Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:

Deliverable **D9.3** / [R] / [PU]

Work package **9** / Lead: **[CETAF]**
Delivery date [M25]
DOI: 10.5281/zenodo.3632535

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

Stakeholder Roundtables

DELIVERABLE D9.3

Karsten Gödderz (CETAF)
Ana Casino (CETAF)
Emily van Egmond (Naturalis)
Luc Willemse (Naturalis)
Xavier Vermeersch (CETAF)
Agnes Wijers (Picturae)
Jeroen Bloothoofd (Picturae)
Quentin Groom (MBG)
Helen Hardy (NHM)
Laurence Livermore (NHM)
Naomi, Cocks (NHM)
Myriam van Walsum (Picturae)
Sharif Islam (Naturalis)
Hilary Goodson (NHM)
Tina Loo (Naturalis)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

1 Table of Contents

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Executive Summary

Under the purview of Task 9.3, seven Roundtables were organised to gather focused expertise from outside the Project consortium. The events were organised and planned by the respective task leaders for which the expertise was needed in collaboration with the leader of T9.3, CETAF (Consortium of European Taxonomic Facilities). Globally, the Roundtables aimed to set operative discussions with experts that could contribute to enhance, improve, complement and validate the content produced by the work packages involved. As an alternative mechanism to common advisory boards, they have proven really effective in tackling specific issues that either could not be entirely covered by the Project partners and thus needed to be complemented with external guidance or required concrete contributions to reach the desirable quality levels.

The Roundtables were:

1) Collection Digitisation Dashboard: The discussions centred around the construction of a dashboard showing collection level information to provide an overall and comprehensive view of the status of collections in terms of digitisation and identify priorities to push forward the level and scope of virtually accessible objects. Two groups discussed different aspects of the dashboard with one focused on the end users, parameters and criteria and the other focused on the technical aspects and unifying processes for data. A set of indications regarding both, the data and the user sides as well as a matrix indicating the usefulness of dashboards/visual tools based on the user's categories – and thus their needs – cross-linked with the level of the data considered (from detailed specimen to broad collection) were produced.
2) Analogue 2 Digital: This Roundtable focused on one of the most time-consuming steps within the digitisation process, i.e. the extraction of label information, the different methods available to do so and the subsequent linkage of the extracted data. In the discussions, the different approaches were compared regarding their relative benefits with special attention given to the importance of ethics when working with volunteers.
3) Future of Warehousing and the use of Robotics: This Roundtable supported the reference task in investigating how robotics could decrease digitisation processing time and costs, to improve the retrieval and handling of specimens. It considered opportunities and practical requirements for automated warehousing and the application of robotics in collections storage and handling as well as current uses, challenges and risks in both industry and natural history institutions when tackling those issues. To date, physical robotics and automated warehousing are rapidly evolving technologies with emerging but not yet well-developed uses for collections. Therefore, costs still mostly outweigh benefits, but this is likely to change over time.
4) Robotics and 3D Scanning: Developments in robotics and 3D scanning are booming. Question is, why are they not yet widely adopted by natural history collections? That is one of the questions this Roundtable tried to investigate. Since it was difficult for organisers to attract participants, they implemented a different approach and visited several stakeholders. While the investigated and described techniques are not yet always directly applicable to speed up the imaging of natural

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

history collections, the potentials for those techniques to become common practice as part of the digitisation of natural history collections projects are clear.

5) Partnership Frameworks for Distributed Research Infrastructures: By gathering representatives from multiple RIs, this Roundtable aimed to share experiences, learn from more mature initiatives and identify possible best practices to follow. When it comes to partnerships, challenges arise from the distributed nature of a Research Infrastructure like DiSSCo. They will have to be forged along three main lines: within DiSSCo with the national nodes, with other players – e.g. other RIs and their technical and strategical interfaces, global partners (e.g. via GBIF, iDigBio) that either serve or incorporate DiSSCo's mission – and with the underlying foundational providers such as the European Open Science Cloud (EOSC) for supporting e-infrastructures. The question there is how to organise this based on the RI's ability to procure or rely on services that allow the RI to operate. One advantage of DiSSCo is its fundamental partnership with CETAF, rooting the RI in its own community.
6) Museums specimen and molecular data linkage: The main goal of the Roundtable was to explore possible use cases based on linking data in molecular databases to museum specimens. It is hoped that this effort will result in new data becoming available to an enlarged user base thereby increasing the impact of publicly funded research and services.
7) Cultural Heritage Synergies: This roundtable intended to first establish the digital needs and requirements of humanities researchers by means of reviewing survey results, listening to the user case of a digital humanities researcher, and subsequent discussion between participants. Presentations introduced organisations describing the content, function or services most relevant to humanities researcher needs. The survey demonstrated the need for integrated data resources in the population of researchers working at the science-humanities interface, but more fundamental questions regarding the size, significance and prioritization of this demand remained unanswered and requires further exploration before planning can occur.

Keywords: Digitisation, ICEDIG, 3D, Robotics, Warehousing, Molecular Data, Cultural Heritage, Partnerships, Distributed Research Infrastructures, Text recognition, OCR, natural history collections, labels, dashboards

1 Introduction

Task 9.3 (T9.3) of ICEDIG was dedicated to deploying networking actions, an essential tool in connecting to linked initiatives, contacting potential users or providers of services, strengthening the commitment among all related agents and to meticulously define the type of influence or relationship these partners might have towards the design of DiSSCo as European digitisation research infrastructure (RI).

To this end and within the scope of this task, seven target-oriented Roundtables were held with selected audiences, key technology holders as well as providers and users of the data for the Research Infrastructure (RI) on the one hand and relevant ICEDIG work package partners on the other to determine details of the design. They were organised under the purview of specific tasks to gather much-needed expert insight and commentary on the design and potential use of the DiSSCo RI. The events were organised and planned by the respective task leader for which the expertise was needed in collaboration with the leader of T9.3, CETAF.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

This deliverable D9.3 gathers the reports of the seven Roundtables to provide a reference document for their results. However, since the Roundtables results were already analysed to directly inform the tasks under which they were held and their deliverables, this report has the purpose of merely documenting the roundtables, leaving the enhancements and improvements achieved as an element contributing to the final deliverables of the respective tasks.

2 The Roundtable Tool

During the proposal process, the tool of a *Roundtable* was specifically chosen for the ICEDIG project to gather specific and timely expertise that is not already present in the Project consortium without the need to subcontract or to maintain a steady advisory board. The latter would have been challenging anyhow as the diverse expertise necessary to finetune the results of the different tasks within ICEDIG would have needed a very diverse board. Fruitful discussions would then have been difficult to facilitate as most people would come from very different backgrounds. Therefore, having a flexible group of experts available to extract specific knowledge was seen the superior option.

This decision has proven to be the correct one, overcoming the barrier of access to experts in the process while enabling tackling topics of specific relevance for each set of actions and gathering targeted experts in the domain. As can be taken from the reports below, the Roundtables were productive and brought knowledge into the project that otherwise might have been more time-consuming and complex to acquire. From this experience the Roundtable tool stands as a very efficient and productive one, given that the discussions are focused, well prepared and guided in their remit and that the required experts participate.

In some instances, gathering prominent expertise has been very difficult, e.g. from the private sector, with which the ICEDIG-related community, i.e. the natural sciences research institutions have difficulties to establish sustained contact. In this case, our partner Picturae showed great initiative to overcome this obstacle and tried a different approach. Instead of trying to collect knowledge collectively from different private agents at once (around a physical table), it was decided to do the opposite, i.e. accessing the source directly at their own premises and getting answers to focused questions as to enable gathering relevant and to-the-point information from a variety of relevant sources.

3 Roundtable Reports

3.1 Roundtable One - Collection Digitisation Dashboards

3.1.1 Summary

The Roundtable is framed under Task 2.3 and was held on the 11th of June 2018 during the first ICEDIG All-Hands meeting in Leiden, the Netherlands. Twenty-one people attended, consisting of a mix of ICEDIG participants and external experts. A general introduction was given by Luc Willemse (Naturalis) on the scope of the Collection Digitisation Dashboard, which is to be designed within ICEDIG Task 2.3. The focus is initially on a dashboard showing collection level information to identify which collections has been digitised already and which collections still need to be digitised. Elspeth Haston (Royal Botanic Garden Edinburgh, RBGE)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

then explained what is happening regarding internal dashboarding at RBGE. Wouter Addink (Naturalis) explained how different dashboards will come together within DiSSCo. Finally, Simon Chagnoux (MNHN) spoke about dashboard metrics related to citizen science projects.

After the general introduction, there was a break-out in two groups: the first group focused on the end users, parameters and criteria and the second group focused on the technical aspects and unifying data.

In the first group, end users and their user stories were identified and listed. These were supplemented with what data elements (parameters) would be necessary to be displayed in a dashboard for each user story. A next step is to further identify which data elements are associated with each user story and whether user stories can be grouped based on the data elements. Together this will provide the basis for different kind of visualisations, including a dashboard, as indicated by the conceptual model on collection digitisation visualisations.

In the second group, some technical aspects of the dashboard and how to bring together the data were discussed. Discussions were started from the data side, instead of the user side. The main conclusion is that it is essential to have a standard for the description of the collection, as to date this only exists for specimen level data. This is a requirement, so all data can be unified and presented in the dashboard in a harmonized manner. Also, collection level data is already gathered in several ways, including the annual reports of institutes, so it would be good to combine these efforts to feed into the dashboard.

When regrouping again after the break-out, the chairs of each of the subgroups gave a summary of the outcomes. In the general discussion, it became clear that there are several initiatives that are related to collection description standards (e.g. the group of TDWG tackling description standardization as Natural Collections Description - NCD) and collection digitisation dashboards (e.g. a task group on CDD recently started by Naturalis;) It will be good to keep in contact and have an open communication to make sure we combine efforts and no duplication takes place.

3.1.2 Introduction

Within the ICEDIG project, we will prepare a preliminary design of a Collection Digitisation Dashboard (CDD), with the main purpose to make digitised and not (yet) digitised natural history collections visible and discoverable across Europe. The CDD is intended to be an online tool that allows one to quickly obtain reliable and complete information on which taxa and/or geographic regions have already been digitised within a natural history collection held by an institute, and to what degree. Bringing together the data to feed within the CDD is paramount to this aim, as is determining which parameters are most useful to present within the CDD. Expert opinions on these topics regarding the design of a CDD were obtained by having a Roundtable.

On the 11th of June 2018, a Roundtable was organised on the topic ‘Design of a Collection Digitisation Dashboard for European natural history collections’ during the first ICEDIG All-hands meeting held in Leiden, the Netherlands. For this purpose, twenty people attended, consisting of a mix of ICEDIG participants and invited external experts (Appendix 4.1).

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.1.2.1 Aims

The main aim was to prepare a preliminary design for the Collection Digitisation Dashboard (CDD) with the main purpose to make digitised and not (yet) digitised natural history collections visible and discoverable across Europe.

For this purpose, we aimed to discuss the following elements during the Roundtable:

- Discuss the possibility of combining current and future digitisation dashboards, including the CDD, within a larger European data research infrastructure.
- Identify potential audiences and end-user requirements of the CDD.
- Define the parameters on which to apply the CDD, including how to calculate the values to be shown in the CDD based on the underlying data.
- Identify what data is needed and in what format this needs to be delivered to create the CDD.
- Identify how to bring this distributed data from European institutes together and keep it up to date.

3.1.2.2 Roundtable Setup

First, a lunch was organised to informally introduce the participants to each other. The Roundtable was officially started with a general introduction of the topic and the framework followed by several presentations given by experts on the different dashboards used in different organisations, the usefulness of this tool and the existing mechanisms used in different related fields (e.g. citizen science initiatives and collections management systems).

After this general introduction, two subgroups were formed: the first subgroup focused on the end users, user stories and parameters, while the second subgroup focused on the technical aspects and unifying data. The individual group discussions was followed by a presentation made by the chair of each of those subgroups and a final wrap-up of the conclusions reached by each one. Final comments and next steps were presented by the Roundtable convener.

Agenda of the Roundtable on the 11th of June 2018, Leiden, the Netherlands

12.00 - 13.00 h	Lunch and introductions
13.00 - 14.00 h	General presentation - setting the scene
13.00 - 13.20 h	Introduction (Luc Willemse, Naturalis)
13.20 - 13.35 h	Towards a collection dashboard for RBGE and CETAF (Elspeth Haston, RBGE)
13.35 - 13.50 h	Dashboards within DiSSCo (Wouter Addink, Naturalis)
13.50 - 13.55 h	Related dashboard metrics: Citizen science (Simon Chagnoux, MNHN)
13.55 – 14.00 h	Summary and splitting into subgroups (Luc Willemse, Naturalis)
14.00 - 16.00 h	Work on specific questions in two subgroups:

1. End users and parameters (front of dashboard)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

2. Technical aspects and unifying data (back of dashboard)

16.00 - 16.30 h	Presentation of subgroup outcomes
16.30 - 17.30 h	General discussion
17.30 - 18.00 h	Identifying next steps and closing

3.1.3 Main Results

3.1.3.1 Subgroup One – End Users and Parameters

In the first subgroup, as a first attempt, a list of end user groups was drawn up. The following main (potential) user categories for the CDD were identified (in random order): Research, Collection, IT, Governmental, Non-Governmental, Education, Industry, Media, Institution and Citizen science. For each user group, the participants together indicated which information (collection, storage unit, species or specimen) would be relevant for each user group (Appendix 4.1). This shows that collection and specimen level information are considered to be useful to many of the user groups, while storage unit and species level information is of more value to the institute and collection managers.

Subsequently, user stories were captured at a larger, overarching level, following the format 'as a' [*user*] 'I want to' [*do this; know this*] 'so that I [*can do this*]'. For example: "as a collection manager, I want to see all digitised European collections of bees, so I can prioritise the digitisation of bee collections. In total, 22 user stories were collected related to the CDD (Appendix 4.1). Once a complete set of such stories is available information can be added like 'for this I need (data elements)' digitised and non-digitised information on bees at family, genus and species level, which was already started by this subgroup for the collected user stories. In addition, it was indicated by the participants during the Roundtable that an overview of natural history collections at the highest data level would be (to varying degrees) useful to all user groups.

3.1.3.2 Subgroup Two – Technical Aspects and Unifying Data

In the second subgroup, a discussion was held to identify the main technical challenges to be addressed when preparing the CDD. This meeting was structured based on domain events related to digitisation and data. On pink sticky notes, domain events were described related to the dashboard in particular. There is an input and output, and something happens in between with data (domain event), which were written down on yellow sticky notes (triggers). All sticky notes were placed on the wall and different elements were discussed in relation to the CDD (see Appendix 4.1 for details).

The main conclusion is that collection data standards are needed to feed data homogeneously into the CDD. Currently, collection-level information is, if at all, determined by each natural history institution in a different, heterogeneous way. So far, there is no data standard available for collection-level data and a data quality standard for this type of meta-data is also lacking. The natural history institutions that are a partner within DiSSCo will need to deliver collection-level data, following a newly developed collection data standard, each year, which can be visualised in the CDD. Ideally, this information is available in an online environment to be able to feed data automatically to the CDD and keep it up to date.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.1.4 Concluding Remarks

3.1.4.1 Challenges Faced

- Short time to organise the Roundtable and thus, to engage a diverse audience. Experts from the field of statistics or from the field of communications, education or media (visual materials producers), for example, would have been of great help.
- Though the splitting of the Roundtable into two groups helped to tackle different dimensions of the Dashboard as envisaged for ICEDIG, the topic complexity would have needed more time for discussion on its different aspects and additionally an extended time for merging the results coming up from the two different groups.
- To harmonise the different backgrounds of the participants and to centre the overarching discussion on the final dashboard pursued, it would have been useful to provide the participants with supporting documentation (tools available, examples already running).

