Research Ideas and Outcomes 9: e106883
doi: 10.3897/rio.9.e106883

Research Article

Mapping the Digitisation Workflow in a University Herbarium

Karen M Thompson‡, Joanne L Birch‡

‡ University of Melbourne, Melbourne, Australia

Corresponding author: Karen M Thompson (karen.thompson@unimelb.edu.au)

Reviewed v 1

Academic editor: Editorial Secretary

Received: 24 May 2023 | Accepted: 18 Jul 2023 | Published: 18 Aug 2023

Citation: Thompson KM, Birch JL (2023) Mapping the Digitisation Workflow in a University Herbarium. Research Ideas and Outcomes 9: e106883. https://doi.org/10.3897/rio.9.e106883

Abstract

Specimens or objects in natural history collections hold substantial research and cultural value that is enhanced where these items are made digitally available. Benefits of digitisation include increasing open access to collection-based biodiversity data, increasing productivity of scientific research, enabling novel research applications of digitally accessible data, reducing preservation requirements through reduced object handling, and expanding potential for “remote curation” in collections. However, the time available for object and data digitisation is limited for most collections. Well documented digitisation workflows can ensure that curation time is efficiently applied to achieve digitisation outputs, and that digitisation standards are consistently applied within and among projects.

While this case study focused on the generation of digitisation workflows in a medium-sized Australian university-based herbarium, the findings of this study are relevant to collections globally. The curation workflows comprise a set of modular steps required for the digitisation of herbarium specimen data and images. Steps are clearly identified as requiring human-mediation versus those that can be automated, those that require on-site versus remote-access, and those that require transfer or transformation of data or files. This clarity enables consideration of the opportunities and challenges for increasing efficiencies for collection-based digitisation, data and file management. The maps provide a contextual framework for herbarium-based digitisation pathways for those who work with

specimen-derived biodiversity data, and an insight into these tools for those who are not familiar with herbarium protocols.

Keywords

collection management, curation, digital extended specimen, digital imaging, digitisation, herbarium, workflow

Introduction

The key arguments for effective digitisation of herbarium specimen sheets are the same as those for all natural history and cultural material collections – specimens or objects can provide greater research and social value, while their physical integrity is better protected for future applications, if they are readily available in digital formats (e.g., Baird (2010), Kalms (2012) p. 11). Digitisation of collection objects creates significant economic benefits including efficiency in curation time and return on the research-associated benefits; for example, economic modelling indicates that digitisation of the 80 million collection objects in the British Museum would generate a seven to ten times return on investment (Popov et al. 2021). Access to physical specimens remains essential for many types of collection-based research, including for the study or quantification of those micromorphological characters that are not visible even in high-resolution specimen images (Phang et al. 2022). However, access to digital data accompanied by specimen images can enable vast specimen-associated resource use: potential for taxonomic identification of specimens, nomenclatural clarifications, assessment of type designations, and generation of phenology or trait-based data, to name only a few applications of these data.

Digital resources increase and enable open access to biodiversity data as per the FAIR principles (ensuring data are Findable, Accessible, Interoperable, and Reusable; Wilkinson et al. 2016). In doing so, they lower the barriers – financial, institutional, academic, sociological (Hedrick et al. 2020, p. 243) – to participation in research, and thereby expand the field to include interdisciplinary involvement and more diverse voices. Remote discovery and assessment of digitised collection objects may also increase efficiency of in-person research visits to collections, potentially providing a cost saving for research funding (Popov et al. 2021). The increased potential for discovery of collection resources is particularly valuable for small collections, which may receive fewer in-person research visits than larger herbaria due, in part, to the time and cost savings associated with traveling to a small number of larger herbaria over a larger number of smaller herbaria (Marsico et al. 2020). The recent years of pandemic-induced constraints on travel and personal movement have further highlighted the importance of online assets in enabling remote access to objects for research. Interaction with digital objects reduces post-digitisation object handling, and eliminates unnecessary handling, such as retrieving a specimen only to check data. There are also potential curation benefits in terms of reducing ongoing object conservation costs, and of costs and risks associated with the transportation of objects for loans. The digital images also provide a valuable specimen

record, which are rendered the only remaining records if the original items are lost or destroyed (e.g., 'Australian customs destruction of herbarium specimens' (Staight 2017, Stokstad 2017); 'Loss of museum specimens in fire' (Escobar 2018)).

While the benefits of digitisation are widely recognised, the costs of digitisation in terms of labour and resources are considerable. In almost all collections resource availability for digitisation, and therefore digitisation effort, can ebb and flow; priorities follow funds, staffing levels can be variable, and momentum for digitisation projects may be intermittent. Digitisation standards must remain high and be consistently applied within and among projects. This requires that protocols are well documented, and that staff, despite turnover, are well-trained and consistently apply established curation and digitisation protocols. Small or medium size collections are often heavily reliant on a volunteer workforce and may integrate both in-house and outsourced digitisation initiatives, necessitating data and imaging transfer and file format compatibility across software. Digitisation workflows must be flexible and adaptable (without compromising quality), for those workflows are regularly revised and further optimised as obstacles arise and are mitigated, and as best practices evolve. These apparently conflicting requirements are more effectively achieved when digitisation workflows are well documented, contextualised, and understood.

In this paper we share the output of mapping the digitisation workflows efforts at the University of Melbourne Herbarium (IH herbarium code: MELU). Of particular interest to us was the identification of impediments to workflow efficiencies, where this workflow was situated in relation to other workflow descriptions in literature, and developing an understanding of the extent to which image and data collection relies on the physical involvement of a human. For MELU, mapping the curation workflow for digitisation was done in part to streamline the digitisation workflow, identify bottlenecks, and to identify risk points in the data management pipeline for future attention and mitigation. Our intent in sharing this workflow is to contribute the real-life experience of a medium-sized collection to the literature, so other small and medium-sized herbaria may use this as a reference for reviewing or designing their own workflows. Such maps also act as a communication tool for securing resources to enable digitisation work. As Nelson et al. (2015) point out: “The data resources housed in those small or otherwise resource-challenged collections are particularly valuable because they often contain records from areas or taxa that are underrepresented in larger collections” (p. 2).

In the Background section we explore the last decade of workflows in the literature. We then introduce the MELU case study, describe the methodology used to build the workflow maps, and present and discuss the workflow diagrams. In the Discussion section we identify and discuss the similarities of the MELU workflow to others in the literature and the contributions of these streams to accurately representing the complexity of specimen digitisation. Finally, we consider the resources and technologies that are required to meet the increasing bioinformatic challenges associated with curation of specimen-associated digital objects and data.

Background and Context

A large proportion of specimens held in herbaria are dried and pressed plant samples, secured to archival card with labels attached, and stamps or handwriting present – the whole object will be referred to here as a ‘specimen sheet’. The majority of specimens are sufficiently two-dimensional that they can be photographed at a single focal depth. A smaller number of specimens include large, three-dimensional structures, e.g., storage roots, succulent stems or leaves, infructescences, or fruit that are not rendered two-dimensional during pressing. Digitisation of these three-dimensional structures, either attached to specimen sheets or held in separate carpological collections, requires the production of multiple images across a range of focal depths that are then combined to generate a single digital image (examples from MELU in Fig. 1).

Figure 1. doi

High-resolution images of a *Banksia canei* specimen (MELUD121102a) from the University of Melbourne Herbarium (MELU), (online.herbarium.unimelb.edu.au/collectionobject/MELUD121102a). © University of Melbourne, 2023.

The definition of ‘digitisation’ has shifted slightly over time. For clarity, here it is used to refer to:

1. the capturing of a digital image of a specimen sheet;
2. the input of all data present on the specimen sheet into a searchable database;
3. the addition of relevant meta-data for the digital image to the same database.

The collection of digital representations of the physical specimen may be referred to as a "digital specimen" (Nieva de la Hidalga et al. 2020, p. 11). As Haston et al. (2015) (p. 9) highlight, a chief aim of any digitisation process is to enable discovery and use of digital objects and their associated data. In the first instance, discoverability facilitates collection

management and therefore provides significant benefits in the form of curation efficiency and effectiveness. Subsequently, discoverability benefits researchers associated with the herbarium, and extends to resource provision for other researchers, cross- and supra-institutional repositories, and for interested members of the public. The subsequent sharing of these digital assets via the internet (including institutional portals such as The University of Melbourne Collection Online (online.herbarium.unimelb.edu.au), biodiversity repositories such as the Australasian Vitrual Herbarium (AVH; avh.chah.org.au) and Atlas of Living Australia (ALA; ala.org.au), or in community contributed databases such as JSTOR Global Plants database (JSTOR; plants.jstor.org)) is a key element of the overall process, and in many cases is its driver. An efficient, scalable, adaptable, and cost-effective workflow (i.e., series of steps enacted either by persons or machine) for the digitisation of the collection objects in an herbarium is critical to the success of the endeavour. The next section briefly surveys literature, in chronological order from the last decade, discussing such digitisation workflows.

Workflow definition and examples

A workflow can be thought of as chain of "atomised and executable components with the relationships between them to clearly define a control flow and a data flow" (Hardisty et al. 2022, p. 324). Further:

> Digitization workflows span across human mediated processes through data and computationally intensive automation where software tools and services are the actors and intersect field collection techniques, institutional accession policy, differences in curatorial practice among domains, and involvement of the general public in crowd-sourced methods. (Beaman and Cellinese 2012, p. 12)

In their report for the Australian Museum, Tann and Flemons (2008) investigate various means of data capture from images of insect specimen labels, while studying the viability of allocating some tasks to volunteers. Inserting data into an institutional collection management system (in this case, Axiell EMu for internal access only), the workflow options parallel the decisions being made in other collections (Fig. 2). They conclude that there "are some advantages to taking a photograph of a specimen label before capturing the data from that label … the following transcription process is quicker by about 20%, easier and more convenient" (p. 16).

Two years later, in their conference paper Moen et al. (2010) propose a high-throughput workflow for extracting data from specimen labels at the Botanical Research Institute of Texas (IH herbarium code: BRIT; over 1 million specimens at the time). While the paper focuses on meta-data standards, the workflow is notable for its human-in-the-loop aspects (Fig. 3).

In the same year, Granzow-de la Cerda and Beach (2010) analyse digitisation approaches at the University of Michigan Herbarium (MICH) during a 2002-2008 programme focussed on Mexican and Mesoamerican holdings. Nine key tasks were organised in three workflows (Fig. 4), with timing and costing assessed. The authors advocate for the "clear

efficiency benefit of articulating a biological specimen data acquisition workflow into discrete steps, which in turn could be individually optimized" (p. 1830) and reiterate that "the limiting cost for collections computerization is the human labour needed to capture data from specimen labels into a structured database" (p. 1831). This work also demonstrates the benefit of language and geographic familiarity for data transcription efficacy.

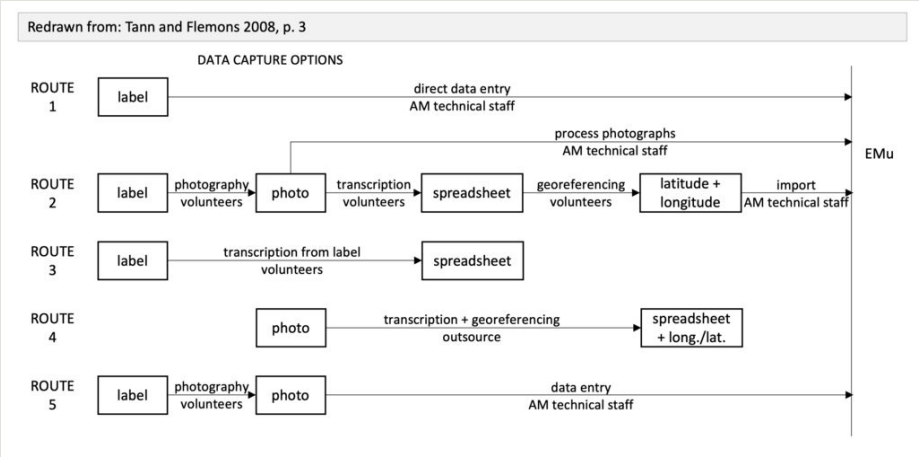

Figure 2. doi

Workflow options for data capture from specimen labels, redrawn from Tann and Flemons (2008), Figure 1 on p. 3.

