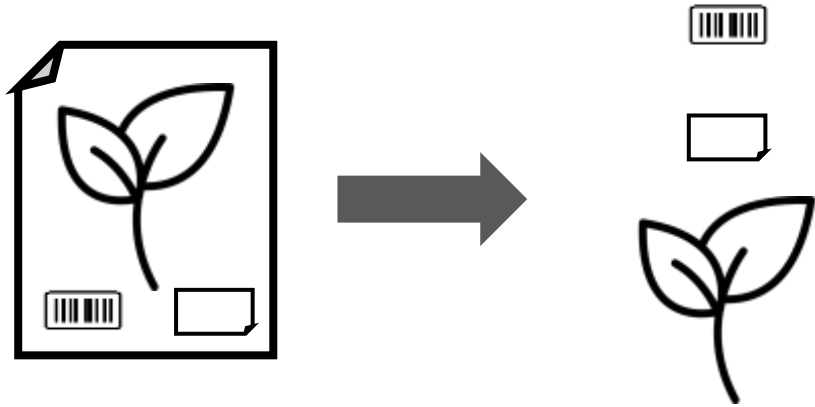


Platform for workflow &  
tool execution

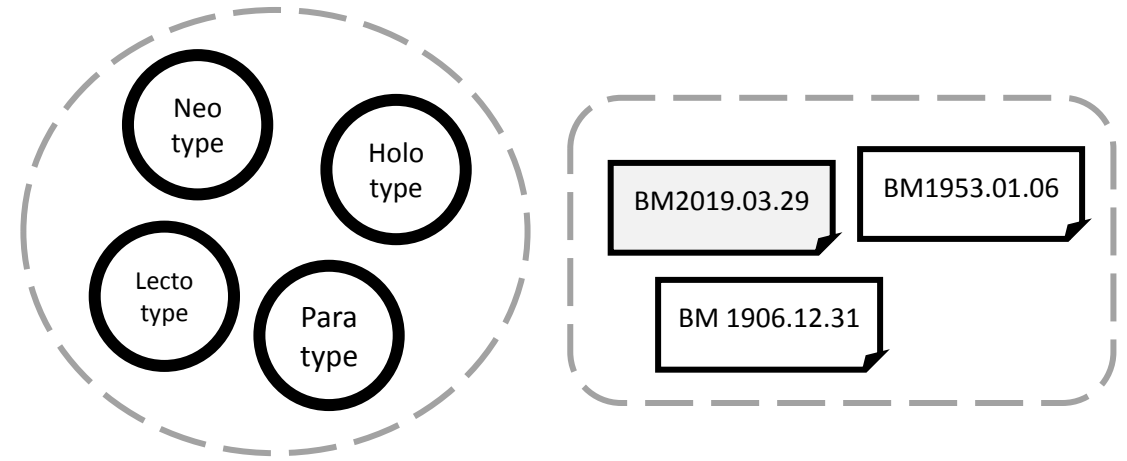


So what will it (hopefully...)  
do?

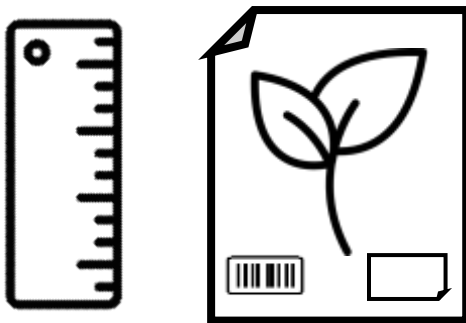
# Allow curators & researchers to create and run repeatable and citable workflows resulting in datasets with rich self-descriptive metadata based on GUIDs and persistent identifiers



Segment and crop parts of images



Group similar specimens and labels  
(based on size, shape, colour, landmarks)