3.1.4.2 Outcomes Produced

- A set of indications obtained from the two discussion groups regarding both, the data and the user sides.
- A matrix indicating the usefulness of dashboards/visual tools based on the user's categories (and thus, their needs) cross-linked with the level of the data considered (from detailed specimen to broad collection).
- The need for collection data standards to feed data homogeneously into the CDD.
- An indicative list of available (visual) tools.

3.1.4.3 Actions Proposed

1. To identify the level of the content to be showed in the dashboard in relation to the audiences identified.
2. To agree upon the metrics and the dashboard architecture (flows, update automatization, etc.).
3. To acknowledge existing efforts and collaborate with them (a.o. CETAF Digitisation group, COST Action Mobilise, SYNTHESYS+ project).
4. To integrate results produced by the Taskgroup CDD.
5. To take into consideration the parameters and criteria used by the NCD of TDWG.
6. To analyse the different visual tools available in the market and consider its feasibility and fitting-for-purpose.
7. To compile more user stories, translate it into user requirements and identify the data to gather and show in the final dashboard.

3.2 Roundtable Two – Analogue 2 Digital: Faster Better Cheaper

3.2.1 Introduction

One of the most difficult and time-consuming steps in the digitisation workflow of museums and herbaria is the extraction of information from labels. Methods to extract these data can be through human transcription or through some form of automated transcription, such as optical character recognition (OCR). These different approaches can be broken-down further, for example, manual transcription may be conducted either by specialist companies;

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

volunteers or in-house technicians. All these possible approaches can have different outcomes in terms of speed, quality and cost.

Another indirect method of annotating specimens is by connecting the data on the specimens to other sources of information. For example, if data on a specimen can be linked to other specimens collected by the same collector on the same date then one can assume that they were collected at the same or an adjacent site. Another example might be connecting the name of collectors to a biographical database. Such a link can help validate dates by cross validation.

The roundtable was part of the Joint Annual Meeting of the Society for Preservation of Natural History Collections and the Biodiversity Information Standards (TDWG) organization. Attendees to the conference come from museums, herbaria, universities and other organizations concerned with biodiversity research, data management and monitoring. The abstract of the roundtable is available on the conference website (tdwg.github.io/conferences/2018/sessions/W03).

The roundtable was attended by approximately 70 people and in addition to ICEDIG partners there were representatives of many organizations involved in collections (e.g. Sydney Herbarium and Botanic Garden Botanical Museum, Berlin); citizen science (e.g. Wikidata, Digivol); communications (Atlas of Living Australia) and digization (e.g. iDigbio).

3.2.2 Presentations

There were four formal presentations at the meeting that are detailed below together with links to their abstracts in the Biodiversity Information Science and Standards journal.

State of Digitisation and Gap Analysis Surveys - Sarah Phillips

> Phillips S, Haston E, Green L, Weech M, Cubey R, King S, Drinkwater R (2018) State of Digitisation and Gap Analysis Surveys. *Biodiversity Information Science and Standards* 2: e25969. https://doi.org/10.3897/biss.2.25969

An Evaluation of In-house versus Out-sourced Data Capture at the Meise Botanic Garden (BR) - Henry Engledow

> Engledow H, De Smedt S, Bogaerts A, Groom Q (2018) An Evaluation of *In-house* versus *Out-sourced* Data Capture at the Meise Botanic Garden (BR). *Biodiversity Information Science and Standards* 2: e26514. https://doi.org/10.3897/biss.2.26514

Service-based information extraction from herbarium specimens – Fabian Reimeier

> Reimeier F, Röpert D, Güntsch A, Kirchhoff A, Berendsohn W (2018) Service-based information extraction from herbarium specimens. *Biodiversity Information Science and Standards* 2: e25415. https://doi.org/10.3897/biss.2.25415

Crowdsourcing, is it a good option for your collection digitisation? – Quentin Groom

> Groom Q, Bogaerts A, De Smedt S, Phillips S (2018) Crowdsourcing, is it a good option for your collection digitisation? *Biodiversity Information Science and Standards* 2: e25410. https://doi.org/10.3897/biss.2.25410

The first presentation detailed the current status of digitisation among members of the Consortium of European Taxonomic Facilities. The full presentation has been archived online (Phillips et al., 2018). The presentation details a survey conducted for the Synthesys project. The survey showed that institutions in Europe vary considerably in their progress towards full digitisation and that much effort is needed in training, standards development and automation. One of the Tweets about the Roundtable referred to this talk and said "*Few standards are used for herbarium digitisation, everyone uses different equipment says Sarah Phillips from @KewScience @#SPNHCTDWGNZ"*. It refers to the diversity of digitisation approaches used by different institutions.

The second presentation examined the experience of Meise Botanic Garden who used a specialist company for the transcription of herbarium specimen labels. The presentation compared this approach with transcription performed in-house by expert technicians. The conclusion was that the two approaches resulted in similar quality data, but there were differences related to the protocols used by the different transcribers. Out-sourcing was recommended but it was also recommended that a clear quality control system must be in place (Engledow et al. 2018).

The third presentation described a workflow for the semi-automated extraction of label information from herbarium specimens. This workflow combines OCR and data services with the data manipulation tool OpenRefine (http://openrefine.org/). Using this combination of tools users can rapidly and reliably digitise label information and link information on the label to external resources. Such workflows could be built into a future digitisation pipeline or run on archives of images in order to enrich their metadata (Reimeier et al. 2018).

The final talk compared all the different options for data extraction and compared them with each other (Groom et al. 2018). It also looked into the future and discussed possible ways to improve the current workflows. Table 1 was presented at the meeting and describes some of the differences between methods that need to be considered. This slide induced some discussions about the relative benefits of the different approaches and of the importance of ethics when working with volunteers.

Table 1. Adapted from the presentation by Groom et al. (2018). A simple summary of some of the considerations when choosing a data extraction method. The column interpretation is because some methods extract verbatim text transcriptions, while others interpret the meaning of the text.

Method	Interpretation	Speed	Engagement	Costs

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Volunteer Crowdsourcing	✓	slow	✓✓	High setup costs, decreasing with use
In house transcription	✗	fast	✗	static
Scientists	✓✓	Fast, but not systematic	✓	Low, as a side product of research
Outsourcing	✗✗	Very fast	✗	Single high payment
Automation	✗✗ ?	Potentially very fast	✗	High setup costs, decreasing with volume

3.2.3 Feedback

Sarah Philips from Kew Gardens wrote the following regarding the meeting.

"*The analogue 2 Digital: faster better cheaper symposium was extremely useful to me. Insights from Meise's work comparing in-House transcriptions to external outsourced transcriptions could directly feed into the ICEDIG report for task 4.2. It was also useful to see common issues coming up including the difficulties of making a direct comparison in quality when transcribers are following different protocols and hear thoughts on how this can be dealt with. It was also interesting to hear about the work done by Fabian Reimeier at the Botanic Garden and Botanical Museum, Berlin, within the StanDAp-Herb Project and the tools/webservices available that we could possibly test here at Kew. This is something I definitely look to follow up on. I also discussed with Fabian about the possibility of detecting certain labels on our herbarium specimens on the basis of colour. Kew has used Orange stickers to indicate when a specimen has been sampled for DNA so I wondered if these could be picked up on the specimen images.*"

3.2.4 Conclusions

In conclusion, the meeting was a useful communication and discussion event that was well attended and reached an international audience. The results will be feed into ICEDIG deliverables, particularly in work package 4. We believe the meeting has raised the profile of ICEDIG with the attendees and will help in the recruitment of experts in future work of ICEDIG and other DiSSCo related projects.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.2.5 References

Engledow, Henry, De Smedt, Sofie, Bogaerts, Ann, & Groom, Quentin. (2018, October). An Evaluation of In-house versus Out-sourced Data Capture at the Meise Botanic Garden (BR). Zenodo. http://doi.org/10.5281/zenodo.1445426

Groom Q, Bogaerts A, De Smedt S, Phillips S (2018) Crowdsourcing, is it a good option for your collection digitisation? Biodiversity Information Science and Standards 2: e25410. https://doi.org/10.3897/biss.2.25410

Phillips, Sarah, Haston, Elspeth, Green, Laura, Weech, Marie-Hélène, Cubey, Robert, King, Sally, & Drinkwater, Robyn. (2018, October). State of Digitisation and Gap Analysis Surveys. Zenodo. http://doi.org/10.5281/zenodo.1443394

Reimeier F, Röpert D, Güntsch A, Kirchhoff A, Berendsohn W (2018) Service-based information extraction from herbarium specimens. Biodiversity Information Science and Standards 2: e25415. https://doi.org/10.3897/biss.2.25415

3.3 Roundtable Three – Future of Warehousing and the use of Robotics for Natural History Collections

3.3.1 Summary

Retrieving and handling specimens by hand is the main cost in mass digitisation. Task 3.3 investigates how robotics could decrease processing time and costs, and alleviate problems with regard to health & safety, as well as supporting on-demand remote digitisation. This covers all stages of the retrieval and digitisation of objects; from preparation to the transport of the object to the digitisation space, the digitisation process itself, the transport back to the collection storage and all actions in between needed to complete the digitisation process.

This meeting considered opportunities and practical requirements for automated warehousing and robotics in collections storage and handling. We heard from industry, and from peers who have undertaken or are undertaking collections facilities projects. The main goal was to stimulate discussion around new approaches, and tour collections to give insight into current uses and challenges. We analysed the opportunities and risks/challenges that a collection may face in considering the use of robotics.

The focus is on physical robotics and automated warehousing. All references to robots/robotics should be taken to mean physical robotics.

3.3.2 Introduction

This Roundtable forms part of ICEDIG WP3, which investigates the physical handling and technical aspects of the mass digitisation of collections, building towards the DiSSCo design and vision for a pan-European collections research infrastructure. Retrieving and handling specimens by hand is the main cost in mass digitisation. This Roundtable investigated how robotics could decrease processing time and costs, as well as supporting on-demand remote digitisation.

The focus is on physical robotics and automated warehousing, not on software 'robots' or automated processes such as Optical Character Recognition – this is covered in other work packages. All references to robots/robotics should be taken to mean physical robotics.

The use of Robotics and warehousing systems covers all stages of the retrieval and digitisation of objects; from preparation to the transport of the object to the digitisation space, the digitisation process itself, the transport back to the collection storage and all actions in between needed to complete the digitisation process.

3.3.2.1 Aims

This meeting considered opportunities and practical requirements for automated warehousing and robotics in collections storage and handling. We invited developers and users who have undertaken or are undertaking collections facilities projects to stimulate discussion around new approaches, as well as touring collections to give insight into and discuss current uses and challenges. We analysed the opportunities, risks and challenges that a collection may face in considering the use of robotics.

For this purpose, we discussed the following elements during this Roundtable:

- Research and curatorial needs - benefits and risks of offsite warehousing, including possibilities for technology to improve the balance of benefits and risks.
- Which types of collections and storage may be best suited to robotic/automated handling?
- The interactions between digitisation and automated retrieval and handling, including baseline data required, and digitisation on demand.

3.3.2.2 Roundtable Setup

Natural science collections increasingly have to look at new storage solutions and facilities, to protect their collections while managing size and costs – often including consideration of offsite facilities, outside major cities.

This Roundtable brought together parties who have challenges and solutions in the field of warehousing, robotics and natural science collections. While we based discussion on current challenges and needs, we took a broad and long term look at possibilities, as well as developments that may be needed, for example, in collections storage or in the ability of robots to handle fragile objects. We considered the balance between physical and digital access to collections, looking at examples from outside the natural sciences.

In the preparation of this Roundtable, we experienced a lack of response from industry, perhaps owing to lack of familiarity with the collections-holding sector. Where interest and solutions exist, these are often focused on the public-facing aspects of collections i.e. exhibition. R&D is booming in the robotics and automated warehousing sector however, so it may well be that this progress becomes increasingly applicable to collections as well.

Presentations:

Maarten Taborsky, Bruns (NL), ROBOCASE

Link to presentation

Andy Appleyard, British Library (UK), High Density Storage

Link to presentation

Clare Valentine, NH (UK) Considerations in Planning a Major New Collections Facility

Link to presentation

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Collections Tours and presentation of collections:

Ben Price, NHM, Entomology

Link to presentation

Mark Carine, NHM, Herbarium

Jacqueline Mackenzie-Dodds, NHM, Molecular Collections facility

Giles Miller, Principal Curator, Micropalaeontology and SCIC EE

3.3.3 Presentations and Tours

3.3.3.1 ROBOCASE

By Maarten Taborsky, Bruns, Project Director

Bruns, Bergeijk (NL), is specialized in the development, production and installation of innovative mechanical, electromechanical interactive exhibits and complete (interactive) exhibitions for Science Centres, museums and information and visitor centres, including maintenance. works with 90+ professionals, from storytellers to technical developers, craftsman and marketeers, on a fulltime basis. Their ambitious goal is to offer museum and Science Centre visitors a worthwhile experience and maximum educational value. After developing several interactive exhibitions, the company recently developed (partnering with Kisssthefrog) the 'Robocase', a museum display case with a robot arm.

Collection and stories

The museum display is interactive with the public, and brings some of the best stories from collections, that can often be hidden behind the scenes (particularly in the case of smaller items), to the fore. With Robocase, Bruns researched the potential for robotics to personalize a display that can bring an object closer, based on selection by the visitor.

Before developing the Robocase, many questions had to be answered: Collections are vulnerable in regards to lighting; can this be measured and steered with an interactive solution? What is the best way to present the different stories from different collections: historically, thematically, or connected with other collection items? How to attract and seduce people to look more thoroughly at objects; encouraging greater engagement and learning through a more dynamic and personal display.

Answers were used to develop the idea of a robot display case. Curiosity in many ways is indispensable to create the most suitable solutions: storytelling, techniques or materialisations, and combinations of different disciplines. Result: a landmark showcase, a robot that moves objects of the public's interest and even surprisingly opens a drawer.

https://www.robocase.nl/

The process of problem-solving and development in order to realise this vision is very relevant to potential uses of robotics for collections, considering key issues such as safety (for humans and objects being manipulated; longevity of robotic solutions; and functionality. Bruns set up a mock-up and tested the UR 10 robotic arm, which has a combination of robustness, humanness and stability. A ridged construction was combined with a slider, to be able to go across 2.5 meters height and 3 meters width.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Programming and safety issues were the next step. Safety measures are implemented on three levels: (1.) The robot arm itself, (2.) extra sensors in the display case that can be recognised by the (3.) software.

In many instances, there are multiple possible approaches. The grippers were one of the challenges; none were able to deal with objects of different sizes or to detect the exact firmness of the grip. Therefore, an alternative solution was developed, placing the objects on a plate, which the gripper can grasp and manipulate smoothly. Opening the drawers was another challenge, which was solved by making the position of the gripper flexible.

The CMS software that has been used makes it possible to replace the objects as desired, therefore the display can be used for decades.