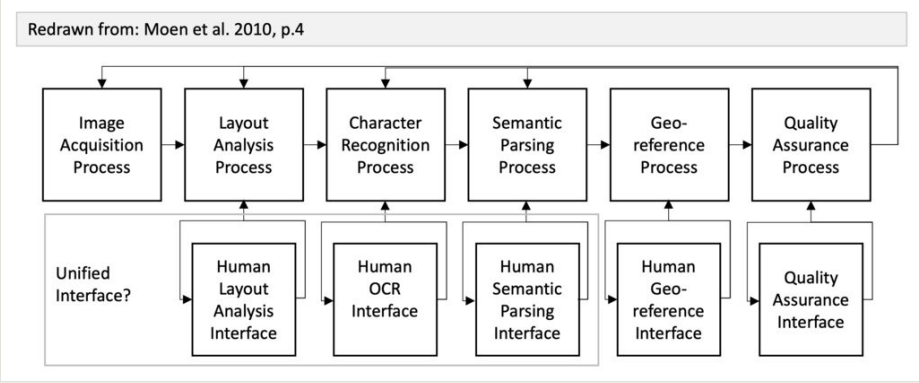

Figure 3. doi

Workflow for data capture from labels, redrawn from Moen et al. (2010), Figure 4 on p.4.

By 2012, digitisation within collections was sufficiently established that the ALA published the *Digitisation of Heritage Materials* guidance (Kalms 2012). The document contains many diagrams, with the workflow for the creation of a data asset being of most interest here (Fig. 5).

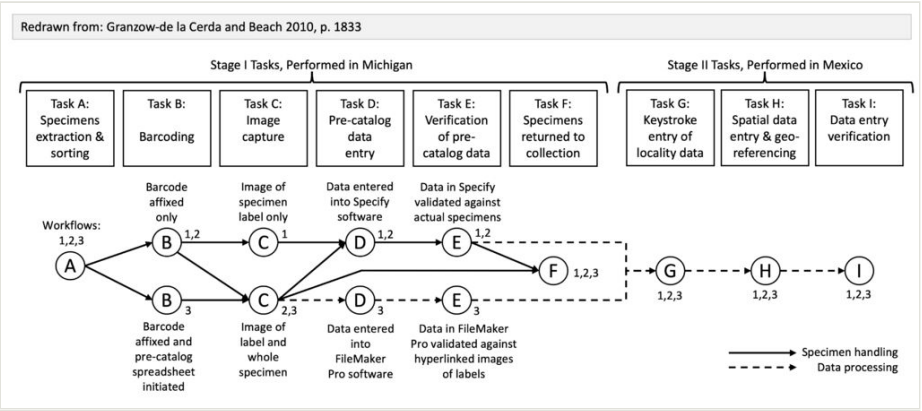

Figure 4. doi

Workflow, redrawn from Granzow-de la Cerda and Beach (2010), Figure 1 on p. 1833.

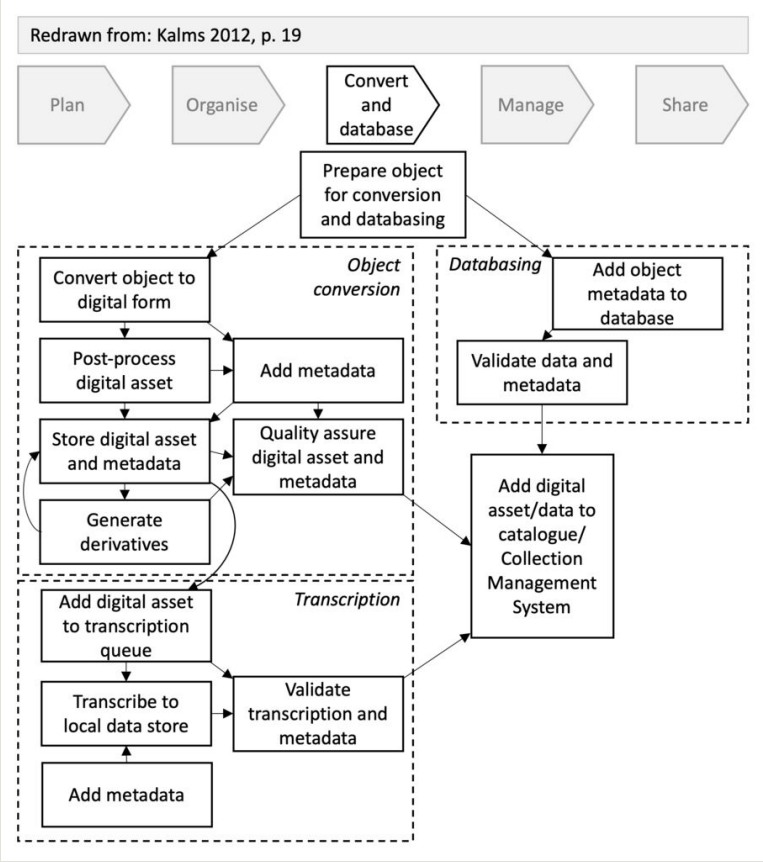

Figure 5. doi

Workflow for creating the digital asset, redrawn from Kalms (2012), Figure 8 on p. 19.

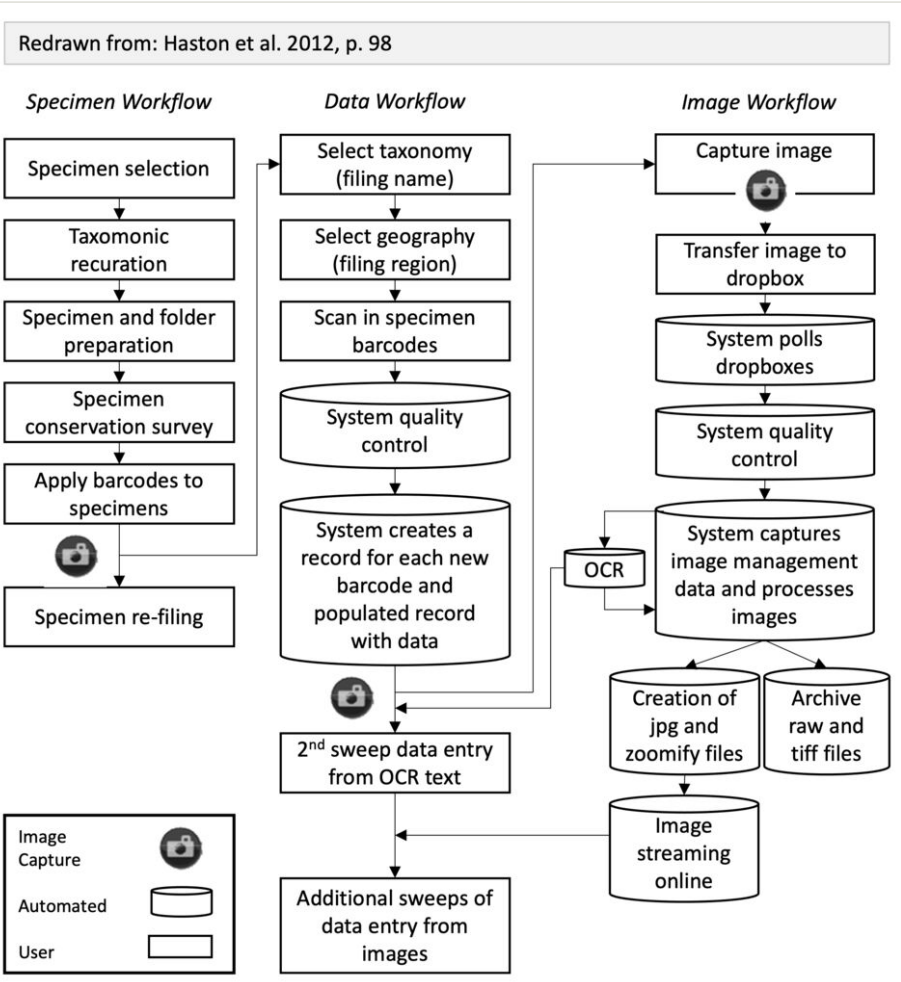

Figure 6. doi

Workflow RBGE (OCR = optical character reading), redrawn from Haston et al. (2012), Figure 1 on p. 98.

A special edition of ZooKeys, with twelve papers, was published in July 2012 (eds. Vladimir Blagoderov and Vincent Smith): 'No specimen left behind: mass digitization of natural history collections.' Haston et al. (2012) detail work done at the Royal Botanic Garden Edinburgh (RBGE), which "given the irregular nature of much of the funding available for digitisation" favoured "digitisation workflows [based] on a modular system which has the potential to be scaled up as funding becomes available" (p. 94). The workflow (Fig. 6) is of particular interest for its splitting of the specimen and data workflows, with the intent of separating the imaging from the data capture. Their text is almost unique in its acknowledgement of the resourcing impost of subsequent data management. Tegelberg et al. (2012) describe a commercial digitisation service (previously named Digitarium, now Bioshare Digitization (bioshare.com)), including a conveyor belt system for faster handling.

Their workflow (Fig. 7) differs from others in its relationship to a ‘customer’, which is understood to refer to herbarium collections, it is included here as the workflow does cover elements of interest. Nelson et al. (2012) describe five task clusters observed as common to the digitising workflows of the varying biological collections (within USA) surveyed:

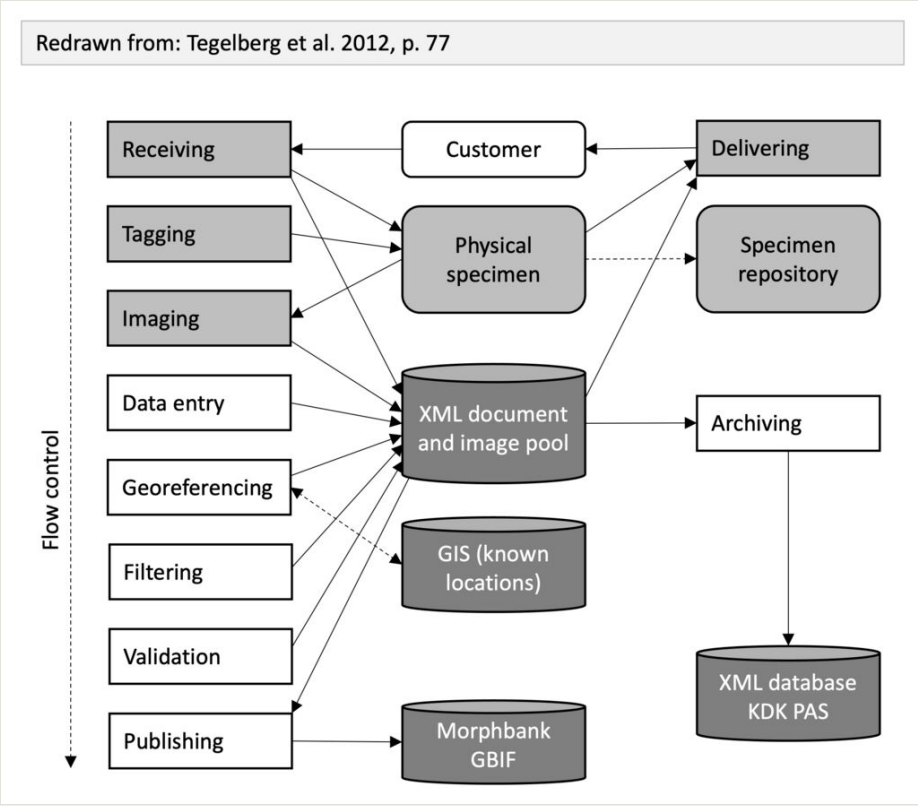

Figure 7. doi

Functional model of digitisation process, redrawn from Tegelberg et al. (2012), Figure 1 on p. 77.