Measure specimens and labels



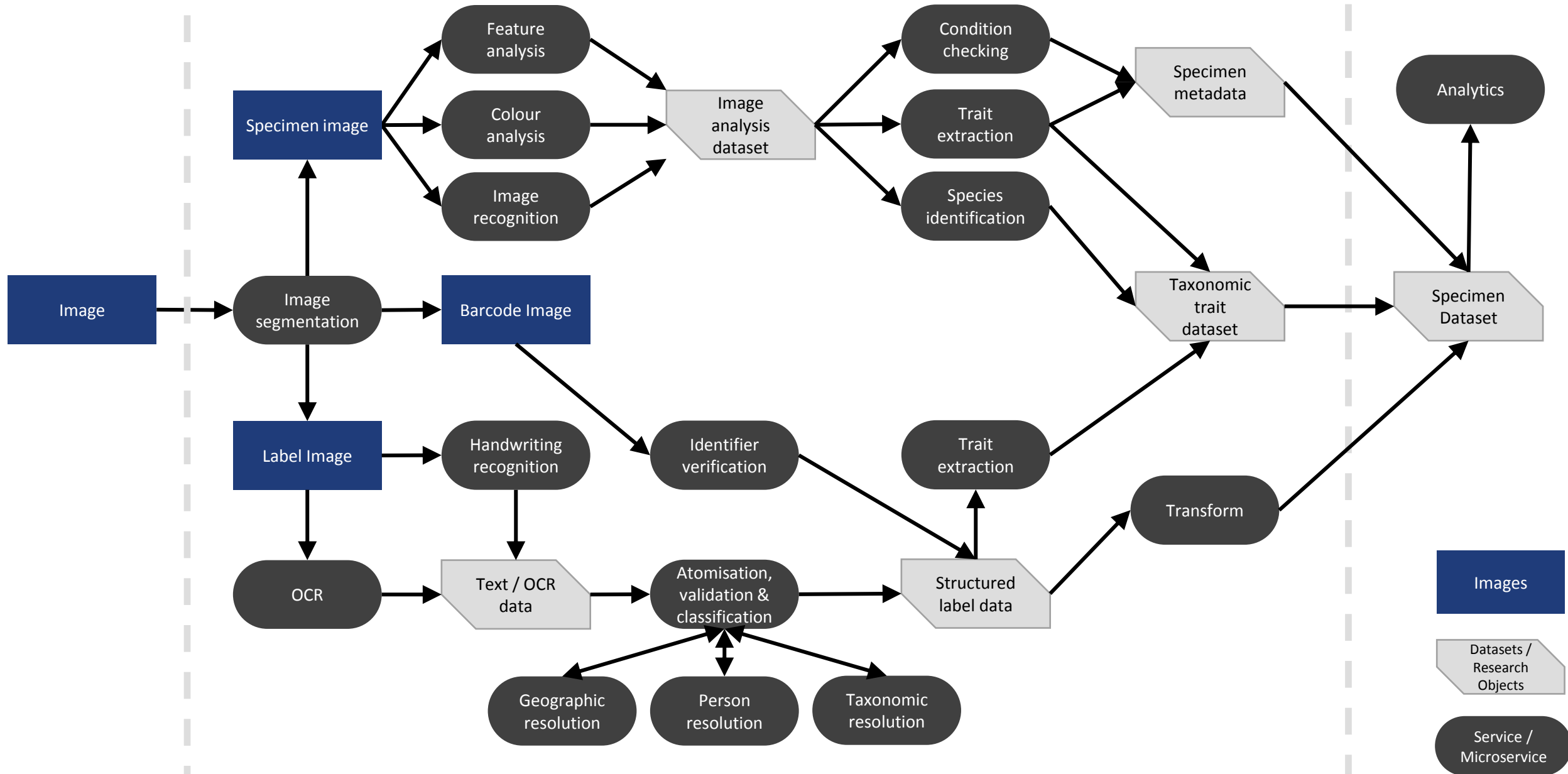
Georeference text

**What might  
workflows look like?**

External

# Specimen Data Refinery Workflows

External



Original diagram by Matt Woodburn – Thanks!



Locality: SITE157761 (*Saint Helena*)

Type: TYPENonType (*Non-type*)

Specimen ID: 01687366

Storage Location: LOC816449 (*Drawer 75*)

Taxonomy: TAX1429066 (*Quadriceps hopkinsi*)



Processed and  
imported into  
institutional  
systems  
(CMS, public  
portal)

Easier

- Condition checking of specimens  
(e.g. gum chloral/phenol balsam discoloration, verdigris, pyrite oxidation)



Microscope  
slide with gum  
chloral  
discoloration

- Natural language descriptions of  
specimens  
(e.g. for public, curators, researchers)



This is a Matchsafe. We acquired it in 1980. It is a part of the Product Design and Decorative Arts department. Its dimensions are Overall: 6.4 cm (2 1/2 in.)