All of the challenges for this Robocase in terms of safety, movement, size, weight and lighting are also relevant to collections behind the scenes. The robotic arm has been shown to move smoothly enough to transport a house of cards, and objects are moved on a plateau, so the robot does not touch them - is that an answer for moving collection items? The Robocase suits smaller objects up to 29kg. Could this work with 3D digital surrogates? Could this be a display-model for the 5 biggest Natural History Museums, and even communicate between the different collections?

3.3.3.2 High Density Storage

By Andy Appleyard, British Library, Head of Operations North

Andy Appleyard and his colleague Alison Selina shared their findings on the complete process of developing automated, high-density, off-site storage facilities for the collections of the British Library, and the daily practise of the use of these buildings, including the 24-hour retrieval service to users in London.

The British library has storage space in London and in Boston Spa, with regular vehicle transport between the two premises.

The British Library site at Boston Spa has two high density storage buildings. The first for books and the second primarily for newspapers, with a third also planned. These buildings also contain back-office functions including conservation and a reading room.

Development of these facilities was a major investment; however, the key driver was the difference in cost for 1m shelf space in London vs high density store at Boston Spa (10x times more expensive in London standard storage).

Day and night shifts are needed to cover demand for retrieval, approximately 1,000 items per day. Night technical support is too expensive, but processes less likely to cause issues are now done at night, and operators can do simple restarts themselves.

The programme required 5 project streams, (1) Construction and automation, (2) System integration, (3) Collection preparation and moves, (4) Delivery service (for retrieval and returning items) and (5) London Buildings disposal. It has been followed by extensive continuous improvement and troubleshooting work. The second building drew on lessons learned and experienced fewer bugs/issues.

The first building is 10 years old next year, having started development in November 2002 and delivered a steady state operation in 2011. The building contains a loading dock on the first floor,

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

working space on the second floor and the rest is low oxygen storage space. Volumes are held in plastic boxes, with a barcode on each item that links to the catalogue. The space is temperature controlled (16 degrees +/- 1%) and humidity controlled (52 degrees +/- 5%).

A key difference to natural sciences collections was that almost all material was already digitally catalogued at item level, so they just needed barcodes to be added to link items; boxes; catalogue; warehouse retrieval software. This was still a major preparatory stage. Serials were the only exception to item level cataloguing and are filed in runs so they can be found, albeit with a bit more work.

The system is self-learning and moves higher use content to the front. The software is externally provided and maintained.

Delivery and return can be done in 24 hours. This demand profile was one of the important values, the system needs to match the speed of retrieval possible as capacity is fixed. As with natural collections, what is high-use/popular can be subject to trends and changes.

Major lessons learned were:

- To have a test environment before construction and moves
- Not to take any shortcuts with item identification and barcoding
- To use chain rather than belt conveyors
- To avoid large plastic containers that warp over time
- To consider the human needs for the site e.g. natural light in rest areas.

For the newspapers there is a digitisation partnership with FindMyPast who sell subscriptions to genealogy users. The material is available to the British Library but only for use behind the scenes and in reading rooms. At the end of the contract they expect to make it fully public. Material digitised for preservation is also outsourced but made available on the internet for free.

3.3.3.3 Considerations in Planning a Major New Collections Facility

By Clare Valentine, NHM Collections Leader, Life Sciences Collections

Clare Valentine is one of the senior leaders of the NHM's Collections Programme, examining options for future collections storage including major offsite facilities.

This is driven by collections conditions and volume, but also raises a much wider range of questions including digital & physical access; usage including research, exhibitions and loans; research 'journeys' across collections e.g. parasite-host; predator-prey etc; and adjacency of relevant kit e.g. for molecular and chemical analysis, imaging etc.

Data is key and the programme is drawing on the Museum's 'Join the Dots' collections assessment data, building on a methodology used by the Smithsonian and capturing use, importance, accessibility and condition of collections. Many collections are relatively highly used, but with relatively low predictability of use (and where there is low use the reasons for this are not always clear) – this has contributed to a change from initial assumptions of 'deep storage' to consideration of a much more active site that may include relevant research teams and/or facilities.

Now NHM is considering how to maintain links in future to the 'home' site - Naturalis decided not to move any whole sub-collections off site for this reason.

The benefits required for the collections are:

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- Improved conditions for collections care
- Improved physical access to collections
- Improved digital access to collections
- New ways of curating and working on the collections
- Increased efficiency of the Museum estate through space rationalisation
- More use of collections in public outreach
- Benefits to wider world (Conservation, Environmental change, Sustainability)
- Improved opportunities for working with collaborators

3.3.3.4 Collections Tours and Presentations

By Ben Price, NHM, Senior Curator in Charge for Entomology, Insects

Mark Carine, NHM, Principle Curator in Charge Algae, Fungi and Plants

Jacqueline Mackenzie-Dodds, NHM, Molecular Collections Facility Manager

Giles Miller, Principal Curator, Micropalaeontology and SCIC EE

Attendees were taken on tours of herbarium collections including the special collections room housing Sir Hans Sloane's bound herbarium volumes; entomology collections including Hymenoptera and wet collections; and molecular collections storage including very low temperature storage and robotic analysis of vials. There were also short presentations from the curators at the meeting.

These tours and talks illustrated the topics discussed during the rest of the Roundtable. During these, questions and observations focussed on the following topics:

How to set priorities?

- Collections and their use continue to change and evolve significantly – for example the NHM entomology collection may acquire more than 100,000 new items in some years through fieldwork; legacies and other acquisition; as well as sending high volumes of material on loan, and hosting hundreds of scientific visitors. These are all challenges for curatorial and digitisation prioritisation.
- Topicality can be an important factor in use of collections e.g. cyanobacteria were low use for many years but now increasingly used because of relevance to climate and environmental change. Not all collections are equally appealing to an audience, but opportunities sometimes arise to raise the profile of collections e.g. through digitisation on demand as part of the SYNTHESYS+ Programme.
- NHM (as all collections) often have to seek external funding for digitisation of collections. Funding should not be the immediate cause of prioritisation; a balanced portfolio is needed of mass digitisation and deeper/more bespoke projects such as capturing all the data on type specimens or 3D digitising particularly significant items.

What has worked well so far?

- At the NHM, they strive to make the collection more accessible, digitisation is an important part of that. Botany has been experiencing some of the transformational benefits of digitisation for longer than some other areas –the JSTOR global plants initiative has made a big difference to botanic research and collections management.

  Engaging communities with collections by getting them involved with digitisation can benefit their understanding. For example, crowdsourcing of label transcription, or sharing stories about collectors on social media – but this needs careful measurement of costs and benefits (e.g. data quality of crowdsourcing means this is not cost-effective for NHM currently).

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- Robotics are already deployed in high-throughput sample manipulation within some molecular collections, with human interaction e.g. to confirm that a stage has been successful and to move to the next stage.

Challenges

- Collections differ in size, fragility, the way they are packed/ stored, identified or not. There is no unambiguous solution for mass-digitisation of the complete collection.
- All collection items need to have a unique identifier associated with digital records e.g. collections management system – this is a huge challenge in itself, and a precondition for use of robots or warehousing actions.
- To find funding for digitisation is one of the biggest challenges. Evidence of the impact of digital collections is growing but there are substantial time lags and a lack of predictability in outcomes e.g. in research use, which affect funders' ability to understand the importance of these projects in relation to more tangible initiatives such as public displays.

3.3.4 Topics Discussed at the Roundtable

The presentations and tours were followed by further roundtable discussion covered a wide range of topics, focusing on how natural science collections can work together and support one another across very similar challenges. Some of the key areas of discussion are outlined below.

3.3.4.1 General Findings

Use of collections

To understand what we are aiming for in digitisation and the level of digitisation we want to reach, we have to develop knowledge about the use of collections. We need to think about whether we can have the collection in house or, work with digitised and imaged materials. Looking at what part of the collection is used by hand, it is in fact quite small. Browsing the collection in person is still important to visiting researchers.

Digitisation will help but is certainly not the answer to everything. Browsing the collections plays an important role. Visiting researchers want to have a look at particular collections. This can invoke interest in other related areas. To be able to fully engage with a collection, based only on what has been digitised, a certain proportion of that collection needs to have been digitised. Only then can a researcher truly see the full potential and advantages of digitisation.

Costs

Cost will often outweigh benefits of automated processes. But technical developments are fast. By looking exactly at what is required and presenting this to the market, these developments can be made more specific for NH collections.

Sharing experiences; possible solutions

ICEDIG and related initiatives will shed some light over how collections are used and could be used more powerfully in the future. We need evidence of what kind of digitisation replaces or increases the use of physical objects. Is it likely to be specific to the type of collection and type of digital surrogate? For example, baseline 'catalogue' data vs 2D images; full 3D model or other.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The connections across types of material vary and are unpredictable, like parasites/hosts; predators/prey; crop and pest. The ability for a workflow 'storage - trolley - reading room or lab' is more complex for different types of material. Digital potential could be high.

Digitisation of material will change the behaviour of researchers. For example, with plants, JSTOR Global Plants has changed how taxonomists approach finding data. But the critical mass of data is an issue, where the majority of information is still missing and where quality is variable and not always transparent. Feeding the databank with reliable information that comes back to your collection is one of the things that can have a major impact on improving the quality of the data. Research data joining up to specimen data has the potential to be very powerful.

3.3.4.2 Warehousing

Many (parts of) collection items have no unique identifier yet, automated retrieval and handling cannot be executed without this data.

Although collections contain high monetary value (e.g. diamonds), high research value, high exhibition value, there are still specimens that are rarely or never used. Size variation can also mean that high percentages of space are taken up by very few items, for example whale skulls. Developing warehouse-style storage involves considerations about potential sites, to minimise risks such as flooding, earthquakes and to manage security. It also involves consideration of which collections should move – those that are less used; the easiest to retrieve or the ones which maximise the opportunity to use data rather than retrieving the physical object.

Attendees did not believe that an approach like the British Library's robotic retrieval would work for most natural history collections, because of the variety of objects and the unpredictability of retrieval, which would often require multiple specimens of different types. However, there are areas such as molecular collections or other collections using vials where robotics is becoming increasingly relevant and where uses are likely to develop. The key point will always be whether the benefits justify the costs.

NHM has moved away from the idea of deep storage. It may remain relevant to particular kinds of collections or environments including molecular; collections with fire or similar risks to humans. One needs to prioritise however as with digitisation that prioritisation is a matrix not a list. Naturalis decided not to put whole sub-collections offsite because of the potential to disconnect from it - 'out of sight out of mind'

3.3.4.3 Robotics

To answer the question which types of collections and storage may be best suited to robotic/ automated handling, we need a better understanding of what robots can do or what automated handling is being suggested. There is no unambiguous solution for mass digitisation of the complete collection.

There are robots that can do almost anything, from very delicate work e.g. medical to large weight e.g. industrial and smooth motion. Robots can deal with extremely delicate situations e.g. surgery - but requires the interface with a human. In the case of collections, it seems clear that the main potential is to handle storage vessels or supports e.g. a plate/platform, rather than collection items directly. We need to establish what the business case is for museums to invest.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Software robots are perhaps more immediately relevant than hardware. Machine learning approaches to extracting data are being explored now.

It is not difficult to imagine storage for many collection types that could be handled robotically, but it is not clear if this is always beneficial. For example, the uses are not like books. Requests would often be for mass material; material of varying types/sizes at the same time; or requests that will change or be added to when the material is seen. This returns us to the question about why? What problem would robotics and automation help us fix? Consistent storage types would be key to robotic handling - e.g. vials; drawers; boxes.

Usage and display of collections are both very low in annual percentage terms, digitisation can be part of the answer, but far from a 'mass demand' – there is limited supply of researchers as well as curators. Big data research most closely resembles 'mass demand' for collections at present.

It is a pity when people don't get to see the vast volume or variety of collections, particularly smaller material that may not have as much impact in traditional displays. There is potential for robotics in display e.g. a robocase. Weekly specimen rotation could show all NHM insect drawers over 1.5 years. This could be a space-efficient display solution for a visitor centre offsite. Another example could be robotic displays around Europe (or beyond) that communicate with each other, telling more stories from collections; how they are stored, researched and digitised. Another thought on visibility is a physical connection to a remote location via robotic/ automated access. For instance, a technologically enabled display could show a virtual model of something elsewhere.

3.3.5 Conclusions and Next Steps

Physical robotics and automated warehousing are rapidly evolving technologies with emerging but not well-developed uses for collections. At present, cost will often outweigh benefits, but this depends on each proposal and is likely to change over time. Key areas of promise appear to be in display of natural science collections, and perhaps handling particular storage media. Work packages across ICEDIG and the related EU projects are producing complementary information covering:

- Common standards, protocols and definitions for digitisation.
- Increased data about the use and value of collections, including differences in the physical use of collections before and after digitisation – when does it increase demand for the physical objects vs when does it reduce/replace demand?
- Software automation for digitisation and beyond, such as automated extraction of label data.

While large-scale adoption of robotics and automated warehousing is unlikely in natural science collections at present, it is likely that pilots in particular areas will be key to understanding future potential, whether public facing; related to particular types such as molecular collections; or related to major storage initiatives. It will be important to support and share such pilots so that consortium partners are aware of how robotics may be applicable to the challenges they face.

Recommendations

- Continue to work towards common standards and definitions for digitisation
- Develop knowledge about the use and value of collections with all partners involved in DiSSCo.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- Map the differences in the physical use of collections before and after digitisation – when does it increase demand for the physical objects vs when does it reduce/replace demand?
- Continue to share successful digitisation projects and the demonstrable benefits of the digitised collections.
- Make sure that the digitisation projects will not only include the technical information, but also the human part (staff, education, workflows, communication).
- Focus on the quality as well as quantity – continue to improve data about uses of digitised collections to determine which elements of data are the most important to research/use.
- Share information on collections moves, pilots of what is possible, achievable and realistic.

In this Roundtable two aspects have remained underexposed:

1. The use of different techniques.
2. How to engage more successfully with industry.

These aspects will receive more attention in the second Roundtable.

Handling and positioning objects robotically are the most difficult challenges for which there is no real solution yet. We will visit research institutes as Fraunhofer, Rijksmuseum Amsterdam, British Library and technical universities to investigate which techniques are most promising for Natural History Collections.

What kind of (rapid) developments are going on in robotics and which robotic techniques are being used in natural history or cultural collections and can possibly be applied to other types of collections? We will mainly research 3D scanning, that makes use of the next generation of autonomous and compliant robots, as well as optical scanning techniques. (arcs equipped with cameras or cameras mounted on a lightweight, compliant robotic arm), photogrammetry, laser scanning, structured light and optical material property scanners.

In an attempt to engage more, we will pay visits to the industry instead of bringing them together at a Roundtable. Together we will look for a possible business case. We will review reasons behind why the business case isn't there. Looking at topics such as lack of familiarity with the collections-holding sector, financial issues and clarity over what the questions from collections are. R&D is booming in the robotics and automated warehousing sector, so how can we make this progress applicable to collections as well?

3.4 Roundtable Four – Robotics and 3D Scanning

Developments in robotics and 3D scanning are booming. The question is, why are they not yet widely adopted by natural history collections? And why is it so difficult to find parties that want to share their expertise with the ICEDIG consortium and develop new possible solutions to image natural history collections in a fast and relatively cheap way?