1. pre-digitization curation and staging,
2. specimen image capture,
3. specimen image processing,
4. electronic data capture, and
5. georeferencing locality descriptions;

ordered into three workflows (Fig. 8). Tulig et al. (2012) describe digitisation efforts at the New York Botanical Garden (NYBG) and demonstrates how they have changed over time. As expected of a maturing process, the complexity of the three workflows increases as the process evolves (Fig. 9). Four years later, Thiers et al. (2016) provide an update on this work at NYBG in which workflow is textually described without workflow diagrams.

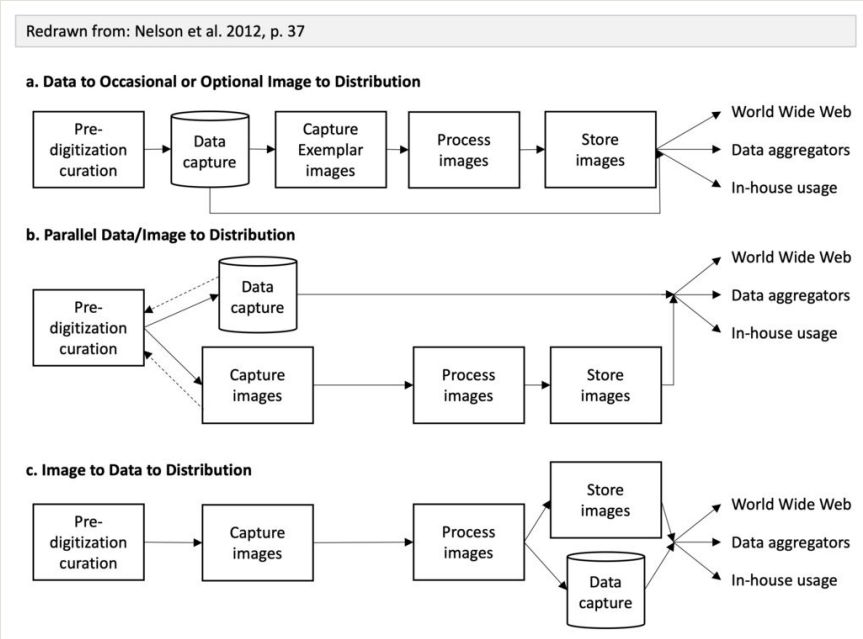

Figure 8. doi

Dominant digital workflows observed, redrawn from Nelson et al. (2012), Figure 6 on p. 37.

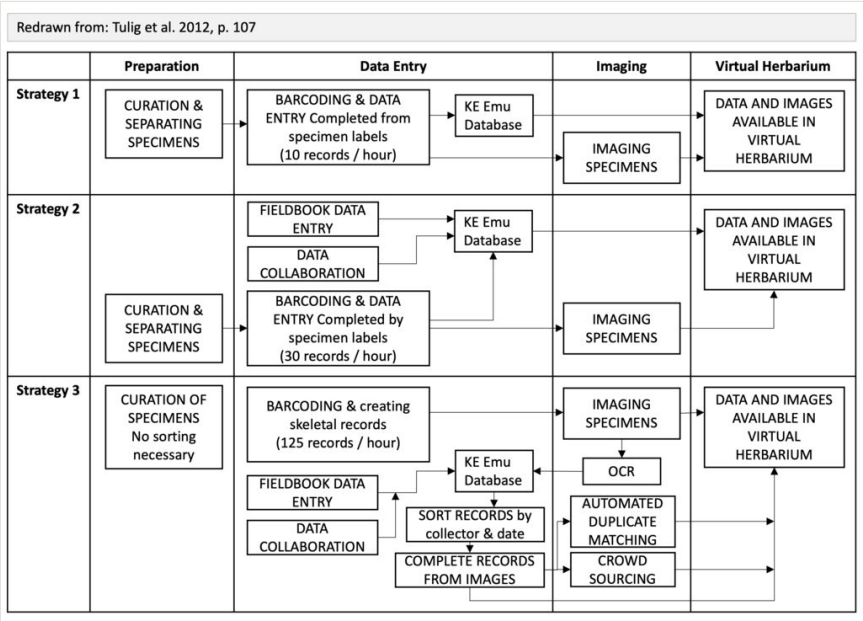

Figure 9. doi

NYBG approaches, redrawn from Tulig et al. (2012), Figure 1 on p. 107.

In the 2013 paper by Barber et al. (2013) the focus is on the development of technology (SALIX: Semi-Automatic Label Information eXtraction) to extract data from a specimen label at Arizona State University (ASU) Herbarium. The workflow within which this specific work fits is also shared (Fig. 10).

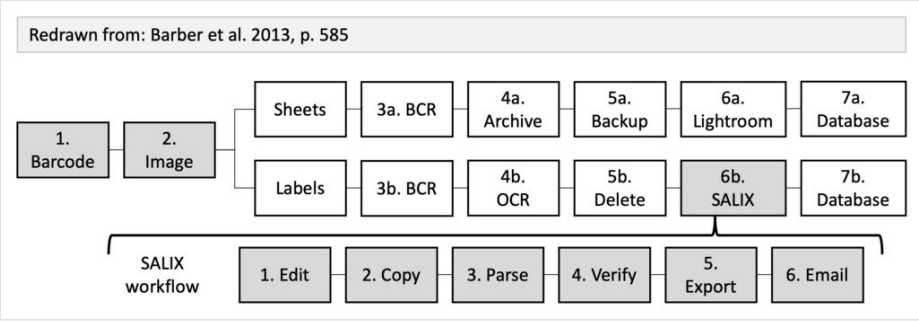

Figure 10. doi

The SALIX method workflow (BCR = barcode reading; OCR = optical character reading), redrawn from Barber et al. (2013), Figure 6 on p. 585.

The Phillips et al. (2014) review of the digitisation practices of eighteen partners of the SYNTHESYS3 initiative (synthesys.info) is an excellent survey of the wider natural history collection landscape. While no diagrams were in the report (likely due to the diverse collection types covered), the authors highlight that "the most common order in which tasks are performed (were all the tasks to be performed)" were those listed below (as per the original text, p. 13):

1. Selection
2. Transfer of material from one area to another
3. Application of barcodes and "other" tasks
4. Full (or partial) data capture
5. Imaging
6. Records management
7. Returning material
8. Quality assurance

The authors also note that "the vast majority of institutions included a full data capture step within their digitisation workflows" (p. 13), and that "the majority of institutions are still capturing full specimen metadata prior to the imaging step in their main digitisation workflows" (p. 18).

Nelson et al. (2015) then refine the 2012 iDigBio information (Nelson et al. 2012) to focus on flat sheets and draw two new visions of possible digitisation workflows: (a) object to data to image, Fig. 11; and (b) object to image to data, Fig. 12. Tasks are separated into fourteen modules (listed below, as per text in Box 1, p. 3) with very detailed descriptions of the considerations involved with each task; available via GitHub (github.com/iDigBioWorkflows/FlatSheetsDigitizationWorkflows).

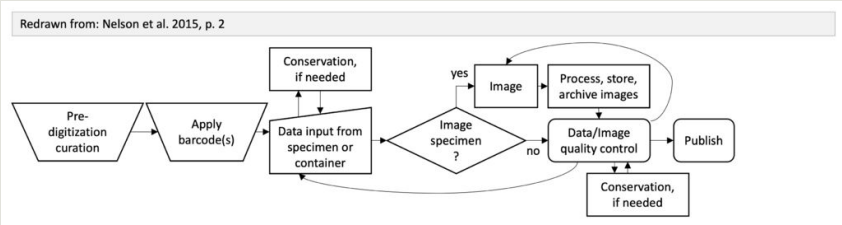

Figure 11. doi

Object-to-data-to-image workflow, redrawn from Nelson et al. (2015), Figure 1 on p. 2.

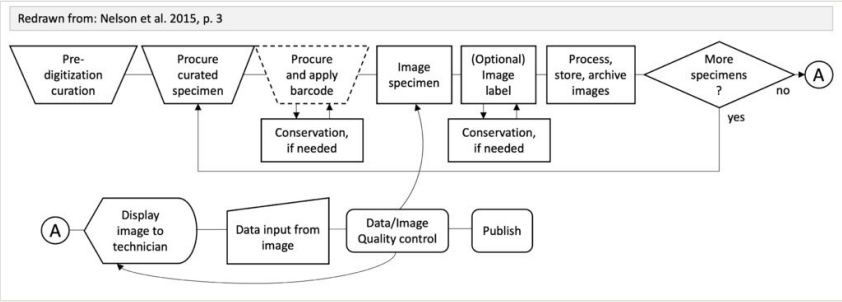

Figure 12. doi

Object-to-image-to-data workflow, redrawn from Nelson et al. (2015), Figure 2 on p. 3.

1. Pre-digitisation curation
2. Selecting components for an imaging station
3. Imaging station setup, camera / copy stand
4. Imaging station setup, light box
5. Imaging station setup, scanner
6. Imaging
7. Image processing
8. Organising and implementing a public participation imaging blitz
9. Image archiving
10. Selecting a database
11. Data capture
12. Organising and implementing a public participation transcription blitz
13. Georeferencing
14. Proactive digitisation

The authors caution that “broad disparities in digitization starting points, institutional infrastructure, curatorial practices, and precise digitization tasks among and within these groups focused on different taxa make the development of a single, consensus object-to-digitized-content workflow impractical” (p. 2).

Around this time the literature appears to shift from the workflows – settling on the *object-data-image* (Fig. 11) and *object-image-data* (Fig. 12) options (per Nelson et al. 2015) – to

focus on the optimisation of specific elements. Many take the image as a starting point and then focus on harvesting data from the specimens or labels (e.g., Triki et al. (2020), White et al. (2020), Owen et al. (2019), Kirchhoff et al. (2018), Haston et al. (2015), Drinkwater et al. (2014), Mononen et al. (2014)); some include streamlining imaging process (e.g., Sweeney et al. 2018, Tegelberg et al. 2014); and others extend the use of the digital images (e.g., Corney et al. (2018), Carranza-Rojas et al. (2017), Unger et al. (2016))

While the excellent paper of Nieva de la Hidalga et al. (2020) focusses on automating image manipulation at the Meise Botanic Garden (MGB), it includes a digitisation workflow (Fig. 13) and detailed task listing. The modular design of that workflow allows for the integration of in-house and outsourced digitisation efforts and focuses on achieving consistent quality standards and recording an audit trail to ensure scalable image production. While a digital image is again the starting point in Walton et al. (2020), the increased modularity of the digitisation process is evident (Fig. 14).

