Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/135180/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Hardisty, Alex ORCID: https://orcid.org/0000-0002-0767-4310, Livermore, Laurence, Walton, Stephanie, Woodburn, Matt and Hardy, Helen 2020. Costbook of the digitisation infrastructure of DiSSCo. Research Ideas and Outcomes 6 , e58915. 10.3897/rio.6.e58915 file

Publishers page: https://dx.doi.org/10.3897/rio.6.e58915 <https://dx.doi.org/10.3897/rio.6.e58915>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

information services
gwasanaethau gwybodaeth

4. Discussion

The process of data collection for this study revealed complexities in gathering and assessing accurate cost data. First, there were inconsistences in how workflows are named and categorised. In asking for the specimen type, one institution used 'mycological' and another used 'fungarium' to describe digitising their fungi collection. The first institution categorised this workflow as a herbarium collection and the latter as 'Other'. This is indicative of limitations and inconsistencies in the terminologies used to describe collections and, subsequently, how they are categorised and analysed. For the purposes of this study, both were categorised as 'fungarium'.

An inconsistent approach to describing collections of physical specimens is a wider challenge that the natural science community is attempting to address. While many efforts have been made within and across institutions to generate and share collection descriptions data, the lack of common standards, data model and vocabularies remain a significant barrier to making these datasets comparable and interoperable. The terminology issues described above are a result of this lack of consistency and standardisation across institutional practices.

The Biodiversity Information Standards organisation TDWG, (https://www.tdwg.org) is developing a new Collection Description data standard to support harmonisation of data across these various resources, and using collection descriptions to underpin specimen digitisation activities is one of the major use cases for the standard (Raes et al. 2019). The scope is intended to encompass the use of common terms and vocabularies to describe properties such as the type of objects within a collection, their preservation method and their taxonomic classification, which would provide more specific and relevant data for mapping to the appropriate digitisation workflows. If adopted by institutions and applied to the costbook framework, this should largely resolve the issue of collections being classified differently in a way that reflects each individual institution's organisational hierarchy.

Secondly, different workflows were broken out into separate cost books. However, some institutions recorded the same number of employees across multiple workflows and, in some cases, the same time and costs associated with different collections. It is unclear if these were separate but identical costs that could thus be summed, or if they were the same costs and thus a double counting of the same data.

ICEDIG recommends working towards harmonisation of approaches to costing digitisation. This will become more important as various kinds of decision about digitisation are made e.g., prioritisation, allocation of certain types of mass digitisation to specific facilities, budgeting, authorization of on-demand digitisation requests, etc.

For categories of collection where digitisation has been carried out by a significant number of institutions, it's reasonable to look at the spread of costs achieved and to focus on transferring knowledge and learning points from those institutions of low cost to those where costs are higher, in an effort to increase cost efficiencies.

For categories of collection where digitisation has been carried out by only a few institutions, the aim should be to spread best practice to institutions embarking on digitisation of these categories as a means to avoid repeating past mistakes and accelerating progress towards efficient (low-cost) digitisation across institutions in those categories.

Recommendations on capital equipment choices, whilst probably appropriate for DiSSCo to give guidance on, is out of scope of the present document.

Based on this costbook exercise an ambitious baseline for mass digitisation of pinned and herbarium sheets would be less than €0.50 PPS per item. This is based on a very limited sample of institutions and workflows so should be taken as indicative only. There is not enough data to make suggestions on baseline costs for digitising other specimens but in order to meet DiSSCo's mass digitisation goals we need to encourage and support continuous improvements to drive that cost down and to increase throughput without increasing per item cost. In practice, also, digitisation projects vary widely, and the degree of data captured should relate to the project aims – where more data is most appropriate (e.g., a key project aim is full georeferencing or some kind analytical treatment of an object) it may well be appropriate to accept a higher baseline cost.