Roundtable three taught us that possible solutions in robotics and 3D are very diverse, so to compose a programme that brings it all together in one Round Table is impossible. Since natural history collection holders have not yet defined a well-developed business case, we wanted to focus on the overview of possibilities and not just on one topic. We felt the need to explore different options, compare them and advise on what offers the most.

Developments in robotics are going very fast, the developers like to focus on the applications that are immediately obvious. It is less interesting for them to participate in a program that does not look

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

for that focus. We therefore decided to turn the program around and visited the developers instead of gathering everyone centrally. At some places we already had a solid contact, at others we had to find out if they were willing to receive us.

It turned out to be an excellent choice, that has brought much information, mainly on 3D developments.

Instead of organising a Round Table we have visited 4 different parties:

British library on 3D scanning
National Museum of Wales on 3D scanning
Rijksmuseum on 2 + 3D Photography, practice and prophecies. Meeting up with different people from industry on different subjects of digitisation and robotics
Frauenhofer on 3D scanning

3.4.1 Open House CultArm3d at Getty Research Institute

Participants:

Charles Walbridge (Minneapolis Institute of Art), Emily Pugh (GRI), Pedro Santos (Fraunhofer), Tassie Gniady (Indiana University), Thomas Flynn (Sketchfab), Myriam van Walsum (Picturae), Mark Lindeman (Picturae)

Target: Artefacts from cultural heritage collections
Model: Presentation to Getty research institute
3D Digitisation Method: Photogrammetry
Type of Station: high resolution PhaseOne mirrorless cameras (iXG, 100 mpix), Canon and Nikon DSLRs, Diffuse D50 lighting (ring light or closed lighting cylinder), Universal Robot UR10, Turntable (transparent, 33 cm Ø)
Publishing: Online, downloadable data

Background

The CultArm3D-P is a colour-calibrated, autonomous scanner that captures the geometry and texture of arbitrary objects using photogrammetry. It stands apart from conventional scanners with high-resolution results and independent view planning. Intelligent algorithms use a first scan to determine which further poses make sense, allowing capture of an object with the optimal number of scans. This approach enables the system to measure objects quickly and autonomously – without having to be taught or own the CAD model

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The CultArm3D has been developed by Cultlab3D in close collaboration with industry and the cultural field. Cultlab3D is the research lab of the Competence Center Cultural Heritage Digitisation[1] at the Fraunhofer Institute. It was presented at the Getty Research Institute.

Outline

Ideally 3D capture should give consistent results in geometry and texture. Traditional handheld, and even tripod based, photogrammetry can't do this. This demonstrated system should be able to flexibly combine geometry and texture as well as optical material properties, to capture objects of various sizes in an automated process. It should generate true-to-original 3D representations with accuracy down to micrometre level. With this solution, photogrammetric capture comes with a lot of variables that are either consistent or specifically optimised for each object. It elevates photogrammetric capture from subjective (even artistic) to scientific and standardised.

Another goal is to achieve a watertight mesh, but not going as far as faking data to close holes. Fraunhofer prefers software that does no interpolation: MicMac. This is the real observed, scientific data. However, they use other photogrammetry software as well because they want the capturing hardware to be independent of photogrammetry software. The strictest requirement is that the photogrammetry software is command line accessible.

Final preparation for web viewing is done through their own software, currently a Fraunhofer spin-off company: RapidCompact by DGG, includes decimation, UV unwrapping and normal maps.

In theory the system is quite camera independent, as long as the camera can be controlled programmatically. This means it is adapted to use high resolution PhaseOne mirrorless cameras (iXG, 100 mpix), Canon and Nikon DSLRs as well as laser and structured light scanners. The benefit of the PhaseOne cameras is that the focal plane is completely reproducible which makes it possible to pre-calibrate. Through the use of the robotic arm, the exact position of the optical centre is known, which almost completely removes the need for camera position calculations. This removes a significant part of the alignment calculations so that alignment can be sped up significantly and is

[1] https://www.igd.fraunhofer.de/en/competences/technologies/3d-scanning

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

more reliable. This also means that the scan is already scaled to real world size, in contrast to uncontrolled photogrammetric capture.

The system is fitted with a glass turntable (manufacturer: PhotoRobot) so that the underside of many objects can be sufficiently captured. Two positions below the turntable have been pre-programmed to minimise refraction from the glass as well as ensure that the arm does not hit the turntable. Objects with complex undersides (or objects with no definable undersides which includes many natural history specimens) will still need to be repositioned to properly capture it from all sides.

The capturing system is not compatible with Mac. The computer specifications for capture are very low: it can run on a lightweight laptop. However, the requirements for the processing stage are a bottleneck. A typical capture results in [600?] images.

The safety of the scanned object and of humans in the vicinity needs to be assured to make this robotic system viable. To do this, multiple layers of security measures are present.

Process

Calibration is only necessary when the whole system is moved, or a change in camera, lens or turntable is made and is designed to be simple. After calibration, a human operator is only required to load and unload specimens, and to start the capturing process.

The capture is done in multiple phases (De Stefano et al 2016). The first phase is called the pre-scan, which is an initial capture of the object to determine shape and size and position, used for safety measures and planning for the next phase. The second phase is called next best view planning: based on the basic volumetric model from the pre-scan the best layout of photos is planned and captured. In the future, a third phase may be added to capture even higher detail images of certain areas, or around occlusions not detectable in the first phase.

Because the camera coordinates are known, the captured masks and depth maps, processing is designed to become fully automated.

Future developments

While the current versions are functional, Fraunhofer is still working on improving various aspects and development of additional functionality: turntable of their own design, make the robotic arm height adjustable to make a larger object size range possible and the ring light needs to be CE certified.

Essential to any 3D capturing project is a plan for presenting the models. This is often done online. Downloadable data is a potential solution, but often download speeds are restricting the accessibility of the full resolution data, as well as local computing power. Fraunhofer is developing an online viewer that allows private hosting. Another solution that is being researched is “geometric similarity measurement and retrieval”; query a database of models based on shape, independent of metadata or 3D resolution (Tausch et al 2016).

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.4.2 2 and 3D Photography Practice and Prophecies

Participants:

International experts in the field of 2d and 3d imaging of cultural history artefacts. For ICEDIG: Myriam van Walsum, Picturae, Mark Lindeman, Picturae, Agnes Wijers, Picturae

Target: exchange knowledge and ideas on technological digital developments in 2D and 3D photography

Model: Conference at Rijksmuseum, Amsterdam

Background

The Rijksmuseum organised on 8, 9 and 10 May 2019, the third conference 2and3D Photography as a meeting place for the international heritage photography community.

They feel it is essential to have a platform in a time when technological digital developments change rapidly. The conference focusses on the cultural heritage image professional, for and by people active in the field.

Outline

Digital photography is nestled right in the centre of art history, science and conservation. Challenges can only be confronted if workflows are aligned in standardized processes which guarantee the quality of the work. Issues regarding digital curation and sustainability of the image material are current and must become part of the way in which digital heritage is examined. The conference was built around multiple topics: sharing, science, daily practice, 3D applications for cultural heritage, colour management and workflow management.

This report will focus on 3D applications for cultural heritage topics; three lectures will be described, in which used techniques or spin-offs of those techniques could be applied for natural history collections.

3.4.2.1 Integrating Spectral and 3D Imaging for Monitoring Heritage Objects

By E. Keats Webb

Keats Webb is a researcher at the Smithsonian; the world's largest museum, education, and research complex, with 19 museums and the National Zoo—shaping the future by preserving heritage, discovering new knowledge, and sharing resources with the world. Their collections vary from arts to anthropology and natural history. At Smithsonian's museum conservation institute[2], the imaging studio uses a variety of techniques, including spectral and 3D imaging, to enhance or reveal details that might not be seen normally.

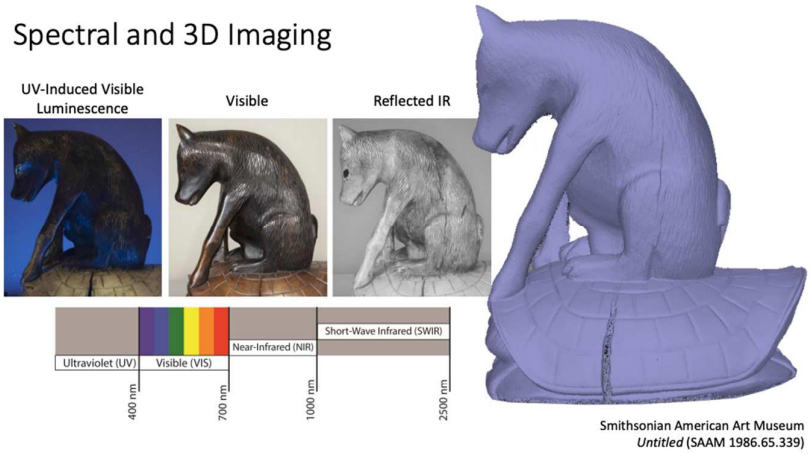
The focus of Webb's presentation is on accessible techniques that can be used by all conservators and heritage professionals. Her presentation describes the use of a modified digital consumer camera to record beyond the range of visible light[3]. The camera is used for techniques such as reflected infrared imaging, which can record under drawing details found below the visible layers of

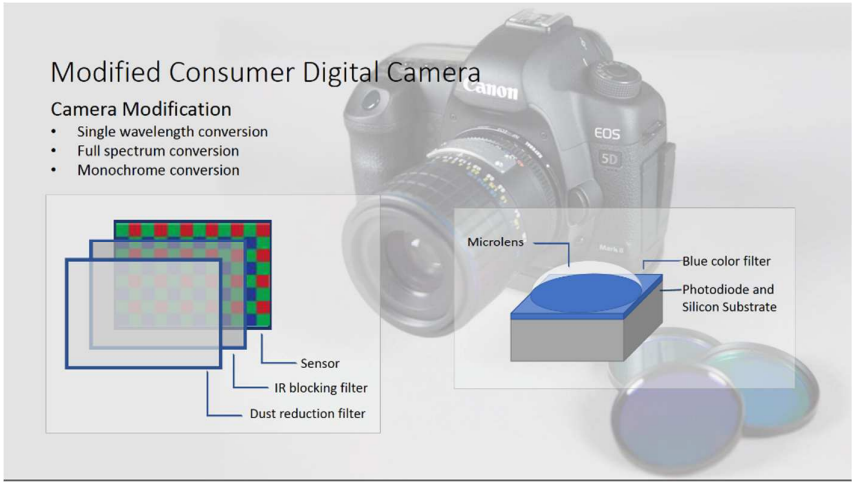
[2] https://www.si.edu/MCIImagingStudio/Multispectral
[3] https://www.maxmax.com/filters

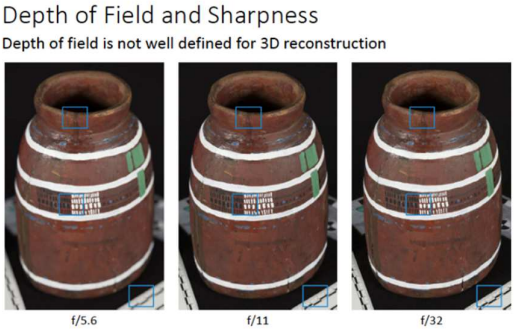
Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

paintings. The modification allows for this technique to be more easily used and more widely applied to collections.

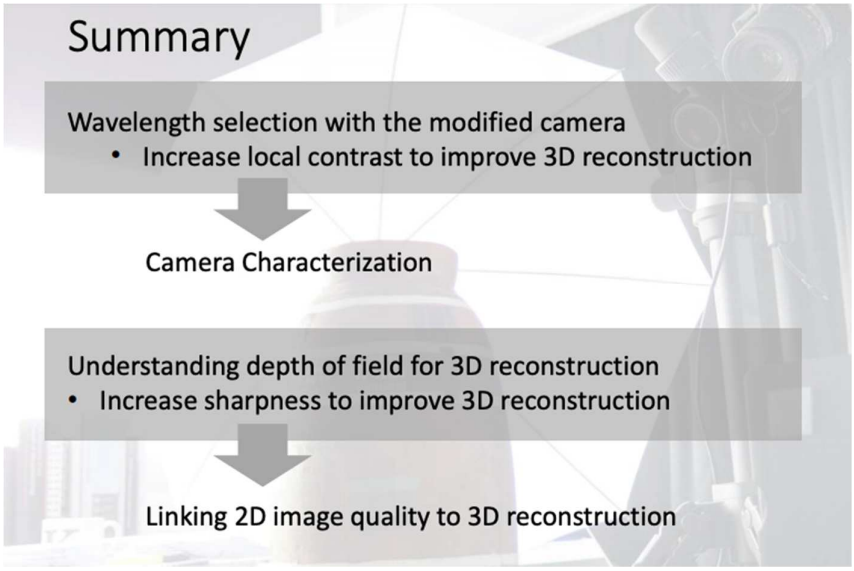
When depth of field is not well defined for 3D reconstruction, sharpness is increased to improve the 3D reconstruction, linking 2D image quality to the 3d reconstruction.







Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483



3.4.2.2 3D digitisation of baroque ceilings, a combined laser scanning and photogrammetric approach

by Dr. John Hindmarch[4] and Prof. Mona Hess[5]

The Baroque Ceiling Project (*Corpus der barockenDeckenmalereiin Deutschland (CbDD)*[6] conducts research on the interior decoration of walls and ceilings created between 1550 and 1800 within Germany. The original project ran from 1966 –2009. In 2014 project was continued and extended with funding for 25 years, to make all the available documentation about the objects, their complex history, and iconographic contents accessible for further research, and for other disciplines on the fields of arts and humanities.

The assignment contains 5000 buildings, including preserved, destroyed and restored examples that need to be documented and analysed. Texts, historical and new photographs, drawings and 3D scans and models. And art historical, architectural and historical contexts. All information must be made available to public online via innovative digital technology and according to Open Access principles.

Schloss Arnstorf is a small village in southern Bavaria. The upper Castle Arnstorf is a moated late Gothic castle dating from the 16th and 17th centuries. It is a private residence with many beautiful ceiling frescos painted from 1714 by Melchior Steidl.

The Kaisersaal (Emperor's Hall) has the most important ceiling, but since the castle is a private residence the art historians from LMU (Munich), photographers from Foto Marburg and the team from Bamberg university had three days access.

[4] https://www.uni-bamberg.de/ddt/team/john-hindmarch/
[5] https://www.uni-bamberg.de/ddt/team/mona-hess/
[6] http://deckenmalerei.badw.de/das-projekt.html

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The CbDD has concentrated on 2D imaging in the past for capturing paintings, since photography will provide better detail, quality, resolution, colour etc. But in this specific case another approach was needed. Many frescos, particularly ceiling paintings are very 3 dimensional and have *trompe l'oeil* effects which can only be experienced from particular viewpoints and via movement.