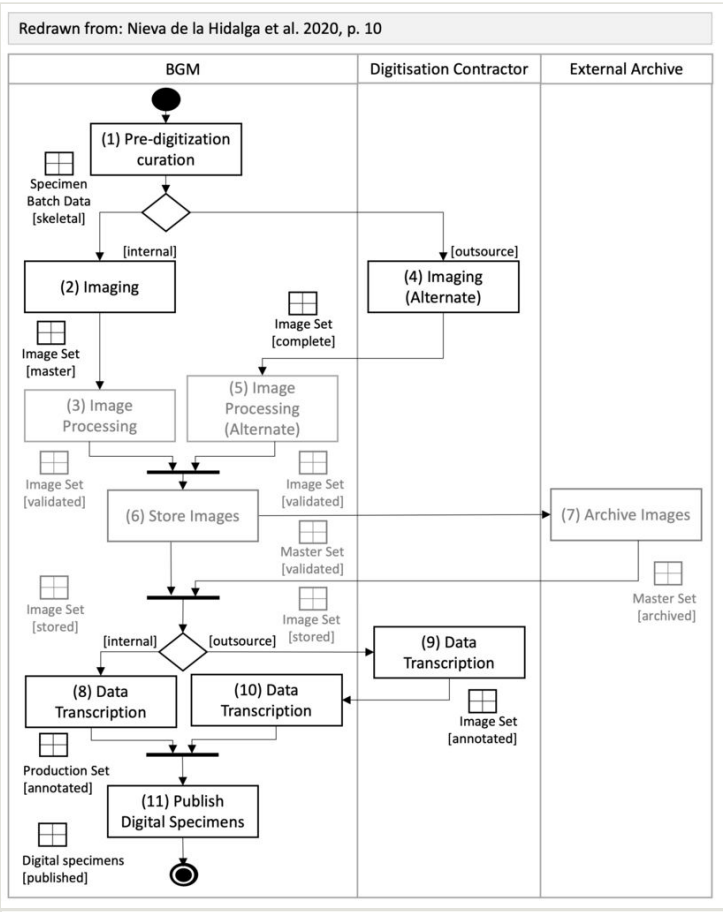

Figure 13. doi

MBG digitisation workflow, redrawn from Nieva de la Hidalga et al. 2020, Figure 2 on p. 10

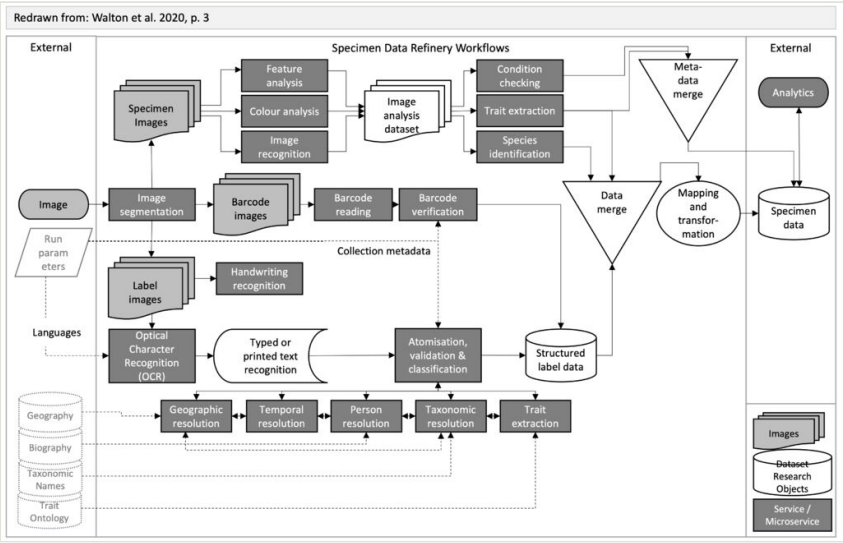

Figure 14. doi

Potential workflow, redrawn from Walton et al. (2020), Figure 1 on p. 3

With a few notable exceptions, most digitisation workflows available in the literature are generalised, and understandably so, for this facilitates their uptake and adaptation. Surveys or applications of workflows tend to focus on large institutions or conglomerates; and large-scale processes for flat sheet herbarium specimens appear to have converged on conveyor belt systems with manual transcription, such as that used by the National Herbarium of New South Wales in Australia which incorporates Picturae (digitisation) and Alembo (transcription) (Cox 2022).

In this paper we present the detailed workflow paths for digitisation of MELU collection, as a real-life case study and contribution to the literature for medium-sized institutions. Though, echoing Nelson et al. (2015) (p.2), it is worth noting advice from Barkworth and Murrell (2012) (p. 60):

> There is no best approach for digitizing herbaria; there are multiple effective approaches. The needs and resources of large research herbaria with multiple type specimens and collections from many countries and multiple centuries differ from those of small herbaria serving a forest district or a teaching institution. … Adopting theoretically suboptimal procedures for digitization may be the best procedure if the resources needed for adopting a better procedure are not available.

Case Study: the University of Melbourne Herbarium

Established in 1926, the University of Melbourne Herbarium (MELU) is the largest university herbarium in Australia, with an estimated 150,000 specimens. Taxonomic diversity spans plants and fungi, as well as historic botanical objects and artwork. MELU is

a research and teaching collection, and the collection's strengths reflect University of Melbourne academic expertise and teaching activities. Digitisation efforts at MELU commenced in 2003 with the establishment of a FileMakerPro (claris.com/filemaker) database that was accessible online (N. Middleton, pers. comm.) and ramped up significantly from 2014 with the transfer of these data into the Specify collection management system (Specify Collections Consortium, Lawrence, KS; specifysoftware.org) and subsequent digitisation efforts (G. Brown, pers. comm.). In 2012, the equipment and software for the generation of high-resolution specimen images and standard protocols for image production were provided to MELU through the JSTOR Global Plants Initiative, which enabled the generation of high-resolution digital images. In 2020, MELU transitioned from a local networked collection management system (CMS) accessible on-site in the Herbarium on the Parkville campus to a CMS hosted on a virtual machine which enabled access on-site or remotely.

Digitisation rates at Australian herbaria are high, partially as a result of digitisation efforts concentrating on Australian specimens during the 2000s to support the development of what is now the AVH. AVH was created in 2001 (Nelson and Ellis 2018, p. 2) under the auspices of the Council of Heads of Australasian Herbaria (CHAH; chah.gov.au), to deliver knowledge and information of plant, algal, and fungal biodiversity. Digitisation rates at University herbaria are more variable, reflecting variation in the size of collection and curation priorities, and access to digitisation resources, including staff or volunteers and camera equipment. Currently around 21% of the holdings at MELU are digitised, with ongoing digitisation efforts focusing on taxonomic strengths of the collection (e.g., Myrtaceae, Fabaceae, the algal collection). The NCW Beadle Herbarium (NE, ca. 113 K specimens) at the University of New England has all specimen data fully digitised. Substantive data digitisation has also occurred at smaller Australian University herbaria (e.g., The University of Newcastle (DMHN), James Cook University (JCT), La Trobe University (LTB), Macquarie University (MQU), University of Wollongong (WOLL)).

The University of Melbourne Herbarium Collection Online (online.herbarium.unimelb.edu.au) was created in 2018, recognising the previously untapped potential for increased access to and engagement with high-resolution specimen images, including to enable data reuse. Specimen data can be searched or browsed, georeferenced specimens are mapped, and plant features or the collector's handwriting are visible in the high-resolution images. The Collection Online links directly to the Specify CMS to provide access to MELU data in real-time, to facilitate viewing, and enabling the downloading of the full-size high-resolution images (ca. 250 MB per image). The Collection Online has been pivotal for expanding access to the collection, with user statistics documenting consistent national and global use of this resource. MELU also provides all digitised material (data and specimens images) to the ALA – "a collaborative, digital, open infrastructure that pulls together Australian biodiversity data from multiple sources, making it accessible and reusable" (Atlas of Living Australia (ALA) 2022). ALA developed from the AVH (Nelson and Ellis 2018, p. 2) and is the Australian node and a full voting member of Global Biodiversity Information Facility (GBIF; gbif.org) (Atlas of Living Australia (ALA) 2022).

The digitisation protocols employed at MELU have evolved over the 20+ year history of the endeavour. For data transcription, protocols follow the standards developed by Biodiversity Information Standards (TDWG; tdwg.org) and Darwin Core (DwC; dwc.tdwg.org). For production of high-resolution images, MELU images (refer to Fig. 1 for a representative image) follow Global Plant Initiative (GPI) protocols (JSTOR 2018), which require each herbarium sheet to include:

1. biological specimen
2. colour chart
3. scale bar
4. labels
5. barcode (where applicable)
6. institution name

MELU specimen sheets include the unique catalogue number of the format "MELU" followed by a letter, seven digits, and single letter, e.g., MELUD121102a. In line with the teaching remit of the University, MELU has a volunteer program that provides training in curation protocols and management of research associated with biodiversity specimens to approximately 25 volunteers annually. Student volunteers are significant contributors to MELU digitation efforts, which means that delegated processes must be carefully documented and detailed to ensure consistency in execution.

Methodology: building the workflow maps

The workflow maps described in this paper were developed as an element of a collaborative project between MELU and research data specialists from the Melbourne Data Analytics Platform (MDAP) at the University of Melbourne. The initial intent for the mapping was to enable the MDAP team to understand the ecosystem within which a specific investigation (into possible methods for machine-reading specimen sheet label data) was situated. Understanding the connections to other elements is critical, especially when focussing on a singular ‘module’ of a digitisation workflow. Taking time to consider the broader context early in the process encourages forward-thinking, avoids developing the work in a direction that may limit future usefulness, and facilitates identification of potential extensions or reuses of components.

The suite of workflow maps detailed in later sections was the outcome of many conversations, over some months, between the MELU curator and a data specialist with limited herbarium domain knowledge. This was an exercise in trans-disciplinary collaboration, and the utility of the workflow depended on allowing time to develop a shared vocabulary. A key value of a non-botanist taking responsibility for drawing the workflow was that they asked questions to elucidate knowledge that could easily be presumed or remain within the mind of the expert.

The workflows were constructed initially as one large comprehensive map of the multiple curation pathways. It was built ‘naively’ – that is, no predetermined workflow was used as

scaffolding or framing, but instead, the tasks undertaken within the digitisation process at MELU were discussed one-by-one and connections made between them. These tasks were then bundled into 'modules', based on natural break points, when the process could be paused without detriment. In this way, the 'outline' map was created. This mirrors the 'grounded theory research methodology' employed by Nelson et al. (2012), which the authors describe as an "inductive social science research method that begins with data collection and leads to qualified conclusions (theories) about those data," particularly with respect to "constructing categories from the data rather than from hypotheses" (p. 21).

Verbal information about herbarium processes were translated into a diagram by the data specialist, and as that diagram iterated over many conversations it became a tool of discovery and mutual communication. As the team were working remotely, the communication tools and diagrams necessarily took a digital format. In retrospect, this work also incorporated 'visual thinking' methodology, which "rests on the intertwined relation between visual perception and cognition" (Fernández-Fontecha et al. 2018, p. 6). The workflow combines written language, basic visual shapes, and the form of their arrangement into a diagram of relationships to each other, to facilitate communication of a complex set of steps. The form of the elements in the MELU workflow diagrams, while based on standardised workflow shapes, are bespoke to this exercise and a legend is provided to ensure they can be universally understood without context.

Creating the detailed workflow maps for MELU met the original intent of situating a specific task into the broader herbarium landscape. It also led to other positive outcomes, including:

- reduced key-people risk – this information is now available for project development and is accessible for the entire curation team;
- manual tasks, file transformations, and other bottlenecks and points of potential risk, have been identified and prioritised for mitigation; and
- the map is an effective communication tool for management, stakeholders, and to support funding or grant applications.

MELU workflow maps

MELU digitisation practices currently follow three streams:

1. *complete cataloguing*: full data capture via transcription from specimen, with subsequent high-resolution image/s;
2. *data only*: lower-resolution image/s, followed by full data capture via manual transcription, prioritised for later high-resolution imaging;
3. *digitally native data*: specimens with data directly loaded into the CMS via a workbench, then prioritised for later high-resolution imaging; this stream is not mapped in the outline, as it is essentially a subset of Stream 1 that does not require data transcription.

In drawing out the maps of each of the above pathways, MELU Digitisation Workflows are represented in several ways:

- *Outline workflow map*, in which tasks are grouped into modules (Fig. 15).
- *Detailed workflow map*, showing all task details and their connections (here presented in parts for readability; Figs. 17, 18, 19, 20, 21) – this is the core workflow map from which the others are derived. (Suppl. material 1)
- Detailed workflow map overlayed *with tools and technology* (Fig. 23).