5 Recommendations

In addition to the discussion points above we recommend the following:

- Focus on harmonisation of costing approach – standardisation of the methodology for gathering and reporting costs. We recognise that many institutes will have difficulty gathering and providing detailed cost information and that a simpler costing approach may be required.
- Focus on cost improvements (efficiencies) – recommend setting a target mass digitisation cost (per specimen) for different types of collection. If we had to set it today, what would we set it at? A strong focus on cost improvement would be one of several means of accelerating progress in mass digitisation.
- Consider how we can transfer best practice between institutes and digitisation teams.
- Track digitisation costs over time as standard - we currently have limited data on digitisation costs and if more institutes started recording this data we could better identify effective and ineffective practice.

5.1 Expanding scope of gathered costs

Anthropological, Palaeontological, Mineralogical and non-insect Invertebrates collections were not included in the scope of ICEDIG digitisation research. While non-insect

invertebrates are a major collection type, they were erroneously omitted from the scope of Cocks et al. (2020) which was used as the basis of this report.

5.2 Other considerations

The costbook work in ICEDIG will be inherited and expanded upon by the DiSSCo Prepare project, specifically in Tasks 4.1 and 4.2, the “Costbook for DiSSCo” and “Cost model for charging services”, and their corresponding reports.

While not directly working on a costbook, SYNTHESYS+ will be gathering and assessing cost data as part of the new Virtual Access workpackage (Hardy et al. 2020,Smith et al. 2019).

In the subsections that follow, we offer some further considerations that other projects in the DiSSCo Programme portfolio should take into account but they apply to any organised large scale digitisation of collections.

5.2.1 Future development and maintenance for collecting new data

The current method for collecting, aggregating and analysing data from different institutions, based on completing pre-formatted spreadsheet templates becomes cumbersome when the number of responding institutions increases and quantities of data increase. Significant manual work is involved both for the institutions in filling templates and for analysts to work with the returned data.

An alternative approach to spreadsheets

As we noted when considering implementation of the costbook template (see section 2.6), alternative approaches are available and should be considered. One such is Airtable (https://airtable.com/), a modern and flexible spreadsheet-database hybrid offered ‘as-a-service’ that allows teams to collaborate in the contribution and analysis of data. With both free and paid options, Airtable presents like a cloud spreadsheet (like Google Sheets) but also supports linking between sheets to form basic relational data structures, providing some of the benefits of a database. Table 10 summarises some pros and cons of using Airtable as the basis for future costbook work, based on the experiences of NHMUK, which has a paid subscription to the product/service.

Regardless of whether Airtable is the specific correct product to adopt, the key learning point is that reliance on old-style spreadsheet products, distributed and managed as files among participants is no longer necessarily the most flexible, efficient or sustainable approach to gathering, collating, analysing and using actual cost information. The recommendation here is that DiSSCo should consider alternatives to the Excel/Google spreadsheets approach for modern management of cost information. However, any change from using commonly used software to a new webform or database will require sufficient support to ensure it is fit for purpose.

Recommendation: DiSSCo must evaluate and adopt modern alternative(s) to traditional spreadsheet approaches for the management of cost information.

Table 10.

Pros and cons of Airtable for costbook work.

Pros
1. Data structure and interfaces support one to many relationships in the Institution/Facility/Fixed Costs/ Variable Costs data model, which would probably require scripting (with associated security/permissions challenges) in Excel or Google Sheets 2. Supports calculated fields, enabling spreadsheet calculations to be replicated 3. Provides basic form interfaces for data entry and grid interfaces for data management and querying – this should be intuitive for new and experienced users 4. Accessible online for submitting and managing data 5. Provides an API for programmatic access (e.g., custom forms, power business intelligence reports) 6. Data managed in a single location, which: ◦ Reduces data management overheads (e.g. chasing down multiple Excel files and extracting data from each) ◦ Enables aggregation (roll-up) and analysis (drill-down) across institutions and facilities ◦ Enables future design changes without having to distribute new Excel files and handle legacy versions
Cons
1. Cannot display calculated fields in form view, only grid view. 2. Cannot edit an existing record in form view, only grid view. 3. Native form views are quite simplistic and linear. 4. Must have a different data entry form for each table, rather than a consolidated form where, for exampleone can add a facility, and then multiple variable costs records, without leaving the form. 5. Airtable is not free. The paid option is needed to gain access to all functions.