3D scanning

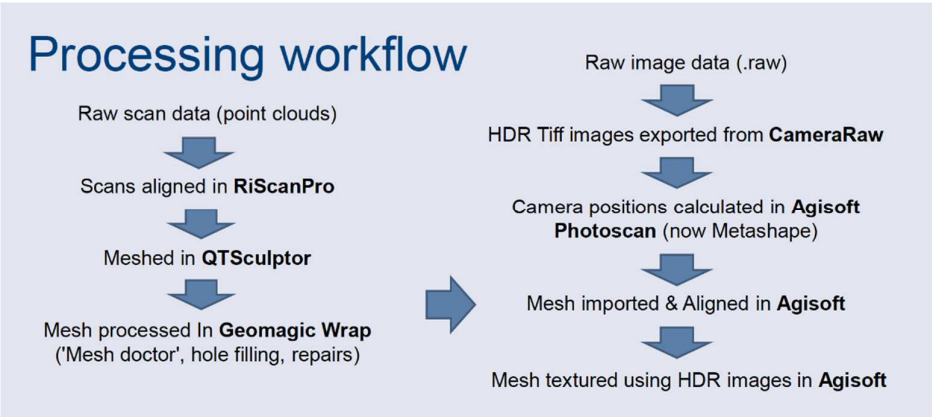
So, the Kaisersaal was scanned from four positions, with aligned point clouds that consists of over 250 million points. Targets placed around the room are automatically identified and used to align individual scans

The used equipment, Riegl VZ-400i Time-of-flight terrestrial laser scanner produces 500,000 points per second at a range of up to 250m 5mm accuracy, and with a resolution of 5mm at 5m. The DSLR on the scanner takes seven images per scan with a fisheye lens.

The photos were taken by Foto Marburg using a CANON EOS 5DS R, full frame 51Mp camera with an EF 15mm fisheye lens, large DoF, since a few photographs needed considerable overlap. Despite of the DoF, a massive distortion blurred some textures.

Workflow

Laser scan has great geometry and terrible textures, while photogrammetry gives beautiful textures but poor geometry. So, why not combine the 2D and 3D data? The 3D geometry data were processed from the laser scan, the scans point clouds were aligned in RiSCANpro software. QTSculptor was used to orient scans and mesh data. Geomagic Wrap was used to edit mesh and meshDoctor was used to restore holes in floor and errors around statues. The crew defined 3D coordinates for obvious points and identified points in multiple photos In Agisoft. The coordinates were entered manually. After importing this in 3D Mesh the texture was as normal.



Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

This workflow is also suitable for objects. All data will be available on CbDD's website and can be used by art historians, architects, heritage preservation professionals, all data will be made available under CC-BY-NC-SA-4.0. A virtual Reality model has been created in Unreal Engine by the university of Bamberg's Cultural Informatics group and can be used with HTC Vive.

3.4.2.3 Smithsonian 3D Open Source Pipeline – from Preservation and Processing to Authoring and Delivery

Vince Rossi, Jon Bundell & Ralph Wiedemeier

As 3D digitisation becomes more common in collections documentation, there is a growing demand for tools to address the specific needs of 3D data stewardship. Central to this is the identification and creation of metadata models to describe these digital surrogates. Just as importantly, such metadata should describe the 'raw' source data from which 3D models are derived and should document the technical processes going into data collection and model creation. There are many institutions and organizations actively working in this area and the Smithsonian's publishes the internal 3D metadata model to add another data point to this critical discussion. They hope this will fuel further discussions within the community, helping everyone identify critical elements which can be standardized to facilitate robust data sharing, and pushing towards the creation of an interoperable metadata model.

Background

The Smithsonian is scaling up 3D digitisation efforts across the institution, moving from low volume, high complexity projects towards high throughput production work aimed at capturing entire collections of objects as 3D models. For these activities, it is important to ensure that the collected data and resulting models are robust, reusable, and easily accessible. To meet these needs and to facilitate digitisation activities, robust IT infrastructure that enables the preservation, management, and delivery of 3D data is essential. To this end, the Digitisation Program Office (DPO) is developing a 3D data-centric suite of open source tools including a content management system, automated processing service, and web delivery platform. Underlying these tools is a metadata model which describes and brings transparency to the complex relationships between raw 3D capture data and the resulting models, ultimately making this data more durable and reusable.

Design Considerations

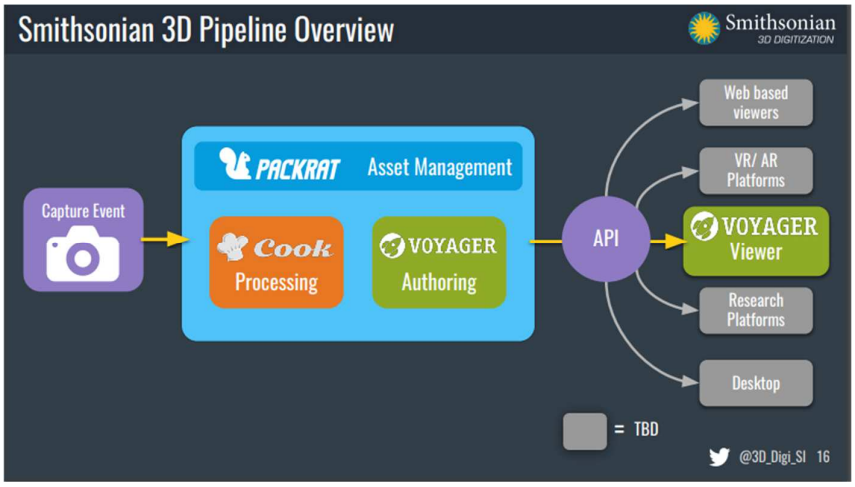
Before the first line of code was written for the tools mentioned above, the basic metadata model underlying them was well under development by an interdisciplinary working group of practitioners, informaticians, librarians, and archivists from across the Smithsonian Institution. The group focused primarily on modelling the metadata needed to fully document a 3D capture event, and specifically focused on photogrammetry capture. Photogrammetry was chosen as a test case because it's an accessible, non-proprietary technology and the 'raw' data, image files are also non-proprietary and have existing best practices for preservation. Additionally, there is a high amount of complexity around how a photogrammetry project can be executed compared to other 3D capture methods. The hope is that, by addressing photogrammetry first, this model will already account for many "edge cases" in capture techniques and will be straightforward to extend to other capture types.

What's Next?

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The Smithsonian has developed a suite of open source tools to produce, manage, and deliver 3D assets from the Institution's collections at an ambitious scale. The beta versions of these tools have been released to the GLAM community as an open source project.

All of these tools are being made freely available to other museums, learning institutions, and commercial entities worldwide, including our new 3D viewing platform and authoring tool, specifically developed for museums and educators.



To support these efforts in access the Smithsonian developed a 3D data repository and a processing automation tool. These systems are built around a metadata model designed to increase transparency and usability of the data and models they produce.

https://cdn.foleon.com/upload/109/p18_2and3d-si-presentation.9a7d78b044f9.pdf

3.4.3 3D Imaging at the British Library

Participants

Donald Cousins, CYREAL, Antony Grant, The British Library, Abraham Nieva de la Hidalga, Cardiff University, Paul Rosin, Cardiff University

Target: Cultural heritage artefacts from the collections of the British Library

Model: Partnership with CYREAL for the digitisation of cultural heritage artefacts

3D Digitisation Method: Photogrammetry

Type of Station: Camera-based digitisation station. Custom-built with off the shelf equipment

Proprietary software for acquisition. Outsourced processing, 3D model building and publishing

Publishing: Sketchfab and linked to pages posted on the BL website[7]

Background

The British Library has an active digitisation program and long experience in digitising 2D artefacts such as printed materials, photographs, and maps. However, it also holds other types of artefacts which can be scanned in 3D.

[7] For instance: https://www.bl.uk/hebrew-manuscripts/articles/the-digital-life-of-a-hebrew-manuscript

In April 2018, Imaging Services installed a custom-built 3D digitisations station. The station is camera based and a robotic turntable. This system enables quick and efficient 3D imaging which is now partially outsourced developed. In this model the image sets from the artefacts are sent to CYREAL for building the models and publishing on Sketchfab. The British Library is still working on devising clear operational workflows for 3D modelling, from Imaging Services, digitisation and preservation perspectives.

3D digitisation project

The British Library in partnership with CYREAL[8] installed and customised a camera-based 3D digitisation station. The station consists of a robotic turntable, illumination equipment and five DSLR cameras connected to a desktop computer and two monitors. Custom software provided by CYREAL controls and synchronises the turntable and the cameras. There is minimal intervention of the operator which only needs to place the artefact, with colour targets on the turntable and start the image acquisition process on the desktop computer.

The modular design of the station means that it can be customised to include more cameras to increase the quality of the models produced and speed up the acquisition process. This station can produce images for either 360° views or 3D models. The imaging in 2D is done onsite, and the 3D reconstruction model is done afterwards by CYREAL.

[8] https://www.cyreal.com/

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Views of BL digitisation station. Top complete view of the station equipment. Lower left the robocase system and the workstation controlling the image capture. Lower right turntable with the artefact being digitised.

The cost of the digitisation station is 6,000.00 € (at 2018 prices) and CYREAL charges an extra fee of 22.00 € for each model created. This can offset the need for hiring and training a digitisation expert, while achieving high quality results from the beginning of operation.

Initially, the British Library has started publishing some models of cultural heritage artefacts online using Sketchfab[9]. However, they are considering different projects which could make use of the models, such as virtual reality recreation of authors offices, creating virtual artefacts which can be handled by users, or using augmented reality to manipulating artefacts next to their display cases within an exhibition. Projects like these are still being refined and analysed, and that CYREAL will also collaborate with the British Library in implementing them

3.4.4 3D Imaging at the National Museum of Wales

Participants

Richard Bevins, National Museum of Wales, Sally Carter, National Museum of Wales, Alex R. Hardisty, Cardiff University, Abraham Nieva de la Hidalga, Cardiff University, Paul L. Rosin, Cardiff University, Xianfang Sun, Cardiff University, James Turner, National Museum of Wales, Mike Wilkinson, National Museum of Wales

Target: Natural history specimens from the museum collections Fossils, Geological samples, Meteorites

Model: Inhouse end-to-end digitisation

3D Digitisation Method: Structured Light Scanning

Type of Station: Handheld Structured light 3D Scanner. Manual turntable

Publishing: Sketchfab[10] and linked to pages published on the museum website[11]

Background

The National Museum Wales has an active digitisation program and long experience in digitising cultural heritage artefacts and natural history specimens. For instance, they have contributed

[9] 23 models published at: https://sketchfab.com/britishlibrary
[10] 74 models published at: https://sketchfab.com/museumwales
[11] For instance: https://museum.wales/blog/2016-09-23/Step-into-the-3rd-Dimension/ and https://museum.wales/blog/2017-07-05/Down-2-Earthwith-an-impact/

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

herbarium sheets type specimens for the global plants initiative. Additionally, the museum also publishes their specimens with images online on the museum website[12].

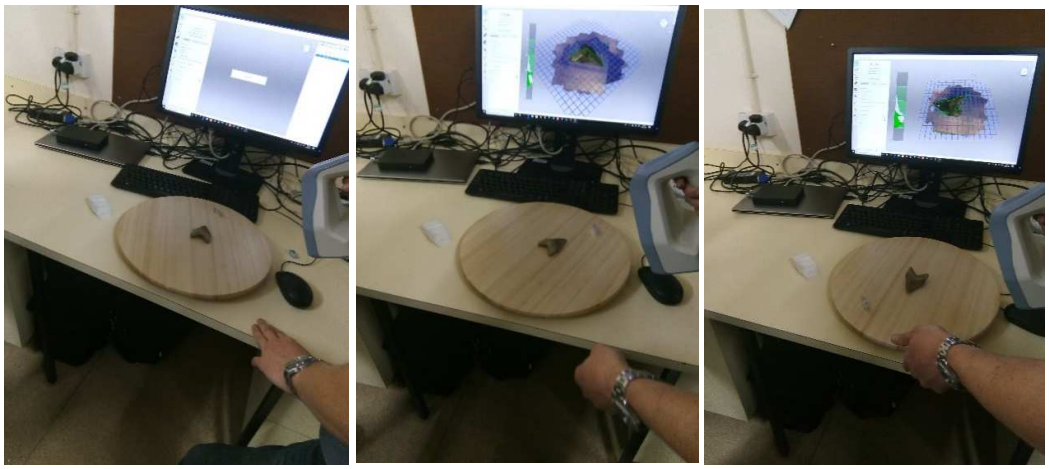
The interest in 3D digitisation of specimens led to the acquisition of an Artec Spider Scanner for documenting relevant specimens and artefacts. The scanner is hand-held (can be mounted on a tripod or a robot arm) and the digitisation operator moves it around the specimen and rotates a manual turntable. This system enables quick 3D imaging with default setting which requires minimal processing after acquisition. The National Museum Wales is still defining their operational workflows for 3D digitisation.

3D digitisation project

The National Museum Wales has a dedicated digitisation program. They consider that 3D digitisation is a valuable medium for publishing relevant specimens, but not for mass imaging of the collections. They follow a similar approach with all their digitisation projects.

The selection of the structured light scanner acquire by the National Museum Wales was based on research and recommendations from other institutions already doing 3D digitisation. The 3D scanner comes with proprietary software which allows some processing; however, it also allows exporting to multiple formats, and this facilitates using other software if further processing is required.

The scanning software provided by Artec controls the data acquisition and coordinates a colour camera which allows acquiring texture/colour information simultaneously, producing coulored models. Apart from rotating the specimen and moving the scanner around, there is no other intervention from the imaging technician the model is saved on a project folder in the host computer file system.



Views of digitisation station consisting of manual round talbe, scanner, and host computer in three differnstages of the scanning process..

The cost of the scanner like the one used at the museum is 19,700.00 € (at 2019 prices[13], however for education and research Artec offers a reduced price of 17,730.00 €). The scanner requires little training for operation and the software is easy to learn, which can offset the cost of training or hiring a digitisation expert, while achieving high quality models will require further use of more advanced processing software.

[12] https://museum.wales/collections/online/

[13] https://www.artec3d.com/portable-3d-scanners/artec-spider.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The National Museum of Wales has started publishing some models of specimens online using Sketchfab[14] and producing articles for their website linking to those models. However, they are also piloting other uses which could make use of the models. For instance, currently they are running an augmented reality pilot[15], borrowing handsets which use augmented reality to show some of the specimens and artefacts in the galleries.

3.4.5 Conclusion

The choice to visit different developers and users of 3D techniques, turned out to be very practical and effective. Though the investigated and described techniques (see also Deliverable 3.7[16]) are not always directly applicable yet to speed up imaging of natural history collections, the potentials for those techniques to become common practice as part of NHC digitisation projects are clear.

Different aspects were closely investigated, both from the point of the user (British Library and National Museum of Wales) and from the developer (Fraunhofer). E.g. determining the attributes of physical specimens which make them candidates for 3D digitization and identifying the appropriate 3D digitisation technologies which may be applicable for each type of specimen.

At the 2 and 3D symposium at the Rijksmuseum, it became clear that natural history collections can learn a lot from research on imaging that has already been done in the cultural field. Multiple topics as science, 3D applications for cultural heritage and workflow management were shared with the audience by clear example projects. Time between lectures was efficient, since so many experts were together.

This modus operandi of visiting instead of inviting, also gave us good insight in the ways in which the current 3D digitisation techniques are used and how they can be improved and made applicable for natural history collections. It furthermore made clear that large benefits can be achieved if guidelines are developed in consultation with manufacturers, industry and cultural heritage. Guidelines that inform areas of research and development for larger digitisation projects in the context of DiSSCo.

3.5 Roundtable Five – Partnership Frameworks for Distributed Research Infrastructures

3.5.1 Introduction

This roundtable was organised to discuss the challenges and the specificities around the development of robust and effective partnership frameworks for RIs, during their preparatory, implementation or operational phase.