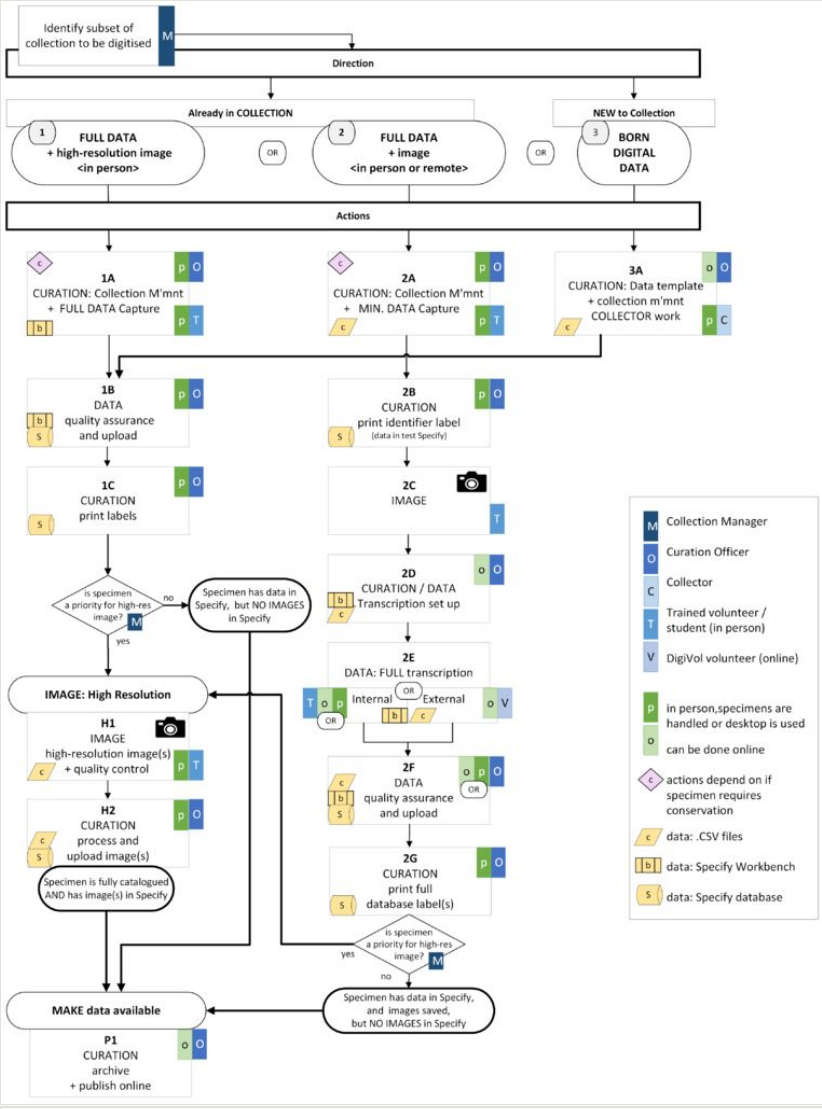

Figure 15. doi

MELU Digitisation Workflow: Outline map.

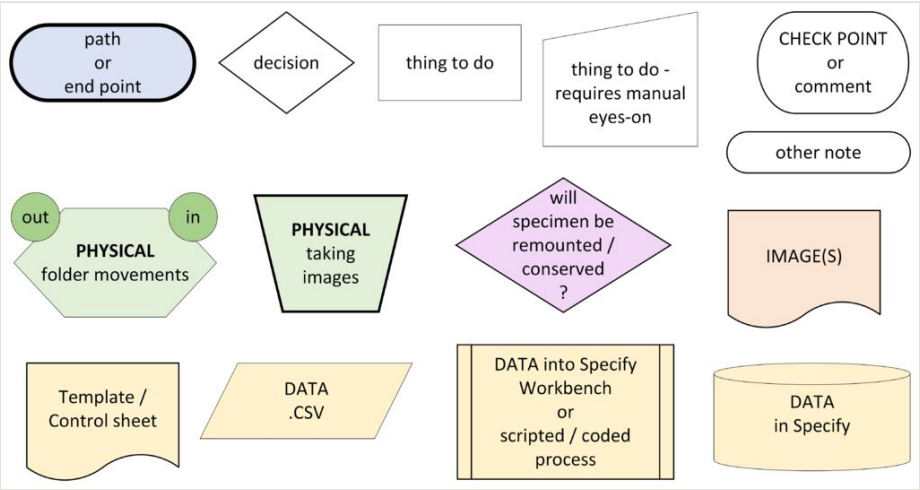

Figure 16. doi

Legend of shapes used in MELU workflow detailed maps. Green elements (hexagons and upside-down wedge shapes) require human physical actions; yellow (the four shapes in the bottom line of the figure) are technology-driven elements; other colours and shapes allow for easy identification in the workflows.

These workflow maps may give the impression they are set in stone, but of course they are representations of evolving processes. Nor should they give the sense that digitisation is a one-off task – e.g., any time the nomenclature of a taxon is updated, if the identity of the specimen changes, or if the specimen requires conservation after damage, then the digital data record needs to be updated (with QA) and/or new images taken (processed, uploaded with QA, and archived). In maintaining accurate collection data, it is essential to maintain version control records to prevent divergence of data on the physical specimens and in the CMS.

Outline

The MELU digitisation workflow outline (Fig. 15) summarises significant detail for the digitisation streams, and groups steps into modules. Using the descriptions from Nelson et al. (2015) (Fig. 11 and Fig. 12 respectively), Stream 1 (*left*) is effectively an object→data→high-res-image pathway, and Stream 2 (*middle*) an object→image→data pathway. At MELU, Stream 1 has historically been the established pathway for digitising specimens that were accessioned in the collection but had not yet been digitally catalogued. This stream facilitated digitisation of specimens that were previously recorded in the print catalogue only. Stream 2 was established to increase efficiency of data transcription within the collection. This stream was developed during 2020-2021, within the 12-month period immediately prior to the onset of the Covid-19 pandemic and became the primary pathway for specimen digitisation during the lockdowns enforced as part of the COVID-19 pandemic response in Victoria, Australia. Each stream is further discussed below.

Stream 1 - data transcription from the specimen

The details for Stream 1 digitisation workflow are included below (Fig. 17; using shapes defined in Fig. 16). The first task (1A) involves: collection management and curation tasks of assigning an accession number to each specimen; setting up a data collection template (using Specify Workbench, which is an environment in Specify allowing for temporary holding of data prior to formal uploading into the CMS); the most recent application of a taxon name on the specimen is checked to ensure that the taxon name is current (as recognised in the Australian Plant Census; anbg.gov.au/cpbr/program/hc/hc-APC.html); all data is input into the collection template. Data capture into the Specify Workbench enables the data entry person to receive feedback about data quality from the functionality built into Specify (including indication of taxa, agents, and locations that are not currently recorded). As all data are captured in Specify, Stream 1 has the benefit of not requiring the transfer of files among software, therefore limiting version control requirements and the risks inherent with these file transfers.

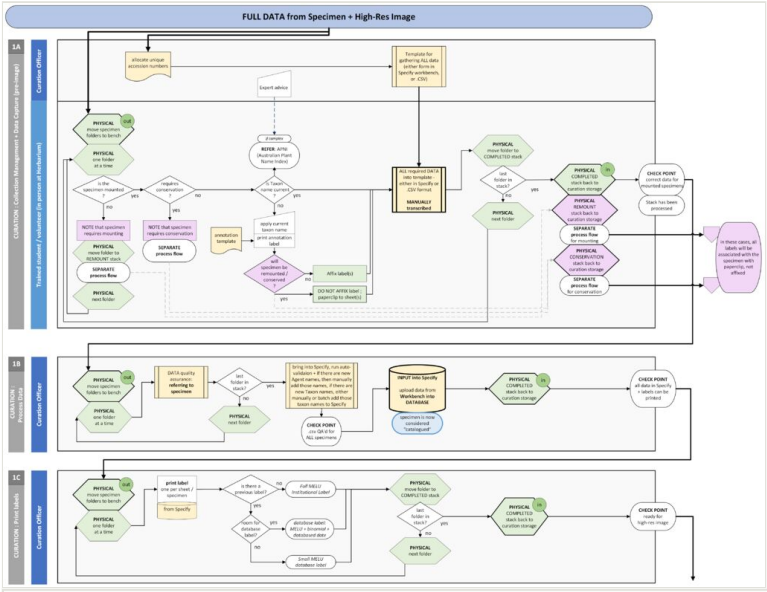

Figure 17. doi

MELU Digitisation Workflow: Detailed, parts 1A, 1B, 1C.

Next, (1B; Fig. 17) quality assurance (QA) of data in the Specify Workbench is undertaken by the curation officer or curator (note this is not undertaken by the trained volunteer entering the data), prior to uploading to the Specify CMS. Labels are then printed (1C; Fig. 17) and either affixed to the specimen or paperclipped if specimen conservation is required prior to imaging.

There may be a gap in time between the data collection in Stream 1 and taking the high-resolution image/s (section H; Fig. 18). Steps H1 and H2 include the image capture and

quality control (MELU uses Leaf Aptus imaging hardware and Leaf Capture software), and actions are required to archive the image/s for backup and then to connect the image/s to the data in Specify CMS. Following this, the next step (step P; Fig. 18) is the scripted provision of data/images in Specify to the online MELU collection portal and ALA.

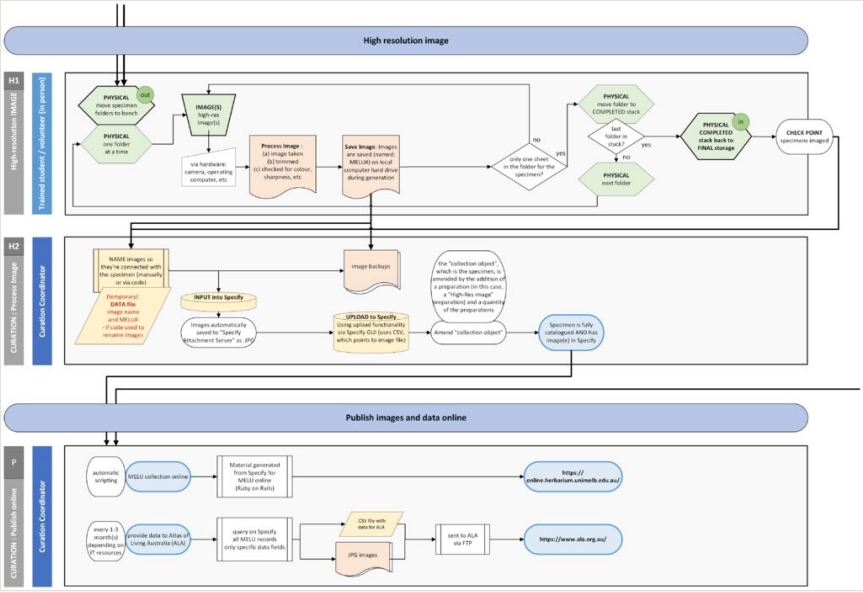

Figure 18. doi

MELU Digitisation Workflow: Detailed, parts H + P.

Stream 2 – data transcription from an image

The initial task (2A) in Stream 2 (Fig. 19) involves steps similar to 1A, but with only skeletal data being collected (MELU accession number, the current taxon name, the number of specimen sheets associated with the collection object, and the number of images for each collection object). The data is collected using a .CSV file (not Specify Workbench), and the current taxonomic name is only checked and recorded if transcription will be completed by volunteers who are not affiliated with MELU (e.g., citizen scientists).