Standardising currency

Several currencies have been used throughout the cost gathering and analysis work. The NHM UK entered their data in £ sterling. Other institutions entered their data in € euros. For summation, conversions were done to the EC's PPS Purchasing Power Standard. However, we failed to foresee that we might want to do some analytical calculations, for example stating specific cost components, such as depreciation as proportions (%) of a total annual cost. This involves going back and re-manipulating specific parts of the data.

A more helpful approach would be to convert from the currency used for data entry to PPS for each data item entered, at the time of entry. This would facilitate the kind of calculation exampled above.

Recommendation: In cost gathering, budgeting and accounting, DiSSCo should convert, *at the time of data entry* from the currency of data entry to the standard currency used for accounting purposes.

5.2.2 Sharing best practice

As we noted in the results and analysis, there are clear differences in costs that are most likely a consequence of the differing workflow approaches adopted by different institutions. Constant innovation leads ultimately to either/both higher throughput efficiencies and/or lower costs.

It is evident from anecdotal comments received during the task that practices for recording and breaking out costs, levels of detail of cost records and maturity of accounting for work vary considerably among the responding institutions.

Two elements to communicate best practices about:

1. Best practice accounting procedures so that quality and level of detail/accuracy of costing data improves
2. Innovations that lead to higher efficiencies/throughputs and lower costs

How then should DiSSCo distil, promote and support dissemination of best practices from established workflows in institutions with high efficiencies and low costs to other institutes that might benefit?

5.2.3 Treating costs separately from charges

Costs must be treated separately from charges. A cost model is not the same as a charging or business model, and the latter is not part of the present task. Nevertheless, in the end, cost calculations cannot be considered in isolation from a business/charging/organisational model, because of the influence of DiSSCo governance decisions and policy on requirements for digitisation, data access and availability. Digitisation can be required to a certain level. Some data may be more immediately available than other data, according to scientific demand and difficulty to retrieve (faster and easier versus slower and more time-consuming).

In-depth analysis of potential business models is described in Hardisty et al. 2020, being tied closely to questions of DiSSCo organisation and governance. Nevertheless, it is most likely that the DiSSCo business model will use the fundamental assumption that data must be ‘free at the point of use’ i.e., at no charge to the ultimate end-user. Within such a constraint, various charging models are conceivable, including for example: i) a research subscription model, whereby an institution or project wanting to provide its research staff with access to digital collections data pays a subscription for DiSSCo user membership; ii) an open-access model, whereby those demanding digitisation of specimens pay for that e.g., through a funded digitisation programme; or iii) an extension/re-orientation of the current loans and visits model, whereby costs of organising loans/visits are re-allocated providing FAIR access to digital content. This latter option assumes that as more specimen data becomes digitally available, the need for physical loans and visits diminishes.

Any business model must, however, take both depreciation and amortization into account.

5.2.4 Depreciation and amortization

Depreciation of equipment

Depreciation is the process of allocating the capital costs of a tangible asset (such as digitisation equipment or storage systems) over time. It's a measure of how much of the value of an asset has been consumed to a point in time (usually, the end of an accounting period). Note though, that usage of such equipment can usually extend well beyond the depreciation period. Depreciation is well understood and, especially for IT infrastructure, is typically allocated over three or four years using a straight-line method (i.e. the same amount in each year).

Depreciation is used in statutory accounting for matching costs against income and hence for calculating annual profit or loss. Its use in management accounting (as considered here) is as a means of reflecting the true cost of digitising specimens in years following those in which a digitisation facility was established.