Attendees consisted of experts from (developing and operational) RI networks and organisations, and e-Infrastructure organisations.

During the RT, participants shared experiences and planned approaches in partnership development, discuss prioritisation mechanisms for core infrastructure operations. Existing partnership frameworks will be used as a guide to discuss best practices in maintaining flexible and robust operational and strategic frameworks.

[14] 23 models published at: https://sketchfab.com/britishlibrary
[15] https://museum.wales/cardiff/whatson/Museum-ExplorAR/
[16] https://doi.org/10.5281/zenodo.3469531

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The context was through an update on the future of the EOSC from Andreas Veispak, the head of DG Connect, European Commission, and a long-term observational perspective from the chair of the ESFRI innovation group, Jean Moulin.

The round table took place in two sessions:

1. Addressing **Coordinating of governance and prioritization across stakeholders and national boundaries**
2. Addressing how to best **Leverage the existing infrastructure landscape and strengthen development between e-Infrastructures and research infrastructures**

Questions being contemplated:

- (bi-directional) How do you engage with research infrastructures and e-Infrastructures hitherto not known?
- how do you create partnerships with industry and infrastructures, to scale up from innovation to production level services?
- how do you collect/exploit synergies to achieve greater harmonization in service development and provisioning?
- how do you map and implement uniform policy practice across a diverse and dispersed political landscape?
- What operational functions have you put in place to govern across national borders? What works well? What ongoing challenges does this model bring (if any?)
- What mechanisms do you use when prioritizing scientific and operational tasks?
- How do you balance institutional strategic objectives with strategic objectives across a consortium?

By the end of the meeting, we had touched upon the key parameters of the issue and highlighted the dimensions that RIs management bodies need to consider, from a practical point-of-view for developing a comprehensive RI-specific partnerships framework.

Dimitris Koureas opened the meeting by introducing the DiSSCo Research Infrastructure (RI) which focuses on scientific collections as the fundamental assets through which we understand and describe biological and geological diversity. Each specimen in these distributed collections is the source of a significant amount of data. This data is there to underpin the work of other environmental RIs, such as LifeWatch, eLTER, etc.

DiSSCo is also about changing the way Institutions holding natural science collections (NCS) do business. Aiming to act as if were a single (virtual) organization which means working out common policies, workflows, and procedures.

When it comes to partnerships, challenges arise from the distributed nature of the RI. It will exist in a complicated landscape of legal limitations and national interests that will have to be navigated. Partnerships will have to be forged along three main lines: within DiSSCo with the national nodes, with other players – e.g. other RIs and their technical and strategical interfaces, global partners (e.g. via GBIF, iDigBio) that either serve or incorporate DiSSCo's mission – and with the underlying foundational providers such as the EOSC for underlying e-infrastructures. The question there is how to organise this based on the RI's ability to procure or rely on services that allow the RI to operate. One advantage of DiSSCo is its fundamental partnership with CETAF, rooting the RI in its own community.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.5.2 EOSC Status and Perspectives – Andreas Veispak

The EOSC aims to be a trusted and open virtual environment with seamless access to services addressing the whole research data life cycle. This includes federating and connecting existing and planned RIs, making data FAIR and ensuring its long-term preservation, as well as offering services to find, access, combine, analyse, and process data. It will be federated, open, transparent, and based on the rule of law, preventing lock-in by individual service providers and Maximising digital capacities available to researchers. It will also support public authorities in informed policy development and implementation, including for key societal challenges and help stimulate the emergence of a competitive EU cloud sector.

Basing the EOSC on the rule of law could be a unique selling proposition globally as it is a differentiating characteristic from the USA (individual companies being free in their reign) and China (CCP steered). Realising it will also contribute greatly to the planned strategic autonomy within the data economy, where Europe currently demands one third of the resources but only supplies 2% of the computing resources. It is therefore crucial to act quickly here and allocate the necessary resources.

To be as inclusive as possible, the EOSC plans to be accessible through a non-exclusive, simple, universal access point, Governed by clear and unambiguous Rules of Participation, and inclusive across borders and disciplines. It is being built on three pillars - data storage and management, network, computing resources – and the involved parties are working on bringing together very different actors (research community, commercial sector, public authorities) in a coherent ecosystem.

So far, the EOSC has been a supply side push, namely through ESFRI, which has not led to high acceptance numbers. Now the EC is trying to add a demand side stimulus, e.g. by developing climate and sustainability policies. There are also initiatives to link together the principles of open science and the digital single market and to regulate the digital single market much more, like transport or energy. RIs therefore should prepare early on to comply with upcoming regulations. To develop the EOSC further, currently the Working Groups and the Executive Board are working on the federated core. Defining the core only builds a trustworthy base while leaving enough freedom for other actors to define the surroundings.

Currently, there are 30 active Horizon 2020 projects contributing to the EOSC with a total investment of half a billion Euros. 300 services from 100 providers are already being offered in the portal. However, this mode of financing has brought about its own problems, with a distinct lack of provisions for sustainability or the maintenance of IPR after the projects end. This leads to a researcher's dilemma: What is the confidence that an average researcher will have in the continuation of various services and pieces of infrastructure over the long-term? Nonetheless, there are two important calls upcoming, INFRAEOSC-03-2020 and INFRAEOSC-07-2020 on creating commonalities and defining the federated core and widening services, respectively.

The upcoming 18 months will be critical for shaping the EOSC. One vision is to have FAIR Digital Objects implemented, turning the principle into practice, by the end of 2020. This is also the timeframe for multipole reviews and reports on the EOSC by the Commission and the Member States.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.5.3 Distributed Research Infrastructures: Issues and Options – Laurence Lenoir

From her experiences in the ESFRI Innovation Group, Laurence Lenoir could report that during the preparatory phase, it is the main aim for RIs to change their way of thinking and managing from a network to an important infrastructure, a single legal entity. The preparatory phase project must be used to consider legal aspects, governance, the functioning of a distributed RI, the role of the nodes, which services will be provided (access, training, calls?), which income sources can be unlocked, how costs can be distributed, etc.

From her point of view, it is also paramount to involve the ministerial level early on because they will have the last word in many decisions later on. This can include giving them concrete tasks. Otherwise a gap opens between institutional needs and ministerial plans that will be difficult to bridge.

She also listed several other important points:

- Keep close contact with cluster projects, as they are aiming for common services to shared needs.
- Consult the document on good practices for sustainable RIs. The KPIs and monitoring processes to evaluate RIs are currently being developed and each RI will have to undergo a 10-year evaluation to see if they can keep the ESFRI label.
- Engage prospective user communities early on and go beyond the direct scientific community. It is very important to be associated with them to get their feedback to ensure the sustainability of the RI.
- Do not underestimate the contributions needed from the institutional level. It needs to come from their business mission to be successful, even if the main pillar of funding is governmental in the later phases.
- Be ambitious but remain realistic with a progressive plan, from which ministries can understand what is needed and how they can convince funding authorities to release budget to the RI.

3.5.4 Dispersed Nodes in Practice – Jerry Lanfaer

ELIXIR connects national bioinformatics centres and EMBL-EBI into a European infrastructure for biological research data. It is operational since 2016. Within ELIXIR, the central hub is funded by national contributions but returns a large proportion of that money to nodes to stimulate coordination activities across the nodes. This leads to harmonizing across countries, institutions, etc. One track of funnelling funds to nodes are joint bids for funding e.g. form H2020 that are based on the hub-node-partnerships.

Another track are services that are developed and operated at national level by nodes but offered internationally through the hub which organises these via a 5-year technical programme in which the hub acts as the coordinator. The hub thus does not do much development or operations by itself; most is done at node level. Hundreds of services may be offered by the nodes, while the hub is content to offer on key infrastructure services.

In terms of governance, the ELIXIR director works with the heads of nodes committee and reports to the ELIXIR board which meets twice annually. Additional to the scientific community and national

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

node governance arrangements, there is an industry advisory body that helps to maintain links with industry.

3.5.5 Aligning Scientific Priorities across Borders – Luca Pezzati

Luca Pezzati, Scientific Co-ordinator of E-RIHS reported from their experience, two years into the Preparatory Phase. The E-RIHS consortium consists of partners from 28 countries with several countries joining recently because they could be convinced it was a good idea. For the initial RI, they set themselves a KPI of 12 founding members which is a hard-enough task, even out of a pool of 28 member states. It is very important to know the partners in order to formulate a proper business plan.

E-RIHS tries to manage complexity by imposing conditions (such as national agreements on national nodes in place) on the partners. They also try to create complementary regional initiatives e.g., in USA, Mexico or Brazil because no such thing exists there yet. In the US, after a lengthy exchange process to focus and establish cooperation of over 4 years, a group under a single coordinator formed. Managing complexity simply needs time and a lot of negotiations. In Brazil, a formal association of laboratory coordinators exists and is the main contact point. They try to be inclusive to other countries in South America and Argentina is on the cusp of joining the initiative.

3.5.6 Making Open Science Work – Tim Smith

Open Science will work once scientists automatically share the data underlying their work to make it transparent, according to Tim Smith. Other scientists can then see and use that data, they enable a different kind of peer review and even citizens can access them. Generally, the move towards open science already has established a much more collaborative way of doing and publishing research. It is however important to note that sharing is not the same as publishing or preserving. Functioning services are needed for each of those.

To enable true open science, these services need to be usable and have open APIs. This way, every act of reporting of science can link back to the data underpinning it. Open APIs often come with hundreds of connection requests for new services which always includes some one has not considered before. This means that not all cases can or have to be known beforehand.

Opening up digital objects and giving them easy to use APIs also stimulates new ideas on use and re-use of data. E.g. a peer-reviewed journal of brief ideas. The objects need to be permanent, the services will then develop around them. This also means that they – services and enriching layers – in turn do not need to be permanent but their added value such as feeding citations back to the object or enriching the linked data will. They can appear and disappear to fit a specific purpose. Therefore, one should focus on building elements form which environment can build itself, not focused on building RI

One contentious item when discussing Open Science often is data sovereignty interests. To this end, InvenioRDM will be a one-click RDM in a box product which allows Zenodo deposits to be held locally on a self-installable instance. This will allow peer-to-peer preservation and empower institutional repositories which are then linked to repositories around the world.

3.5.7 Discussion - I

In the subsequent first round of discussion, several items were addressed, like the insight of Tim Smith, that non-permanent services that enrich digital objects with higher longevity will become the

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

norm. Users, with researchers among them, do not develop loyalties to these services but simply sue them for short periods of time and then move on. Therefore, services cannot necessarily be designed in their entirety in advance. Whatever becomes available should then be exploited and the desirable parts and results must be kept in permanent stores. Closely linked to this development, Smith expects a bigger prominence for micro-transactions and pay-as-you-go options to pay for these and other services. While it is likely that the use of micro-transactions is widespread in the business world within 1-2 years, he expects it also to be used in the science domain within 4-5 years.

Another item was the importance of Trust in what an RI is offering to be successful. As a user, once tends to imitate the behaviour of one's peers and only in a second step one looks outside the institution or country. Trust and longevity are particular important for activities in the heritage domain where preservation and observing FAIR principles have to be secured for decades and centuries. Currently popular services, like Amazon Web Services (AWS), have not been around for very long. Companies tend to fluctuate. Still, they are being trusted more so – through clever marketing – than long-established research or public institutions. But this natural fluctuation is an issue for long term preservation of data. When choosing the proper service for an RI, other things to consider were brought up:

1. Ministries fund large paths of the RI. Choosing a private service like AWS over public service providers could have negative implications.
2. Private-public collaborations need more control and regulation. When spending public money, an extra layer of transparency is appropriate.
3. RIs do not only procure services, they are also offering services themselves. Therefor, a web-service that does not enable this is not going to be sufficient.
4. Commercial providers often consider the commercial value of research data to be zero and therefore refuse liability or compensation in event of loss.
5. Commercial services are not built for research sue. Level of trust is therefore usually higher (e.g. EGI through person-to-person support vs. AWS) once the users make the choice for a public provider.
6. When it comes to sensitive data, (e.g. in life sciences), data export beyond national boundaries is often prohibited, making it difficult to choose a private provider if the servers are out of country.
7. For securing your processes, e.g. the preservation of research data, against a private company discontinuing its services, it is often necessary to invest in some kind of insurance. This could be accomplished by procuring a similar service from a different company. If this additional insurance cost can be saved by procuring the same service from the public sector, this should be very much part of the budgeting equation.

The question then is how public infrastructures – e.g. EOSC – can be encouraged to invest more in building that trust. On top of that, RIs need to become competitive as well in the second layer of which the quality of services, trust in data, the fitness for purpose and the ease of use are just a few categories added to the first layer of long-term preservation.

A linked item was the situation around the procurement of services. For now, this has mostly been explored for the procurement of computing time, e.g. by some ELIXIR nodes but not coordinated across the nodes. None of the present RIs procure common e-infrastructure services at present. This

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

is at least partly because many RIs are still in an early stage where the focus is on developing core services that largely will be operated by the RI itself. Computational capacity, storage or AAI are some good examples of services that can easily be procured elsewhere. However, these services currently are contributed in-kind as a service to the RI from one of its partners which further explains the lack of procurement activities at the moment. It currently also seems simpler to operate these kinds of relations on the national level instead the European.

Other items of note during the discussion were that:

- it is very crucial to a RI's success to obtain and utilise the user requirements at the earliest possible moment, and that
- a workshop on EOSC-ESFRI inclusion has already been held (January 2019 in London) and a second one scheduled for early 2020.

3.5.8 The Future of Synchronisation and Sharing Infrastructures – Jakub Moscicki

The CS3 community is a worldwide, bottom-up community working for five years on cloud services for synchronization and sharing. Members come from all domains including national research agencies, HPC centers, organisations, universities, and companies of all sizes. They do so without any to-down funding, showing that people participate because the issue is relevant.

Now for the first time they do have a funded project, CS3MESH4EOSC, to capitalize on the value created by the community. The project is set to start in January 2020 and aims to deliver a global service for European participants without creating anything new for the users. Objectives are to deliver a global collaboration service for European research, education and public institutions and to create an environment for collaboration on technology, applications, use-cases, software and operations.

The CS3 community realised that various on-premise enterprise file sync and share services (Dropbox-like) work very efficiently, that any different services in use and that usage usually spreads by word of mouth dissemination. That is why they are working on connecting large existing data repositories to make them interoperable via standardised, vendor-neutral APIs to integrate services rather than inventing something new. Users on different sites would then be able to collaborate with ease. They have a good foundation in a large exiting user-base (300-400,000 users) and a good insight into the user needs of the research community.

The project very much contributes to the EOSC. A result of the project is collaboration across sites for file sharing and synchronisation, irrespective of specific technologies in use locally. Added to this, there will be a layer with applications on top, e.g. interactive data science environment based on Jupyter Notebooks, distributed processing workflows and large dataset sharing, collaborative editing of content for office documents, open data applications such as FAIR metadata collections or Open Data publishing systems based on OpenAIRE. The Science Mesh could then be integrated into the EOSC-hub catalogue.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.5.9 Coordinating Initiatives across Research Performing Organisations and e-Infrastructure Providers – Lene Krøl Andersen

Lene Krøl Andersen led with the example of a global vegetation dataset that is stored in Svalbard with a lifespan of 100 years. It is generally hard to discover, even though it is available in GBIF. So, the question becomes how to make it easier to discover and link these and other datasets?