Next, (2B; Fig. 19) data is temporarily input into a test (duplicate) Specify CMS, to enable a label containing the MELU accession number, or any annotation labels to be printed, if required. Then (2C; Fig. 19) an image is taken, using a fixed camera mount, mobile phone camera, plus wireless camera remote control. The time required for lower-resolution image generation is significantly less than that required for generation of the high-resolution images. While these initial modules require physical interaction with the collection specimens, it is no more than required for that of Stream 1. The images are then reformatted into the file size required by the software used to view the images for transcription.

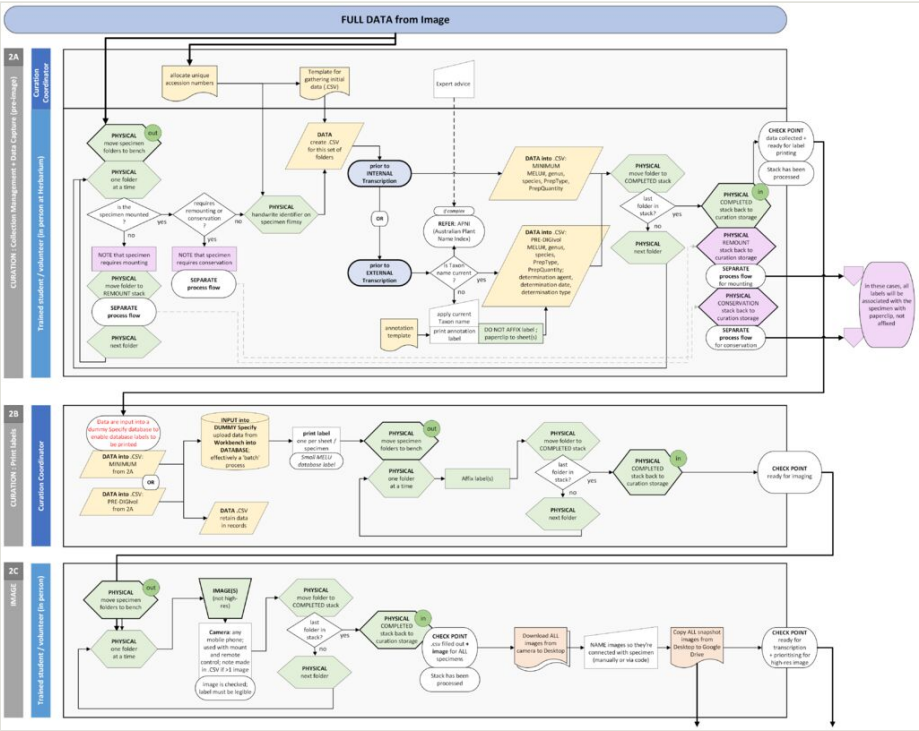

Figure 19. doi

MELU Digitisation Workflow: Detailed, parts 2A, 2B, 2C.

The next steps are to set up for and engage in the manual transcription of the data from the specimen images (modules 2D and 2E, Fig. 20). After module 2C, MELU's workflow can be undertaken entirely remotely if necessary or preferred (right side stream), either by internal (trained volunteers and students) or external (DigiVol; digivol.ala.org.au) transcribers. The potential for remote completion of workflow steps increases flexibility within this stream and reduces the number of times specimens are handled, the latter potentially reducing the average overall time required for progression through this transcription stream (recall the observation in Tann and Flemons 2008). The option of transcribing the data from the specimens in person in the herbarium remains because experience and surveys suggest "a clear preference … for working with physical specimens" (Drinkwater et al. 2014, p. 27). Where data are transcribed externally (i.e., by DigiVol citizen scientists), the data file (a .CSV file) is downloaded from the online interface through which the data are captured for upload into Specify CMS.

The penultimate task in Stream 2 is 2F (Fig. 20), quality assurance and data upload. Quality control of the specimen data can reference the specimen images, so the data can be uploaded into Specify remotely. The final task in this stream (2G; Fig. 20) is printing and attaching any new labels that are required, either full institutional labels where specimens were lacking that label, or annotation labels for accepted name changes. In the remote pathway, the taxon name is checked and an annotation label printed and affixed prior to

imaging (2B). If the transcription was undertaken by internally trained volunteers or students the taxon name was checked immediately prior to data transcription (2C), therefore annotation labels are printed and affixed after that step.

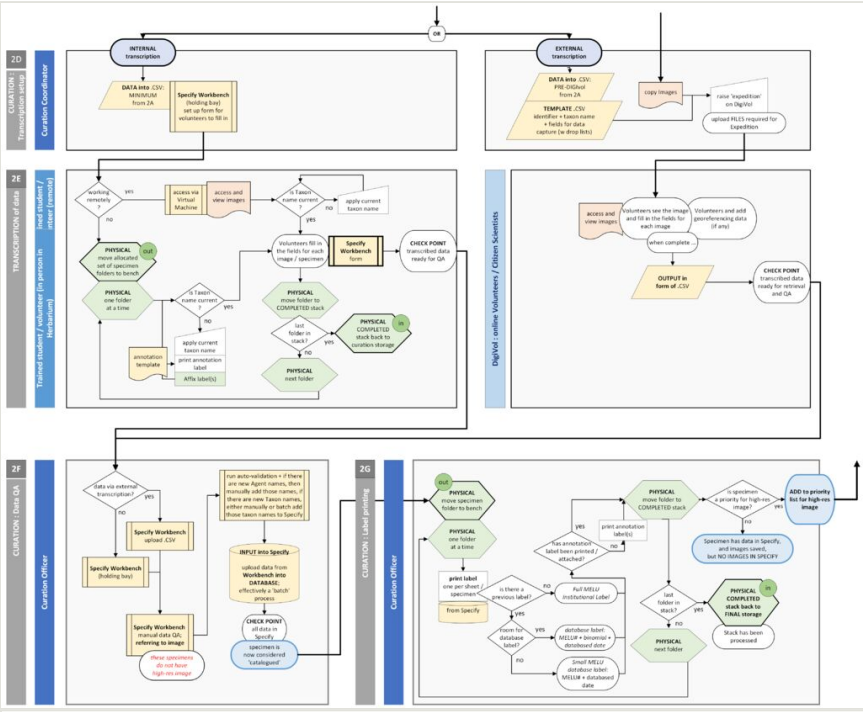

Figure 20. doi

MELU Digitisation Workflow: Detailed, parts 2D, 2E, 2F, 2G.

At the end of Stream 2, all collection data are uploaded into Specify CMS. The lower-resolution image/s are retained in collection records but are not uploaded into the CMS and are not publicly shared. In this way, this stream does not always immediately complete the entire digitisation workflow, as high-resolution images may not have yet been generated. The final decision regarding generation of high-resolution images suitable for online sharing is made based on collection curation priorities and staff/volunteer availability.

Stream 3 - born digital

Increasingly, data is entering herbarium databases soon after collection via digital records kept by the collector. Stream 3 (Fig. 21) outlines this emerging protocol; noting this is for material new to the MELU collection and MELU advises the collector about data format and structure. Data is entered by the collector into a .CSV template provided by the digitisation coordinator. This is more commonly done by referencing field notebooks, though the possibility of using dried specimens is included in the workflow. Following data collection, the file is ready for future integration into Stream 1 at step 1B (Fig. 17).

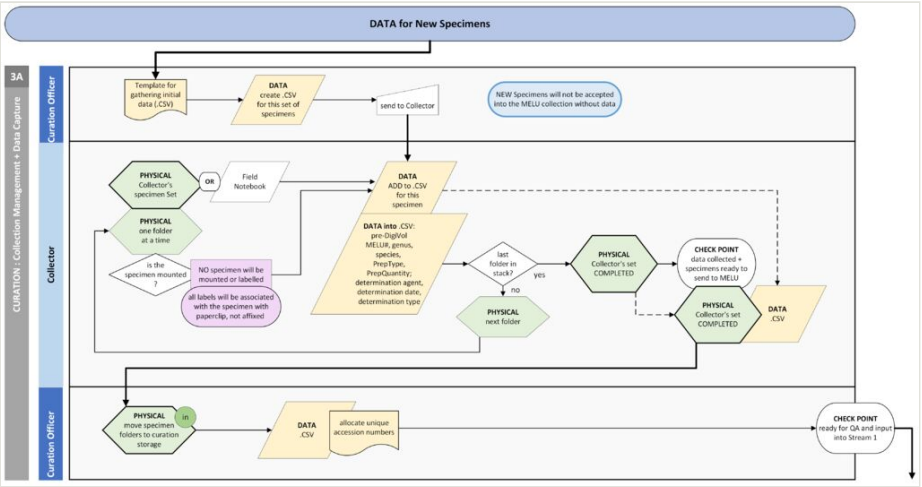

Figure 21. doi

MELU Digitisation Workflow: Detailed, stream 3.

Discussion

Digitisation relies on humans

Many digitisation workflow diagrams observed in the literature do not explicitly distinguish manual or human-mediated versus automated or scripted workflow steps. The MELU outline (and subsequent detailed maps) makes explicit the human-mediated steps in digitisation workflow, particularly the regular and iterative handling of the physical botanical specimens. In the outline map (Fig. 15), human-mediated tasks are highlighted by being marked with a green icon (with a 'p' or 'o' in the box), and by the green hexagonal and upside-down wedge (and scattered rectangle) elements in the detailed maps (Figs 17, 18, 19, 20, 21). This human element is made even more apparent by the proliferation of the green elements in the more complex detailed maps, from which the outline map was distilled. This can be fully appreciated when task text is removed from the overall detailed map for streams 1 and 2, and the green elements of specimen handling are highlighted (Fig. 22).

Specimen handling events in the digitisation workflow include tasks such as selection and collation of specimens for digitisation, generation of lower- or high-resolution images, affixing label, and refiling specimens. The placement of these human-mediated steps, inferred, but rarely annotated as such in workflow landscapes, has significant implications in terms of efficiency in the digitisation workflows. Specimen handling tasks are typically labour-intensive steps and many, such as specimen selection and refiling specimens into the collection cannot be eliminated. However, reducing the number of times the specimens are handled during the digitisation curation workflow, for example by reducing the requirement for (re-)sorting or (re-)filing specimens, introduces efficiencies and time

savings to the overall workflow. Good examples of the timesaving offered by reduction/s in specimen handling are the transition from a data-to-image to an image-to-data workflow and reliance on lower-resolution images, rather than high-resolution images, as a source for specimen transcription. The time requirement for generation of lower-resolution images is significantly less than that required for the generation of high-resolution images. Generation of digital images, albeit lower-resolution images, early in the digitisation workflow, enables sorting and searching of digital images, which can result in a significant timesaving over sorting and searching physical specimens. Generation of high-resolution specimen images is typically still a desired component of the digitisation workflow. Where lower-resolution specimen images are used for data curation, decisions regarding the allocation of resources and time to the generation of high-resolution specimen images can follow collection imaging priorities rather than data capture priorities. Where curation and digitisation resources are finite, as is the case in all collections, such efficiencies in the workflow can release staff time for other essential curation tasks.