Amortization of DiSSCo data

Amortization is the process of allocating the costs of an intangible asset such as data over time (its 'useful life'). The purpose is to match the costs of creating and maintaining data to the value earned from using that data. Or to put it another way, to ensure that expenses are not incurred in maintaining data with no useful value. Like depreciation, accounting for amortization in multi-year business plans for digitisation is good practice. Because of the multi-stakeholder characteristics of the DiSSCo governance and business model, this is a topic DiSSCo must pay attention to – however this is an area of high complexity where evidence is likely to improve over time.

Accounting for amortization in DiSSCo must match the expense of acquiring, preserving and maintaining 'FAIR'3 digitised specimen/collection data with the value of the use that data receives over time, usually in a linear fashion over the period of 'useful life'. Such value, however, can be hard to measure in financial terms - the value of research, education/training and other uses is not usually measured financially, partly because there are no accepted standard methods for doing so. Proxy measures can be useful; such as the number and impact of scientific publications achieved from having the data available; or the number and value of new research grants enabled by digitisation. Such metrics must be tracked from an early stage by the Digitisation Dashboard application.

We know the useful life of physical specimens in collections can easily be measured in decades or hundreds of years. But we also know the usefulness of both individual specimens and collections of specimens varies enormously, according to the scientific and societal questions of the day. What is the useful life of Digital Specimens and Digital Collections? For arriving at a practical basis for valuation and amortization, we must model several scenarios where amortization periods are set at say, 10, 25 and 50-year intervals.

In future, large-scale (mass) and more 'bespoke' digitisation can both be operated more frequently on a digitisation-on-demand basis, i.e. fulfilling demands for specimen

information by immediately digitising it and making it available on request on efficient digital platforms. There are arguments that this is more cost-effective: adapting words from elsewhere5, we could say that immediate digitisation is better than storage, meaning that it is more cost-effective to rapidly digitise and deliver only what is requested than to systematically and slowly digitise and store everything that is collected. In practice, however, experience to date of systematic digitisation is that its benefits are not always predictable – there is a strong element of serendipity e.g. in use of collections data alongside other data via aggregators; and there can be 'critical mass' of data for certain kinds of research ('big data' approaches). Sometimes, demand does not exist until data is made available, and data availability can enable new research paradigms and stimulate future demand. NHMUK's Digital Collections Programme, for example, track citations of digital specimen data – these data have not been created on demand, but the trend in the growth of usage (and therefore benefit/impact) is increasing year on year.

Once digitised, the value of specimen data does not decay quickly. Indeed, the value can even be increased as digitised specimen data is improved and supplemented with links to other information. There are costs associated with this. First, the costs of digitisation; second, the key cost of storage/preservation/serving over long time periods; and third, additional costs associated with data improvement and supplementation. There must be enough steady and measurable benefit over long periods into the DiSSCo business model to balance costs. An additional complexity is across what 'body' of data it is meaningful or accurate to apply amortization– the 'value' or benefit of data tends to increase in the context of other data, whether through an increase in the size of the same dataset; additional data from related collections datasets; or data from other sources and of other types/content e.g. climate data. While each digitisation project may look at their own dataset for amortisation and to estimate costs, the benefits and value do not accrue in isolation. Thus, the approach towards amortizing costs of data for DiSSCo must be examined very carefully and kept under review over time.

6. Conclusions

Considering the complexities of the digitisation process, and its variability among institutions and between different types of collections, we conclude that time spent is an essential parameter informing costing information. Other key parameters are labour rates, consumables and fixed cost elements such as heating and lighting, space rental, etc. Actual costs vary from one institution/country to another and our template offers calculators based on simple inputs. Gathered costs can be normalised to take account of different purchasing power of money in different countries.

Optimal digitisation cost is achieved when the volume and availability of specimens ready for digitisation matches the capacity of the digitisation facility. Having enough specimens ready means the digitisation capacity can be effectively utilised and the highest throughput can be achieved, thus leading to the lowest cost (*notwithstanding other factors contributing to cost and the assumption that the digitisation facility is dimensioned sufficiently for the*

task). Too few specimens ready means the capacity is underutilised, meaning higher cost per specimen.