- Taxonomic trait extraction  
(e.g. phenology, morphology, biological relationships)

Harder



## Task 8.1

### Landscape evaluation

- Evaluate platform based approaches inc. tool/service registries
- Identify sources of data
- Identify & create training/reference datasets
- Assess potential to use EUDAT, EGI, EOSC, AWS etc
- Reuse & integrate reports from previous DiSSCo-related work

D8.1 Report (RBGE)

## Task 8.2

### Development of tools, services & workflows

- *Services include:*
- Optical character recognition data capture and analysis
- Data mining and linkage
- Image analysis
- Georeferencing
- Human interaction and crowdsourcing
- Mature services will be containerized and incorporated into SDR workflow & tool registry

D8.2 Demonstrator (NHM)

## Task 8.3

### Development of SDR cloud platform

- Platform will run workflows in SDR registry
- Controlled user access (ELIXIR authentication & authorisation)
- Create detailed standardized provenance metadata of derivative datasets
- Derivative datasets and outputs packaged as Research Objectives (ROs)
- Ledger log of all ROs & specimen UIDs

D8.3 Demonstrator (UNIMAN)

## Task 8.4

### Data delivery & exploitation

- User acceptance testing (against DiSSCo user stories)
- Training and workshops to promote use
- Delivery of data to collections management systems
- Exploitation of data by third parties

D8.4 Report (Naturalis)



Online open-access  
public reports, code\*  
and documentation



Establishment of  
consortium  
usergroup



Demonstrations and  
examples



Presentations,  
workshops &  
training delivery

- Quarterly meetings of active Task/Subtask Leaders (prior to Exec Board Meeting)
- Regular distributed Task/Subtask team meetings (TBC)
  - Task 8.1: Co-writing/review meetings (Landscape Analysis)
  - Tasks 8.2, 8.3: Short virtual stand-up meetings for service/platform development
  - Task 8.4: Workshops and training (virtual and physical – EU equivalent to DarwinCore Hour?)
- MOBILISE WG1 meeting? (**Q2/Q3 2019** for landscape analysis?)
- Biodiversity Next: Digitisation Next Symposia & others **22-25 Oct 2019**
- CETAF ISTC/DWG (**Jan/Feb 2020 @ NHM?**)

ICEDIG (Active)	MOBILISE (Active)	DiSSCo Prepare (Planned)	Other Initiatives
<ul style="list-style-type: none"> <li>• WP3: Imaging and [data] extraction</li> <li>• WP4: Automated text digitisation</li> <li>• Userstories</li> <li>• Minimum Information about a Digital Specimen (MIDS)</li> <li>• Interoperability with collections management systems</li> <li>• Data infrastructure (WP6)</li> </ul>	<p>COST Action – will have focused working groups with multiple relevant to SDR</p> <p>Also likely to support focused development and testing workshops/meetings</p> <p>Potential STSMs for technical training</p>	<ul style="list-style-type: none"> <li>• WP1: User needs (research &amp; services)</li> <li>• WP3: Capacity enhancement (data readiness)</li> <li>• WP5: Common resources &amp; standards</li> <li>• WP6: Technical architecture</li> </ul> <p>General DiSSCo compliance</p>	<p>ELIXIR (esp. <u>Authentication and Authorisation Infrastructure</u>)</p> <p>iDigBio (related activities)</p> <p>Catalogue of Life Plus (name resolution)</p> <p>...Others!</p>

Also relevant to CETAF Information Science and Technology Committee (ISTC) and Digitisation Working Group (DWG)

- Huge boom in functional applied AI services, consumer apps, open source software and hardware
- Automated *digital* specimen metadata extraction is most effective way of reducing cost – automated physical interactions (robotics, mechanisation) not-cost effective or feasible for majority of legacy collections
- Limited application in natural history collections to date but huge potential
- Existing e-platforms supporting similar workflows in other RIs already exist (see UNIMAN presentation from Goble et al)

Risk/Issue	Mitigation
Untested distributed development with new teams (wider DiSSCo risk)	Regular standup meetings, agreement on style, support via other DiSSCo-related initiatives (inc. DiSSCo PREPARE WP3)
Inaccurate timing of development tasks/deliverables (rough estimates)	Focus on MVPs for Milestones & Deliverables and users (Contractual and Functional)
Inability to hire and retain specialist staff (developers and tech supplier managers)	
Inability to respond to changing user needs and emerging DiSSCo standards (this is a completely new concept and our anticipated user stories may change dramatically)	Microservices and containerisation should allow us to add and modify workflows to changing user needs in project
Constraints with institutional systems / culture prevent integration (collections management systems, media asset management)	Coordinate with related work in ICEDIG, discussions in advance with consortium partners to understand constraints
Long-term support of tools, services and platform	Long-term (in-kind) support from DiSSCo partners
Not being aware of relevant work in other institutes, projects (esp. ICEDIG) and publications	SYNTHESYS+ partners in ICEDIG to monitor and re-use outputs, WP lead to communicate with key researchers / initiatives