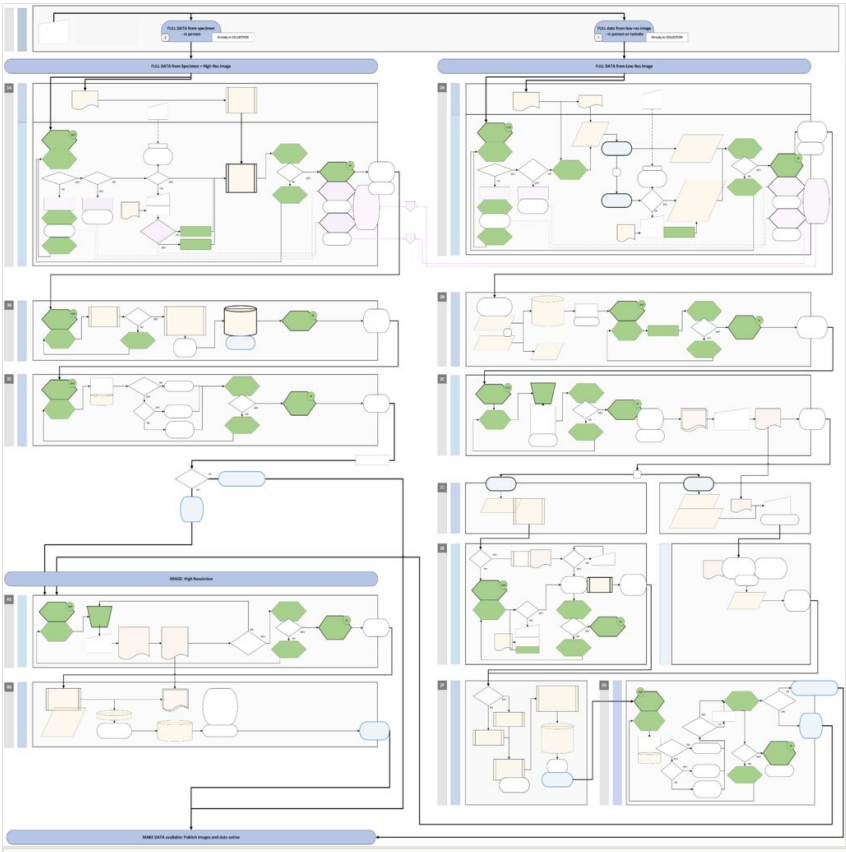

Figure 22. doi

MELU Digitisation Workflow: Detailed, without text and highlighting specimen handling aspects (green).

The coronavirus (COVID) pandemic, during 2020 and 2021, provided the impetus for a pivot to 'remote' curation and taxonomic work (e.g., Phang et al. (2022)) that few would have anticipated was possible prior to that event. As a result, collection managers were required to identify those curation tasks that could be conducted remotely and, conversely, those that could only be completed when workers were on-site in the collection. Both technology and protocol changes were amongst significant adaptations to curation tasks and workflows, in collections globally to ensure ongoing productivity while collection staff were working primarily remotely (Baker 2020). At MELU, some of those changes were in the planning stages and were fast-tracked in response to the immediate need to increase the capacity for remote curation work. While the CMS for MELU was not cloud-based, placement onto a virtual machine made remote access to collection records possible for Herbarium staff and student volunteers. The increased use of an image-to-data workflow (Stream 2) also enabled remote curation work. During the extended lockdowns enforced in Victoria, Australia in 2020 and 2021, a single staff member could safely work in person in the collection, completing specimen pre-processing, generating the lower-resolution images, and undertaking the initial (shorter) preparation steps necessary to enable subsequent, potentially remote, transcription efforts. In additional to transcription of primary label data, georeference data can be generated, and data quality can be checked and errors corrected based on available images (Baker 2020). In this way, MELU staff and student volunteers, like their colleagues in collections globally, were able to continue their digitisation work during what could have been, and was for some, a time of reduced productivity.

Technology and Tools

Any diagram, by its nature, is an abstraction of reality and may appear to imply that work simply flows from one task to the next. Representations necessarily omit detail and seem to suggest that connections are seamless. But it can be these very transitions, between modules and between tasks within modules, that can be the most difficult part of a digitisation project, for they often involve data format transformations, transfers between storage locations or software, which are time-consuming and may be points of highest risk for data loss. Additionally, workflows that require multiple format conversions between input and output data files are often not very resilient to workflow adjustments, which can limit the ease of maintenance and evolution of these workflows over time (Dou et al. 2012).

Detailed workflow maps permit the inspection of the technology required for each digitisation task and, subsequently, requirements for data transfer among software and storage or archival location/s. For example, the complexity of the CMS infrastructure, the software components, and the resulting curation steps involved in the MELU workflow around tasks H1 and H2 are detailed in Fig. 23. For MELU, like many herbaria, the digitisation workflow works within a complex data management system infrastructure; spanning individual objects in the collection, the digital records held in the CMS, image storage on physical hard drives, image servers and archives, and including interactions (data provision) with institutional and external repositories. Understanding the component technologies, and their input and output requirements, are essential for streamlining the

digitisation workflow to ensure format cross-compatibility among software in the workflow series and that standards for storage and archiving of images and their metadata are being met. While this example figure is specific to technology MELU currently uses, it is hoped that other collections can use the workflow maps to facilitate identification of tools enabling each task. Such clarity presents the opportunity to review where the process may benefit from software and/or hardware substitution, because best practices require that these workflows are both sustainable for maintenance in perpetuity, including for database management and migration over time (Thomer et al. 2019).

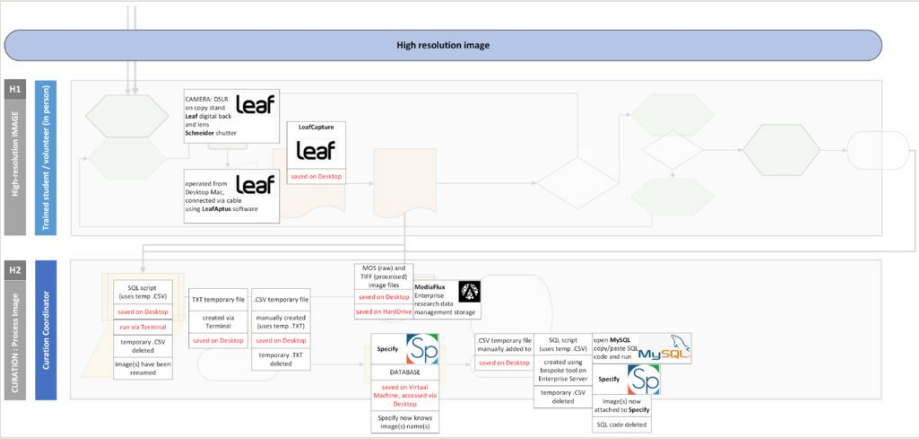

Figure 23. doi

MELU Digitisation Workflow: Detailed, part H: tools and technology currently used by MELU.

Bioinformatic requirements of digital extended specimens

What is evident from these landscape maps is the complexity of data handling requirements for all digitisation workflows. Comprehensively mapped workflows, as provided here, clearly illustrate the complexity and labour-intensity of managing not only the collection objects and their primary data, but also any derived objects and metadata, while also maintaining the links among these entities (e.g., “digital-extended specimens” (Hardisty et al. 2022)). Best practice for voucher-enabled biodiversity research data includes the retention of links among specimens in herbaria or museums and third-party repositories, to maintain an accurate taxonomic context for the derived research data. These links must be actively curated in perpetuity to maintain the significant value of the derived research data for reuse. While the workflow for born-digital specimen data (Nelson et al. 2012, p. 20), is more efficient as transcription of these data is not required, the challenge for these born-digital data is in subsequent required curation of the links among these specimens and their derived objects (e.g., microscope slides created from tissue destructively sampled from specimens or resulting micrographs of these preparations) or their derived data (e.g., DNA sequence or trait measurement data). Well documented and intuitive curation workflows are necessary for the efficient mobilisation and management of

specimen-associated digital objects or data to achieve the Digital Extended Specimen as a fully interconnected network of digital objects on the internet (Hardisty et al. 2022).

While this case-study has focused on digitisation of specimens and their primary data, well-documented digitisation landscapes such as those presented here can provide the necessary framework for subsequent mapping of workflow/s for digitisation of derived specimen objects and for in-house curation of specimen-derived research-associated data that are typically provided by researchers to third-party global repositories (e.g., GenBank, MorphoBank).

Data management represents an increasingly labour-intensive task for curators, which is a challenge for all collections, and in particular for small and medium size collections with limited curation staff. Mapping what currently remains a predominantly manual workflow enabled identification of the steps that hold potential for automation (e.g., H1; the scripted upload of images into the CMS and into the image storage archive). The increased availability of openly available workflows and software architectures with standardised interfaces that meet the information technology and archive requirements of natural history collections and that are customisable to meet the diverse needs of collections (e.g., Kurator, Dou et al. 2012; StanDAP-Herb, Kirchhoff et al. 2018), will be a valuable resource for small and medium size collections to potentially increase the ease and efficiency of these data management requirements.

Further Developments

For the time being MELU still relies on manual transcription of data from specimen sheets. As has already been noted, even the largest organisations (with arguably more funding) also appear to continue to invest in manual transcription. The ongoing engagement of citizen scientists for ‘remote’ elements of Stream 2 has been important for expanding opportunities for ongoing digitisation outputs at MELU. This approach is by no means intended as a replacement for mass-digitisation pathways seen in large institutions, but it is a ‘lightweight’ approach to transcribing from an image and making progress toward digitisation goals. It is suitable for small and medium collections because of this simplicity, and it is cost-effective to apply with minimal tool or technology changes. MELU is currently exploring what efficiencies may be introduced via machine-learning and -reading, believing that “information extraction from specimen labels [is] among the digitization workflow activities which can benefit from greater automation” (de la Hidalga et al. 2022, p. 2). These investigations include optical character recognition (OCR) and applying object-detection machine-learning models to read label data (Thompson et al., forthcoming; Turnbull et al., in draft).

Integration of the archival requirements of the vast amounts of data and digital files that are the result of digitisation efforts is also necessary to ensure these resources are curated and accessible across the research data life cycle. While current efforts continue to focus on the generation of the first set of digital files associated with physical specimens, ongoing study of those physical specimens may require the addition of an annotation label, for example, to denote a change in the taxonomic identity of the specimen or the sampling and

removal of material from the specimen. Version control of images and data becomes even more complex when both specimen data and images are shared with global repositories. New functionality will be required to enable curation and version control of digital extended specimens given the dispersion of objects and their data into multiple databases and repositories and ensure that current and consistent versions of those objects are accessible for curation and use globally (Hedrick et al. 2020).

Conclusions

Time for object and data curation is a precious commodity in all natural history and cultural collections. These digitisation workflows have contributed to ensuring the efficient use of curation time to achieve digitisation outputs and that digitisation standards are consistently applied within and among projects. The time taken to create these workflow maps was substantial, and admittedly more than anticipated at the outset, in part because visualising the pathways was more complex than was initially appreciated. However, the time invested has been worthwhile; they have already contributed significant value to MELU collections. Curation pathways have been optimised as a result of the work required to construct and visualise the documented workflows. Workflow construction provided opportunities for comparison of specimen curation steps among digitisation pathways, which facilitated recognition of the similarities and resulting modularity of these workflows. Significantly, these pathways no longer only exist in the mind of one or two experts and are instead visually available for reference, consideration, and improvement by curation team members. Finally, these workflows have been, and will continue to be, effective tools for communication with stakeholders outside the herbarium. They have illustrated the contextual framework of curation workflows and tasks necessary for collaborations with research data specialists and computer programmers working on tool development based on MELU collection based digital resources including for scripted access for extraction, analyses, and provision of MELU digital resources and data. Additional infrastructure is required, particularly for small- to medium-size collections to meet the increasing demands for high-quality collection associated biodiversity data. We hope these workflows are useful for other herbaria, for comparison, or to serve as a launching point for further workflow optimisation or development.