What an institution wants to know is: When can certain kinds of digitisation be achieved for specific levels of investment? When does it become practical/economic to start digitising a collection? What does it cost to invest for digitisation and to reach a certain level for a collection?

The gathered cost information begins to inform answers to such questions. We have made several recommendations to be carried forward elsewhere in the DiSSCo Programme e.g., as specific work items in the DiSSCo Prepare project, for consideration by the DiSSCo Coordination and Support Office and the DiSSCo General Assembly.

Acknowledgements

We express our thanks and acknowledgement to the following individuals who assisted with this report and the underlying data:

- Hannu Saarenmaa (UH), Ana Casino (CETAF), Xavier Vermeersch (CETAF), Karsten Gödderz (CETAF), Luc Willemse (Naturalis), Michel Guiraud (MNHN), Agnes Wijers (PIC) and Jeroen Bloothoofd (PIC) for contributions towards conception, design and review of the costbook template.
- Lousie Allan (NHM) for attempting the completion of a trial costbook sheet to help us iron out difficulties.
- Quentin Groom (APM), Mathias Dillen (APM), Anne Koivunen (LUOMUS), Kari Lahti (LUOMUS), Sarah Philips (RBGK), Lousie Allan (NHM), Veljo Runnel (UTARTU) and Vanessa Demanoff (MNHN) for filling and returning 22 completed templates.

Funding program

H2020-EU.1.4.1.1. - Developing new world-class research infrastructures

Grant title

ICEDIG – “Innovation and consolidation for large scale digitisation of natural heritage”, Grant Agreement No. 777483

Author contributions

Authors:

Alex Hardisty: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft, Writing – Review & Editing. **Laurence Livermore**: Conceptualization, Data Curation, Formal Analysis, Writing – Original Draft, Writing – Review & Editing. **Stephanie Walton**: Data Curation, Formal Analysis, Validation, Visualization, Writing – Original Draft. **Matt Woodburn**: Software, Writing – Original Draft. **Helen Hardy**: Writing – Original Draft, Writing – Review & Editing.

Contribution types are drawn from CRediT - Contributor Roles Taxonomy.

References

- Allan E, Livermore L, Price B, Shchedrina O, Smith V (2019) A Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides. Biodiversity Data Journal 7 (e32342). https://doi.org/10.3897/BDJ.7.e32342
- Cocks N, Livermore L, Smith V, Woodburn M (2020) Technical capacities of digitisation centres within ICEDIG participating institutions. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e55522
- Dupont S, Price B (2019) ALICE, MALICE and VILE: High throughput insect specimen digitisation using angled imaging techniques. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37141
- Hardisty A (2019) Provisional Data Management Plan for DiSSCo infrastructure. Deliverable D6.6. Zenodo. https://doi.org/10.5281/zenodo.3532937
- Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul D, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e54280
- Hardy H, Knapp S, Allan EL, Berger F, Dixey K, Döme B, Gagnier P, Frank J, Haston E, Holstein J, Kiel S, Marschler M, Mergen P, Phillips S, Rabinovich R, Sanchez Chillón B, Sorensen M, Thines M, Trekels M, Vogt R, Wilson S, Wiltschke-Schrotta K (2020) SYNTHESYS+ Virtual Access - Report on the Ideas Call (October to November 2019). Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e50354
- Nelson G, Paul D, Riccardi G, Mast AR (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
- Raes N, van Egmond E, Casino A, Woodburn M, Paul DL (2019) Towards a Global Collection Description Standard. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37894
- Ranatunga D, Milne P, Birch J (2018) Digitise This! Innovation in Digitisation Initiatives within Australasian Herbaria. Biodiversity Information Science and Standards 2 (e26077). https://doi.org/10.3897/biss.2.26077