## ICEDIG

- [ALICE: Angled Label Image Capture and Extraction for high throughput insect specimen digitisation](#)
- [A Low Cost Approach to Specimen Level Imaging of Natural History Microscope Slides using a DSLR System](#)
- [Interim Report on Quality Control in Imaging DRAFT](#) (Nieva de la Hidalga et al, 2019)
- [Methods for automated text digitisation](#) (Owen et al, 2019)
- [Minimum Information about a Digital Specimen \(MIDS\) DRAFT](#) (Hardisty, 2019)

## SYNTHESYS3

- [Report: Automating data capture from natural history specimens](#)
- [Report: Digitisation on Demand](#)
- [Software: Inselect – desktop application for NH image processing, barcode reading, metadata](#)

## Other

- [BugSnapper: Camera Design for High-throughput Digitisation](#) (Mark Hereld, Argonne National Laboratory)
- [OpenRefine Platform \(StanDAP-Herb\)](#) (Fabian Reimeier, BGBM)

## Teamwork Task List:

<https://dissco.teamwork.com/#tasks/1782802>

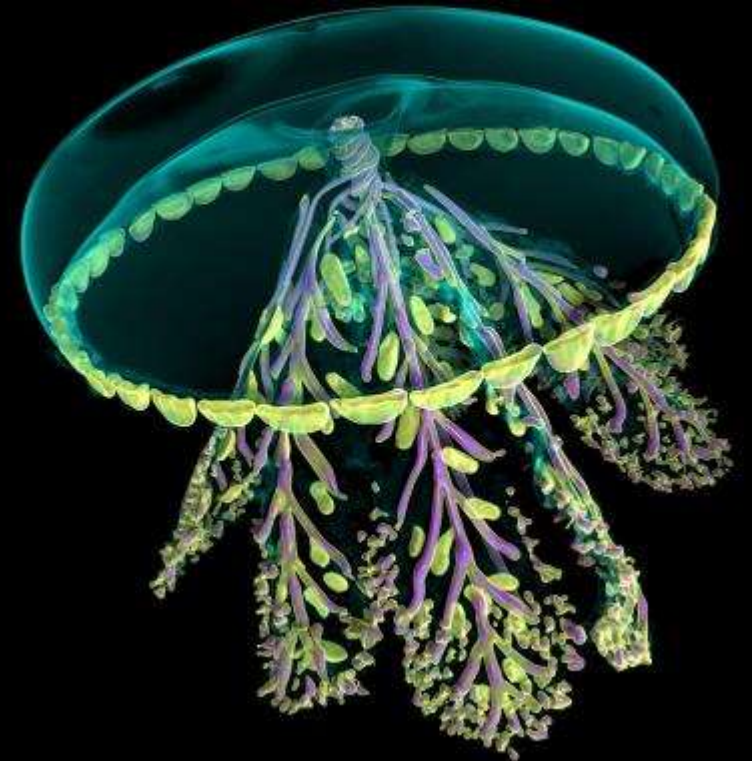
## Grant Agreement

- Objectives: 130
- Description of work: 130-132
- Participation per partner: 133  
Deliverables: 133-134
- Milestones: 134

**Please let me know about any  
potentially interested teams, projects  
or relevant publications!**



**Any questions?**



## **T8.2.1 Optical character recognition, data capture and analysis**

- Recognition of handwritten and printed text

## **T8.2.2 Data mining and linkage**

- Extraction, mapping and linkage of metadata in Subtask 8.2.1 to DwC fields, resolution services (e.g. people, taxa), links to relevant external sources

## **T8.2.3 Image analysis**

- Semantic segmentation
- Colour analysis
- Automated collection assessments
- Automated identification & erroneous determination detection

## **T8.2.4 Georeferencing**

- Strongly linked with "Data mining and linkage" and "Human interaction and crowdsourcing"
- Coordinate transformation and checking services
- Reverse geocoding
- Electronic itineraries (building on gazetteer working in SYNTH3)

## **T8.2.5 Human interaction and crowdsourcing**

- Tools for creating ground truth images and training datasets
- Dictionary/ontology curation
- Integration (inputs/outputs) with crowdsourcing platforms for verification

## **T8.2.6 Workflow and Tools Registry**

- Workflow creation tools
- Registration and curation system for workflows
- FAIR metadata schema compatibility

Platform will:

- Adhere to FAIR principles
- Require an authentication and authorisation system (will include some paid or restricted services)
- Use API-based services (containerised)
- Workflows described in CWL
- KPIs and metrics of usage – EC requirement and give quantitative user information