Acknowledgements

The authors acknowledge Melbourne Data Analytics Platform (MDAP) colleagues also involved in the MELU-MDAP collaboration project: Emily Fitzgerald, Robert Turnbull, Simon Mutch, Noel Faux, Bobbie Shaban; School of BioSciences colleagues: Heroen Verbruggen and Andrew Drinnan; Royal Botanic Gardens colleague Niels Klazenga; and MELU staff member Aiden Webb. The authors acknowledge the University of Melbourne Botany Foundation and the Russell and Mab Grimwade Miegunyah Fund for their financial support for digitisation in the University of Melbourne Herbarium.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Atlas of Living Australia (ALA) (2022) About Us. https://www.ala.org.au/about-ala/. Accessed on: 2022-8-04.
- Baird RC (2010) Leveraging the fullest potential of scientific collections through digitisation. Biodiversity Informatics 7 (2): 130-136. https://doi.org/10.17161/bi.v7i2.3987
- Baker B (2020) Biodiversity Collections, Data, and COVID. BioScience 70 (10): 841-847. https://doi.org/10.1093/biosci/biaa093
- Barber A, Lafferty D, Landrum L (2013) The SALIX Method: A semi-automated workflow for herbarium specimen digitization. Taxon 62 (3): 581-590. https://doi.org/10.12705/623.16
- Barkworth M, Murrell Z (2012) The US Virtual Herbarium: working with individual herbaria to build a national resource. ZooKeys 209: 55-73. https://doi.org/10.3897/zookeys.209.3205
- Beaman R, Cellinese N (2012) Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys 209: 7-17. https://doi.org/10.3897/zookeys.209.3313
- Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A (2017) Going deeper in the automated identification of Herbarium specimens. BMC Evolutionary Biology 17 (1): 1-14. https://doi.org/10.1186/s12862-017-1014-z
- Corney DA, Clark J, Tang HL, Wilkin P (2018) Automatic extraction of leaf characters from herbarium specimens. TAXON 61 (1): 231-244. https://doi.org/10.1002/tax.611016
- Cox L (2022) 'Heavy lifting at Sydney's herbarium: the quest to move and catalogue more than 1m plant specimens'. https://www.theguardian.com/australia-news/2022/jan/12/heavy-lifting-at-sydneys-herbarium-the-quest-to-move-and-catalogue-more-than-1m-plant-specimens. Accessed on: 2022-7-30.
- de la Hidalga AN, Rosin P, Sun X, Livermore L, Durrant J, Turner J, Dillen M, Musson A, Phillips S, Groom Q, Hardisty A (2022) Cross-validation of a semantic segmentation network for natural history collection specimens. Machine Vision and Applications 33 (3): 1-31. https://doi.org/10.1007/s00138-022-01276-z
- Dou L, Cao G, Morris PJ, Morris RA, Ludascher B, Macklin JA, Hanken J, (2012) Kurator: A Kepler Package for Data Curation Workflows. Procedia Computer Science 9: 1614-1619. https://doi.org/10.1016/j.procs.2012.04.177
- Drinkwater R, Cubey R, Haston E (2014) The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels. PhytoKeys 38: 15-30. https://doi.org/10.3897/phytokeys.38.7168
- Escobar H (2018) In a 'foretold tragedy,' fire consumes Brazil museum. Science 361 (6406): 960-960. https://doi.org/10.1126/science.361.6406.960
- Fernández-Fontecha A, O'Halloran KL, Tan S, Wignell P (2018) A multimodal approach to visual thinking: the scientific sketchnote. Visual Communication 18 (1): 5-29. https://doi.org/10.1177/1470357218759808

- Granzow-de la Cerda Í, Beach JH (2010) Semi-automated workflows for acquiring specimen data from label images in herbarium collections. Taxon 59 (6): 1830-1842. https://doi.org/10.1002/tax.596014
- Hardisty A, Brack P, Goble C, Livermore L, Scott B, Groom Q, Owen S, Soiland-Reyes S (2022) The Specimen Data Refinery: A Canonical Workflow Framework and FAIRDigital Object Approach to Speeding up Digital Mobilisation of Natural HistoryCollections. Data Intelligence 4 (2): 320-341. https://doi.org/10.1162/dint_a_00134
- Hardisty AR, Ellwood ER, Nelson G, Zimkus B, Buschbom J, Addink W, Rabeler RK, Bates J, Bentley A, Fortes JAB, Hansen S, Macklin JA, Mast AR, Miller JT, Monfils AK, Paul DL, Wallis E, Webster M (2022) Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. BioScience 72 (10): 978-987. https://doi.org/10.1093/biosci/biac060
- Haston E, Cubey R, Pullan M, Atkins H, Harris D (2012) Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. ZooKeys 209: 93-102. https://doi.org/10.3897/zookeys.209.3121
- Haston E, Albenga L, Chagnoux S, Drinkwater S, 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 Bars G, Livermore L, Mühlenberger G, Paul D, Philips S, Smirnova L, Vacek F (2015) Automating data capture from natural history specimens: SYNTHEYS 3 Work Package 4. URL: https://synthesys3.myspecies.info/node/695
- Hedrick BP, Heberling JM, Meineke EK, Turner KG, Grassa CJ, Park DS, Kennedy J, Clarke JA, Cook JA, Blackburn DC, Edwards SV, Davis CC (2020) Digitization and the Future of Natural History Collections. BioScience 70 (3): 243-251. https://doi.org/10.1093/biosci/biz163
- JSTOR (2018) JSTOR Plants Handbook. URL: http://www.snsb.info/SNSBInfoOpenWiki/attach/Attachments/JSTOR-Plants-Handbook.pdf
- Kalms B (2012) Digitisation: A strategic approach for natural history collections. CSIRO, Canberra, Australia. URL: http://www.ala.org.au/wp-content/uploads/2011/10/Digitisation-guide-120223.pdf
- Kirchhoff A, Bügel U, Santamaria E, Reimeier F, Röpert D, Tebbje A, Güntsch A, Chaves F, Steinke K, Berendsohn W (2018) Toward a service-based workflow for automated information extraction from herbarium specimens. Database 2018: 1-11. https://doi.org/10.1093/database/bay103
- Marsico T, Krimmel E, Carter JR, Gillespie E, Lowe P, McCauley R, Morris A, Nelson G, Smith M, Soteropoulos D, Monfils A (2020) Small herbaria contribute unique biogeographic records to county, locality, and temporal scales. American Journal of Botany 107 (11): 1577-1587. https://doi.org/10.1002/ajb2.1563
- Moen W, Huang J, McCotter M (2010) Extraction and parsing of herbarium specimen data: Exploring the use of the Dublin core application profile framework. Illinois Library. URL: https://hdl.handle.net/2142/14920
- Mononen T, Tegelberg R, Sääskilahti M, Huttunen M, Tähtinen M, Saarenmaa H (2014) DigiWeb - a workflow environment for quality assurance of transcription in digitization of natural history collections. Biodiversity Informatics 9 (1): 18-29. https://doi.org/10.17161/bi.v9i1.4748

- Nelson G, Paul D, Riccardi G, Mast A (2012) Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys 209: 19-45. https://doi.org/10.3897/zookeys.209.3135
- Nelson G, Sweeney P, Wallace L, Rabeler R, Allard D, Brown H, Carter JR, Denslow M, Ellwood E, Germain-Aubrey C, Gilbert E, Gillespie E, Goertzen L, Legler B, Marchant DB, Marsico T, Morris A, Murrell Z, Nazaire M, Neefus C, Oberreiter S, Paul D, Ruhfel B, Sasek T, Shaw J, Soltis P, Watson K, Weeks A, Mast A (2015) Digitization workflows for flat sheets and packets of plants, algae, and fungi. Applications in Plant Sciences 3 (9): 1500065. https://doi.org/10.3732/apps.1500065
- Nelson G, Ellis S (2018) The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B: Biological Sciences 374 (1763): 20170391. https://doi.org/10.1098/rstb.2017.0391
- Nieva de la Hidalga A, Rosin P, Sun X, Bogaerts A, De Meeter N, De Smedt S, Strack van Schijndel M, Van Wambeke P, Groom Q (2020) Designing an Herbarium Digitisation Workflow with Built-In Image Quality Management. Biodiversity Data Journal 8: e4705. https://doi.org/10.3897/bdj.8.e47051
- Owen D, Groom Q, Hardisty A, Leegwater T, van Walsum M, Wijkamp N, Spasić I (2019) Methods for automated text digitisation. Zenodo. https://doi.org/10.5281/zenodo.3364501
- Phang A, Atkins H, Wilkie P (2022) The effectiveness and limitations of digital images for taxonomic research. TAXON 71 (5): 1063-1076. https://doi.org/10.1002/tax.12767
- Phillips S, Green L, Weech MH (2014) Review of digitisation workflows and equipment. URL: https://synthesys3.myspecies.info/sites/synthesys3.myspecies.info/files/NA3%20Del.%203.3%20Review%20of%20Digitisation%20workflows%20and%20equipment_0.pdf
- Popov D, Roychoudhury P, Hardy H, Livermore L, Norris K (2021) The Value of Digitising Natural History Collections. Research Ideas and Outcomes 7: e78844. https://doi.org/10.3897/rio.7.e78844
- Staight K (2017) Irreplaceable plant specimens from France destroyed in Australian quarantine blunder. https://www.abc.net.au/news/2017-05-08/irreplaceable-plant-specimens-destroyed-by-biosecurity-officers/8504944. Accessed on: 2022-8-01.
- Stokstad K (2017) Botanists fear research slowdown after priceless specimens destroyed at Australian border. https://www.science.org/content/article/botanists-fear-research-slowdown-after-priceless-specimens-destroyed-australian-border. Accessed on: 2022-8-01.
- Sweeney P, Starly B, Morris P, Xu Y, Jones A, Radhakrishnan S, Grassa C, Davis C (2018) Large–scale digitization of herbarium specimens: Development and usage of an automated, high–throughput conveyor system. TAXON 67 (1): 165-178. https://doi.org/10.12705/671.10
- Tann J, Flemons P (2008) Data capture of specimen labels using volunteers (Australian Museum). URL: http://australianmuseum.net.au/Uploads/Documents/23183/Data%20Capture%20of%20specimen%20labels%20using%20volunteers%20-%20Tann%20and%20Flemons%202008.pdf
- Tegelberg R, Haapala J, Mononen T, Pajari M, Saarenmaa H (2012) The development of a digitising service centre for natural history collections. ZooKeys 209: 75-86. https://doi.org/10.3897/zookeys.209.3119

- Tegelberg R, Mononen T, Saarenmaa H (2014) High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. TAXON 63 (6): 1307-1313. https://doi.org/10.12705/636.13
- Thiers B, Tulig M, Watson K (2016) Digitization of The New York Botanical Garden Herbarium. Brittonia 68 (3): 324-333. https://doi.org/10.1007/s12228-016-9423-7
- Thomer A, Weber N, Twidale M (2019) Supporting the long-term curation and migration of natural history museum collections databases. Proceedings of the Association for Information Science and Technology 55 (1): 504-513. https://doi.org/10.1002/pra2.2018.14505501055
- Triki A, Bouaziz B, Mahdi W, Gaikwad J (2020) Objects Detection from Digitized Herbarium Specimen based on Improved YOLO V3. Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications523-529. https://doi.org/10.5220/0009170005230529
- Tulig M, Tarnowsky N, Bevans M, Kirchgessner A, Thiers B (2012) Increasing the efficiency of digitization workflows for herbarium specimens. ZooKeys 209: 103-113. https://doi.org/10.3897/zookeys.209.3125
- Unger J, Merhof D, Renner S (2016) Computer vision applied to herbarium specimens of German trees: testing the future utility of the millions of herbarium specimen images for automated identification. BMC Evolutionary Biology 16 (248): 1-7. https://doi.org/10.1186/s12862-016-0827-5
- Walton S, Livermore L, Bánki O, Cubey R, Drinkwater R, Englund M, Goble C, Groom Q, Kermorvant C, Rey I, Santos C, Scott B, Williams A, Wu Z (2020) Landscape Analysis for the Specimen Data Refinery. Research Ideas and Outcomes 6 (2): e56211. https://doi.org/10.3897/rio.6.e57602
- White A, Dikow R, Baugh M, Jenkins A, Frandsen P (2020) Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning. Applications in Plant Sciences 8 (6): e11352. https://doi.org/10.1002/aps3.11352
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Supplementary material

Suppl. material 1: MELU Digitisation Workflow - Whole picture doi

Authors: Karen M Thompson and Joanne L Birch
Data type: Image
Download file (931.13 kb)