- Smith V, Gorman K, Addink W, Arvanitidis C, Casino A, Dixey K, Dröge G, Groom Q, Haston E, Hobern D, Knapp S, Koureas D, Livermore L, Seberg O (2019) SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes 5 https://doi.org/10.3897/rio.5.e46404
- van Egmond E, Willemse L, Paul D, Woodburn M, Casino A, Vermeersch X, Gödderz K, Bloothoofd J, Wijers A, Raes N (2019) Design of a Collection Digitisation Dashboard. Zenodo https://doi.org/10.5281/zenodo.2621055
- van Walsum M, Wijers A, Kamminga P, van der Mije S, van Dorp K (2019) State of the art and perspectives on mass imaging of liquid samples. Zenodo https://doi.org/10.3897/zookeys.209.3135
- Walton S, Livermore L, Dillen M, De Smedt S, Groom Q, Koivunen A, Phillips S (2020) A cost analysis of transcription systems. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e56211
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (160018). https://doi.org/10.1038/sdata.2016.18
- Wu Z, Kahanpää J, Sihvonen P, Koivunen A, Saarenmaa H (2019) Automated Methods in Digitisation of Pinned Insects. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.38260

Supplementary materials

Suppl. material 1: Cost Template Data Extraction Script doi

Authors: Matt Woodburn
Data type: VBA script
Brief description: Short VBA procedure for extracting data from multiple Excel template sheets into a flattened structure.
Download file (1.08 kb)

Suppl. material 2: Costbook Template doi

Authors: Alex Hardisty
Data type: Excel Spreadsheet
Brief description: This costbook template contains separate calculators for establishment (upfront) costs, for fixed costs of digitisation and for variable costs. We strongly recommend that before using again, to modify the costbook template to allocate named ranges to cells.
Download file (31.72 kb)

Suppl. material 3: Cost Books - Flattened Data doi

Authors: Matt Woodburn
Data type: Excel spreadsheet
Download file (27.63 kb)

Suppl. material 4: Cost Books - Original Responses doi

Authors: Quentin Groom (APM), Mathias Dillen (APM), Anne Koivunen (LUOMUS), Kari Lahti (LUOMUS), Sarah Philips (RBGK), Lousie Allan (NHM), Veljo Runnel (UTARTU) and Vanessa Demanoff (MNHN) for filling and returning 22 completed templates.
Data type: Excel Spreadsheet
Brief description: The original 22 responses from six ICEDIG collections-holding institutions (APM, LUOMUS, MNHN, NHM, RBGK, UTARTU).
Download file (213.96 kb)

Suppl. material 5: Cost books - RBINS doi

Authors: Royal Belgian Institute of Natural Sciences
Data type: Excel Spreadsheet
Brief description: Thirteen costbooks from RBINS covering technique-based digitisation costs (e.g. µCT, photogrammetry, structured light and multispectral imaging).
Download file (135.25 kb)

Endnotes

*1 Digitisation scenarios can be characterised along two axes: capability and capacity (or scale).

On the capability axis a spectrum of possibilities for the organisation of digitisation ranges from temporary or permanent inhouse facilities to fully outsourced contracts of digitisation undertaken by commercial companies. On all points of the spectrum, there can be various proportions of professional and volunteer digitisers contributing effort and affecting operating costs accordingly.

On the capacity axis, digitisation activities can range from small-scale, one-off bespoke projects to digitise specific specimens, collections or parts of a collection through to large-scale, long-term mass digitisation programmes aiming to digitise complete holdings of an institution. At multiple points on this axis digitisation-on-demand can also range from sporadic one-off digitisations (special cases) to continuous routine requests for digitisation.

*2 At the time of writing the present article there is no citation available for the proposed MIDS standard. Readers are advised to refer to the Biodiversity Information Standards (TDWG) website, https://www.tdwg.org/ and/or corresponding GitHub pages/ repositories, https://github.com/tdwg/ for the latest status.