This is where EOSC Nordic comes into play, one of the four regional implementation projects for the EOSC. Comparisons between EOSC Nordic and what DiSSCo needs to do can be drawn. Some of the relevant elements are a knowledge hub, service providers, cross-border services, FAIR practices and the user community.

The focus however will not be laying on the technology, but rather everything else, including policy. This comprises e.g. an analysis of where and which EOSC services are blocked at the (data) border, how to make the back-office work for cross-country service provision, how service access from a researcher in another country actually works, etc. It also includes a thorough analysis of which services are being used and to what extend to then only open up the ones that are being used. This process of identifying the core services is currently ongoing.

Another important step is the implementation of FAIR data. To do so they will rely on best practices from other projects (e.g. FAIR is fair) which could be useful for DiSSCo as well.

One of the goals of EOSC Nordic is to demonstrate the potential of the EOSC to discover research data. Stakeholders in this context will also be governmental entities as policy will play an important role in making EOSC accessible and usable.

The GitHub page of the project can be found here: https://nordicesmhub.github.io/about.

3.5.10 Discussion - II

In the second round of roundtable discussions, the business plan and the potential relationships to industry were debated. It was noted that it is important not to change the core of the RI when thinking about a business model or relations to the private sector, to remain an RI primarily geared toward research. That does of course not mean that relations with commercial entities should not be explored and we could expect 5-10% of services to mobilise towards the industrial sector. This would go a long way to show that what we build is important for the industry, too, which in turn could boost the business model. This could then be one of the revenue streams for an RI that need to exist next to the base funding from the member states. Said base funding will not always support all the ambitions. Therefore, flexible income from sponsorships or fees is necessary. This could come in many dimensions of the RI such as training, conferences, and – though less common – in the provision of e-infrastructure services. It should be made clear early on what kind of basic capabilities must be funded by the secure funding from national governments and what is more ambitious and needs additional revenue.

A difference that needs to be considered when setting up the business model is the uniqueness of RIs: the business model is not based on selling of products and services to end-users. When an RI manages to double its user base it did not necessarily double its clients which are usually the national funders. Therefore, it needs to be ensured that products and services offer long-term added value. Even though this uniqueness makes the creation of a business model a difficult exercise it is

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

important to have a notion of operational costs early on, including in-kind contributions. This is to make sure that at least five years of operation can be secured in funding from the beginning.

When reporting form experience, some participants said that when exploring the services to provide to industry, their relations are more with SMEs that with multinational corporations. That is partly due to the fact that large companies tend to delegate R&D activities to smaller companies and then buy the successful ones.

A second main item of discussion was the convergence of the RIs in the common landscape. The different state of RIs and their timing do not help in this endeavour as different facets, like needs and solutions, develop at different speeds. This also applies to the relationship between RIs and the e-infrastructure providers. Nevertheless, RIs will have to invest efforts to identify interfaces with common lines of production from where then common services could emerge. While parts of the construction of an RI is based on the needs of the specific communities, it is always possible to define high-level requirements that apply across different RIs. Implementing FAIR principles early on, especially interoperability, can prevent a disconnect.

Another issue that was identified was the apparent difficulty of the outputs of science projects, be they funded by DG RESEARCH or DG CONNECT, are great for innovation but not at operationalisation. Their lifecycle does not necessarily support going the next step. To become scalable requires a different approach than to innovate.

ESFRI however has been quite successful in going beyond project-oriented R&D and innovation toward delivering products and services already in production quality and sustaining them. This also requires better standards to open up those products and services. It also helps to have a critical mass of users who can help improve and hone services through the iterations towards production quality.

Some lessons can be learnt: In EUDAT, running a high number of pilots with different research communities helped understand and produce services that are needed by many of them. The CS3 approach on the other hand speaks to the emergence of a response to researchers needs that already are practiced but fragmented way and not completely addressed. RIs appear to need some kind of global access layer that sits above all the different services and data repositories. CS3 does not want users to change their own local interfaces, but to simply discover new capabilities being added to their system. This is enabled by the interoperability work.

3.5.11 Conclusion

During the interesting interventions and fruitful discussions, it became very clear that we are just at the beginning of this conversation around the challenges building effective, trustful and long-lasting partnerships. The purpose of this meeting was to bring it into a higher gear and address some more specific aspects, such as relation to e-Infrastructures or how to procure, co-develop, and co-operate services.

Another interesting aspect was the discussion on the funding framework. Maybe this could be improved by integrating all aspects of R&D, including innovation, deployment and operations.

Dimitris Koureas thanked all participants for their participation and expressed his wish to continue this discussion soon. All the present representatives from RIs echoed that sentiment.

3.6 Roundtable Six – Museums specimen and molecular data linkage (DiSSCo/ELIXIR)

This Roundtable was jointly organised by DiSSCo, ELIXIR and CETAF under Task 9.5 ("Liaising with the research infrastructure landscape") and held on 26th and 27th of September 2019 at the Royal Belgian Institute of Natural Sciences in Brussels. The first goal of the workshop was to explore possible use cases based on linking data in molecular databases to museum specimens. The second goal was to consider how this work could be taken forward in practical terms (e.g. funding aspects) as well as involving a broader set of stakeholders. It is hoped that this effort will result in new data becoming available to an enlarged user base thereby increasing the impact of publicly funded research and services.

Fifteen people attended (see Appendix 4.4 for the list of the participants) and among them several ICEDIG participants were present. The Roundtable was set up into two half-day sessions where the first day was focused on data and service landscape and the second day was focused on future collaboration and funding opportunities. The participants agreed that this was a well-organized Roundtable with a focused agenda and specific goal that were clearly communicated to the participants beforehand. Also, various background materials were collated in GitHub before the meeting to facilitate the discussion.

The day one session started with all participants giving a short introduction. After that Dimitris Koureas (Head, DiSSCo Coordination Team) introduced the goal and vision of DiSSCo. He highlighted the importance of natural history collections around Europe and the idea of "Digital Specimens" -- digital representations in cyberspace of physical specimens in the natural history collections. This presentation also included a brief overview of the technical architecture of DiSSCo and the idea of FAIR Digital Object (DO) which provides an actionable knowledge unit for the physical specimens.

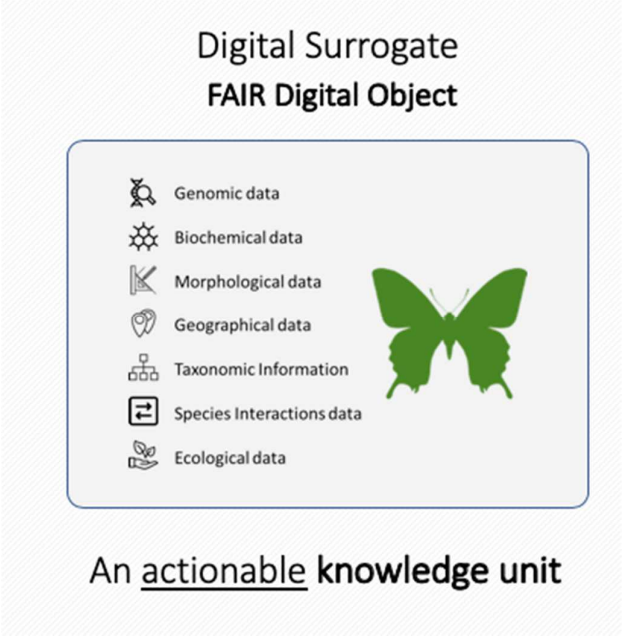


Figure 1: DiSSCo FAIR Digital Object. Source: Dimitris Koureas.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ELIXIR introduction was provided by Jerry Lanfear (CTO, ELIXIR). ELIXIR connects national bioinformatics centres and EMBL-EBI into a sustainable European infrastructure for biological research data. The biodiversity data sources were not in the original scope of ELIXIR but slowly their domain is expanding. A new biodiversity working group within ELIXIR members just had their first meeting. This working group will focus on the issue of FAIR and understanding the complexity of biodiversity at the molecular level that is cross-linked with many fields.

The next presentation was by Sharif Islam (Data Architect, DiSSCo). He introduced various technical materials and outlined the data and service landscape. He provided examples of heterogeneous data sources, workflows, and standards from different domains (for example "traditional" collection and taxonomy-based workflow versus newer DNA sequencing methods) that are relevant for data linkage. The group also discussed a specific example of a specimen (see figure 2 below) and the manner it is currently linked and not linked to various data sources. This provided an opportunity for a lively discussion on workflow, data authority, and data curation. A detail description of this diagram and the related links can be found here.

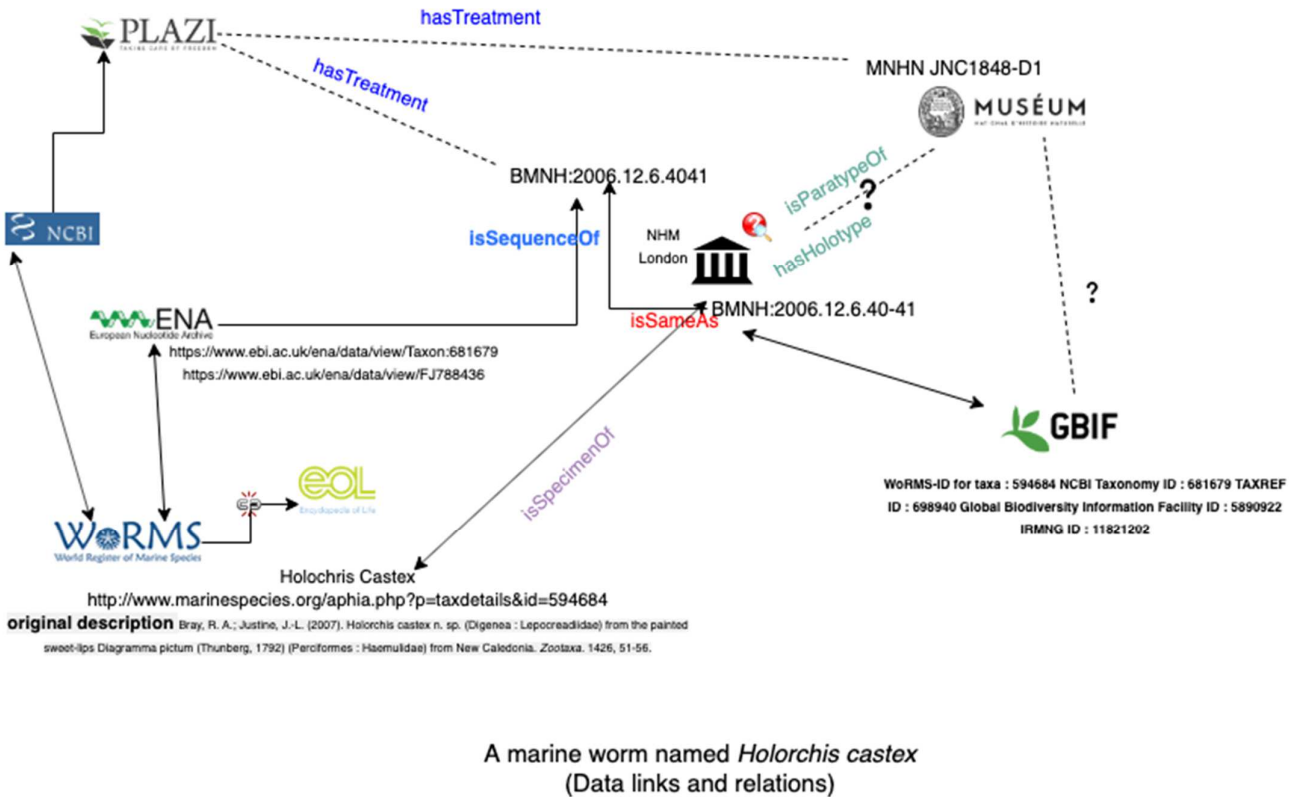


Figure 2: How a specimen of a marine worm is linked. Source: Sharif Islam

Sharif Islam also presented a simple schema that identified the data flow and the service landscape (from field collection to the database). Again, the focus here was to show the heterogenous data sources and workflow.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

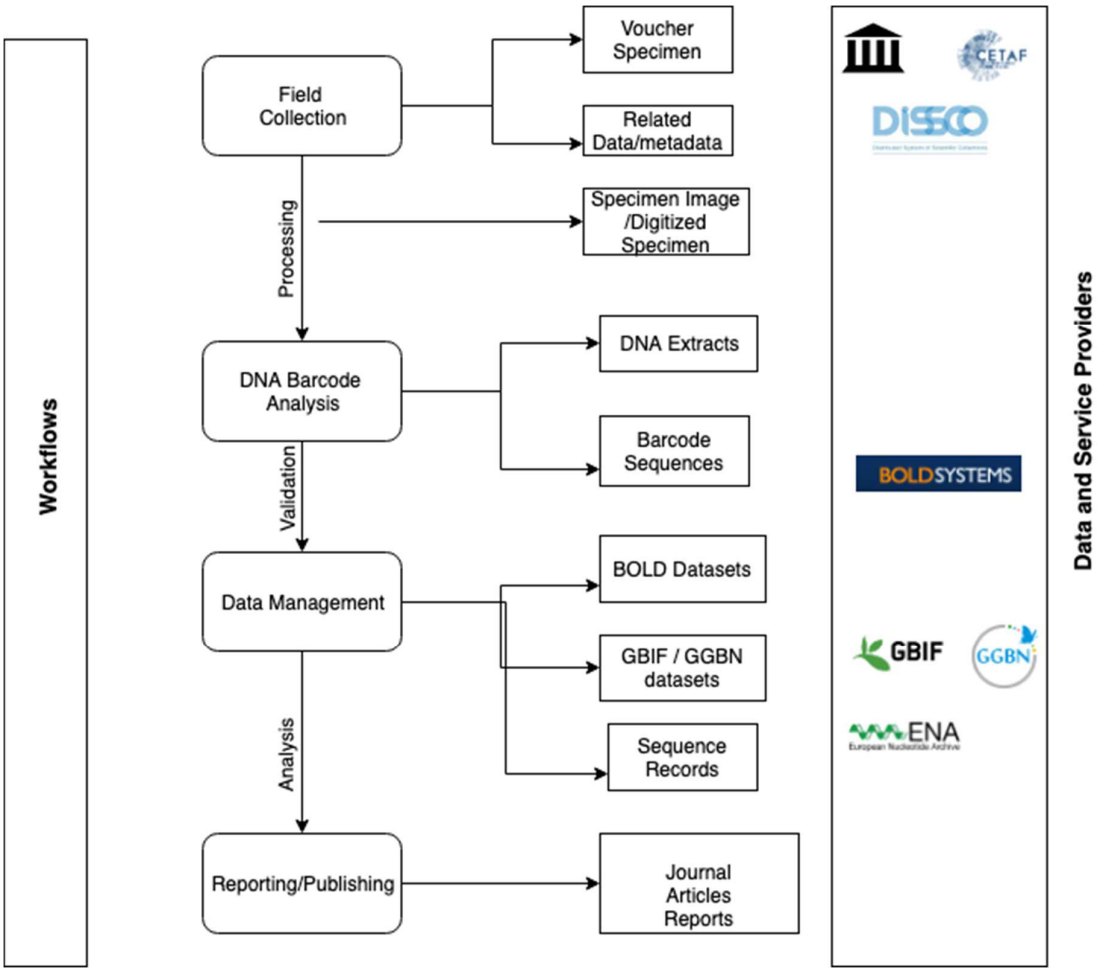


Figure 3: Workflow and service landscape. Source: Sharif Islam

After the formal presentations were finished the floor was open for group discussion. Instead of handling the full data lifecycle in the domain of museum specimens and molecular databases, the group decided to focus on services that revolve around persistent identifiers. This provided a concrete focus for the discussion. The issue of persistent and globally unique identifier is a much-discussed topic in the biodiversity and natural history collections community. Based on the current financial and organisational reality a single solution will not be available for the wider biodiversity and molecular research community to adopt. Therefore, any solution at this point needs to be flexible and applicable to multiple domains. The group agreed on this and understood the technical and organizational challenges. DiSSCo and GBIF are actively involved in this global discussion. ELIXIR's work with Identifiers.org is also relevant in this space. However, the concern was raised that we need to move forward and find a way to take advantage of the current technical solutions that are already available. Often it is more about workflow and organisational challenges than technical obstacles.

In that vein, after some brainstorming, a simple trigger-based workflow was proposed that can be a way to build a global digital specimen catalogue for DiSSCo. In this proposed workflow the DiSSCo

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

NSiD (Natural Science Identifiers[17]) interface could interact with ENA (The European Nucleotide Archive) API (ENA currently holds 7,5 million sequence records that contain some form of voucher specimen identifier-based Darwin Core Triples) or other third-party data providers. This interaction will need to comply with the openDS standard that is under discussion in various DiSSCo projects. For example, using the ENA API the following JSON snippet can be harvested (which contains voucher id "RMNH D 38033"). This can then be used as an input to create a Digital Object for DiSSCo. During the creation of this Digital Object a new globally resolvable persistent identifier will be minted and assigned. At a later period, a DiSSCo member (a collection holding institute that is in charge of RMNH D 38033) can start using this digital object and the persistent identifier to enhance the data and for a variety of other services.


```
{
   "accession": "AM076944",
   "country": "Atlantic Ocean:Caribbean Islands St Vincent",
   "scientific_name": "Stylodactylus serratus",
   "specimen_voucher": "RMNH D 38033",
   "tax_id": "342640",
   "type": "Digital Specimen"
 }
```


This solution provides a solid starting point towards a more comprehensive global specimen catalogue. At the same time, it starts using a robust persistent identifier mechanism that needs to be integrated in the current natural history collection workflow. Based on this idea, the DiSSCo Technical Team will follow up with ELIXIR/ENA for further tests and proof of concept. This concluded a very productive first session where the participants agreed that this was a worthwhile endeavour that produced an actionable technical idea.

On the second day the focus was on identifying funding opportunities and future collaboration. Andrew Smith (Head of External Relations, ELIXIR) joined remotely to provide a brief overview of various EC funding opportunities. According to him INFRAIA-2-2020 – "Integrating Activities for Starting Communities" -- would be a possible source. The group discussed that even though the partners are mature players, the community hasn't yet come together to integrate properly (in particular combining specimen/collection holding institutes, molecular/sequencing databases and biodiversity literature). Donat Agosti (President, Plazi) also joined remotely to discuss and proposed "European biodiversity and genomics data library" which will provide access to data in publications, not access to publications per se. The group agreed that there are various ways to go about this, but it is important to identify the potential coordinators and partners. The group also agreed that in order to transform the ideas described in the Roundtable into something concrete a solid funding proposal is needed. At the end of the session, the following action items were proposed:

[17] DiSSCo is currently proposing a PID schema/service for natural sciences that would be applicable across a broader community. This will be handle based with a top-level prefix. For example (http://hdl.handle.net/20.5000.1025/486a7e883f14f88bba37).

1. Define and brand the starting community (new cross disciplinary community designed to address the need scientists have to assemble complex *corpora* of knowledge around specific Biodiversity informatics challenges.)
2. Describe example use cases. See example.[18]
3. Scope the idea and describe in 1-2 pages for the proposal.
4. Identify and communicate with potential Partners: EMBL (ELIXIR Hub + EBI), SIB (Swiss Institute of Bioinformatics), DiSSCo and institutions, Zenodo, Pensoft, CoL (Catalogue of Life), GBIF.

3.7 Roundtable Seven – Humanities Researcher Synergies with Natural Science Collections and Archives

3.7.1 Summary

The roundtable was organized by Naturalis Biodiversity Center and took place on 14 January 2020 from 11:30 - 17:30 at the Tulip Inn Leiden Centre, Leiden, the Netherlands. There were 13 attendees (Appendix 4.5.1). The official proceedings of the roundtable were preceded by a buffet lunch at the venue providing an opportunity for participants to get acquainted.

Roundtable Objective

To define the digital requirements of cultural heritage researchers

working at the interface of science and humanities

that are needed to facilitate their use of museum, university, and botanic garden

natural science collections, data and associated archives.

The structure of the roundtable intended to first establish the digital needs and requirements of humanities researchers by means of reviewing survey results, listening to the user case of a digital humanities researcher, and subsequent discussion between participants. After a short break five individuals representing global and pan-European online humanities (research) infrastructures [the Biodiversity Heritage Library (BHL), Ariadne, Europeana, the Digital Research Infrastructure for the Arts and Humanities (DARIAH) and the European Research Infrastructure for Heritage Science (E-

[18] User is interested in all information about *Oxalis* in Europe - very few published data accessible at present. Two Virtual Access (VA) services give access to extracted data. User needs access to add more data to that (small) corpus to expand it. Opens content for that person but also makes it more accessible for wider community as well. This should be in form of a FAIR Digital Object. This kind of data is very useful also for people studying evolution, etc. Emphasise benefits to lateral communities, which are newly creatable as a consequence of the services provided.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

RIHS)] made presentations. Their presentations provided an introduction to their organization describing the content, function or services most relevant to humanities researcher needs. A short discussion followed.

This roundtable falls under WP9, Communication and Dissemination, examining communication, liaison, and networking with external actors, and Task 9.4, Link with Cultural Heritage, examining the potential for overlap of biodiversity collections with cultural heritage collections, and exploring the need for making the relationship more explicit by means of evaluating potential synergies.

3.7.2 Introduction

In 2015, a recommendation was made to the European Commission's (EC) Commissioner for Research, Science and Innovation that a key priority for the EU's research and innovation policy should be to sponsor initiatives and venues that "foster, harness, and leverage collaborative interdisciplinarity". The recommendation was part of a policy brief (Allmendinger 2015[19]) authored by a member of the EC's Research, innovation and science expert group (RISE), a high-level group of policy experts who advise the EC Commissioner for Research, Science and Innovation. The author's premise was multifaceted (1) scientific breakthroughs tend to take place at the border of or beyond traditional academic disciplines resulting in new interdisciplinary specialities that entail novel methods of using, producing and disseminating science, (2) there is a paradigm shift occurring in the public academic community from 'blue science' where relevance and application of cloistered research is not immediately apparent, to a more open, demand-driven, participatory paradigm that tackles real world problems and engages different audiences, research communities, stakeholders, and citizens, and (3) trans-, multi- and interdisciplinarity result in a more effective research organization committed to supporting and triggering innovation.

Closer to home, the DiSSCo Prepare Project Proposal contends that the proliferation of natural science research infrastructures and their diverse data mandates that infrastructures go farther than simply providing access to different data classes. A "holistic approach is now required" where cross-linked information from the nano to the galactic "effectively underpins the entire research lifecycle and provides open access to mass and precise data (Hardisty et al. 2013[20])" driving novel, integrative research. While this statement was written in the context of natural sciences, its premise is equally applicable to the cross-domain integration of natural sciences and humanities with socio-historical and temporal context completing the holistic picture.

With the value of interdisciplinarity having been established and advancements in technology and the e-science revolution enabling it, the question then is not whether digitally unifying science and humanities data resources is advisable or possible but whether it is practical and timely. Is the benefit sufficient to justify the cost? Who are the people needing digitally integrated natural science and humanities data resources . . . natural scientists supplementing collecting event data with critical detail, museum preparators needing an original drawing for accurate specimen reconstruction, art historians establishing the temporal and social context of a painting based on the species and age of

[19] Allmendinger, J., (2015). Quests for interdisciplinarity: A challenge for the ERA and Horizon 2020 (EUR 27370 EN). Publications Office of the European Union, Luxembourg.

[20] Hardisty, A. et al., (2013). A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology, 13, 1-23. doi: 10.1186/1472-6785-13-16

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

its wood frame, archaeologists needing dates and collecting localities of ancient stone tools? Is the volume of potential users significant enough to warrant investment or is this a niche population? If their numbers are small, are the needs of cross-domain researchers being satisfactorily met in other ways? Will the numbers of researchers using integrated science and humanities resources expand once the resources are provided? What kind of research is being done at this interface and how well do these research objectives align with the corporate mission or global priorities? Are all of these questions and discussions premature?

Unfortunately, the answers to these critical questions remain outside the scope of this task and roundtable, but the inability to articulate and quantify appropriate answers or solutions does not preclude the practical imperative to begin discussions exploring future potential. So, it is from a broad long term planning perspective that this task and roundtable was approached.

3.7.2.1 Objectives

Potential synergies between the natural science and cultural heritage domains was initially considered from a bi-directional perspective and covered all institutional levels. That is, benefits natural science researchers and museum personnel would derive from links with cultural heritage, and non-research benefits derived by the humanities domain using natural science resources such as using a collection object in an educational art exhibit, were initially considered. Due to time constraints, however, these synergies were excluded. Instead the task and roundtable remained specifically focused on the use of natural science collections, data and archives by humanities researchers in a research capacity.

The roundtable was convened to define the digital requirements of humanities researchers working at the interface of science and humanities needed to facilitate their use of museum, university, and botanic garden natural science collections, data and associated archives.

The primary goals were threefold:

1. To establish the digital needs of humanities researchers by means of
   a. reviewing the results of a survey distributed to humanities researchers,
   b. listening to a use case from a digital humanities researcher,
   c. subsequent discussion between participants.
2. To understand the content, function and services of and potential for future collaboration with global and pan-European online humanities platforms and research infrastructures (BHL, Europeana, Ariadne, Dariah, and E-RIHS) aided by presentations.
3. To discuss and come to a consensus about what the next steps should be.

3.7.2.2 Roundtable Set-up

In anticipation of the roundtable discussion, an online Google survey was designed and distributed to humanities researchers working at the interface of science and humanities to get a picture of their potential use of natural science collection objects, data, and archives.

Survey candidates were selected by means of referrals from colleagues and task team members, online research of museums and research centers hosting cross-domain disciplines, and via survey recipients forwarding the survey on to other colleagues working at the same interface. This may

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

have resulted in some humanities disciplines being underrepresented, however, it's not clear whether this is due to the unstructured selection process or the natural gravitation of the survey to disciplines most likely to use the resources. Response to the survey was acceptable considering most were distributed just before the Christmas holiday break and shortly thereafter and were returned under a tight timeline.

Roundtable candidates were similarly selected, i.e., from referrals of colleagues and task team members and online research. Representation at the roundtable by the humanities domain was generally satisfactory, however humanities researchers were underrepresented owing also to the short timeframe and holidays. Their input was potentially important because their first-hand knowledge and experience would have dynamically enriched the conversation, however, it is questionable whether this would have changed the findings in a substantial way.

3.7.3 Presentations

Altogether eight presentations were made by the following representatives in the following subjects:

Presentation	Presenter
Introduction to DiSSCo	**Dimitris Koureas** DiSSCo Coordinator
Results of humanities researcher surveys	**Tina Loo** ICEDIG Project Officer
Digital needs of humanities researchers	**Andreas Weber** University of Twente (NL), Assistant Professor Department of Science, Technology and Policy Studies (STePS)
Humanities platforms and research infrastructures	**Jane Smith** Vice-Chair, BHL Members' Council and Head, Library and Archives, Natural History Museum London **Marco de Niet** Chair, Europeana Pro Network Association Management Board, Member Europeana Governing Board, Division manager, Research & Education Services / Deputy Director, Leiden University Libraries

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Franco Niccolucci

Project Director, Ariadne-plus,

Scientific Coordinator, PIN Vast-lab (IT)

Tatja Scholte

E-RIHS Dutch National Coordinator,

Senior Researcher, Dutch Cultural Heritage Agency
(Rijksdienst voor het Cultureel Erfgoed, RCE)

Sally Chambers

DARIAH Belgian National Coordinator and Senior Management Team National Coordinators Committee,
Digital Humanities Research Coordinator, Ghent University Centre for Digital Humanities

3.7.3.1 Introduction to DiSSCo, Dimitris Koureas, DiSSCo

This roundtable is part of ICEDIG, a DiSSCo design study, and its governing task (9.4) investigates the link between natural science collections and humanities. DiSSCo was added to the ESFRI roadmap in 2018 and is committed to making natural science collections, data, and associated expertise available and easily accessible.

In the DiSSCo data paradigm, the collection object is at the center of all information derived from its study and that data will be seamlessly provided. The collection object is the basic container of information holding a diverse array of information, e.g., genomic, biochemical, morphologic, geographic, taxonomic, ecologic, and species interactions, making it a powerful scientific tool in its own right. However, in addition to traditional biodiversity data, other data can be extracted from associated labels and archives such as people, places, affiliations, illustrations and stories, around which historic societal context can be reconstructed, making it useful for disciplines outside of biodiversity sciences.

D. Koureas emphasized the importance of a cross-domain interface between natural science collections derived data and the humanities sector and that the scope of the roundtable would focus on the use of that data by humanities researchers. Nevertheless, it was hoped that roundtable presentations would shed some light on shared issues and possible solutions that could inform DiSSCo's development to

3.7.3.2 Results of Humanities Researcher Surveys, Tina Loo, ICEDIG

T. Loo presented the results of a survey distributed to humanities researchers working at the science / humanities interface designed to capture their potential use of natural science collections, data and associated archives. The survey results were presented in four parts (user characterization, user interface with a natural science research infrastructure, use cases, and summary) and were calculated based on a pool of 33 respondents. The presentation and results are appended (Appendix 4.5.3), but the essential findings are provided here.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

USER CHARACTERIZATION

- The respondent was primarily west European (80%) coming from universities and natural history museums (77%).
- Representation across disciplines came primarily from anthropologists (44%) and historians (41%).
- The most frequently used (research) infrastructures were BHL (64%) and Europeana (32%).
- 'Expanding research possibilities' (84%) was considered the primary benefit of digital access to an *integrated* natural science collection data and archive resource, and this access was considered 'most important' (44%) or 'important' (34%).

USER INTERFACE

- Access to natural science collection data and archives was generally not divided along domain-specific lines, that is, 44% of the respondents chose access via *either* an online cultural heritage resource portal (e.g., BHL, Europeana, Ariadne) *or* natural science collections portal (e.g. GBIF).
- The most frequently used search terms in digitally accessing a natural science collections data resource were scientific name (55%) and locality (35%), and most respondents (55%) preferred to download search results in list format (.csv, .xlsx) indicating subsequent data manipulation.
- The respondents were further asked to identify desired functionality that would make them more likely to use natural science collections and archives. Their free text responses were mapped to tangible achievable technological objectives required to fulfil these wishes, and the mapping was validated by DiSSCo's Data Architect, Sharif Islam. Data mobilization and semantic annotation topped the list at 62% each.
- Finally, the respondents were asked to provide use cases describing their use of natural science collection resources in their research including corresponding label data and archives. The use of label data and archives was first separately calculated.
    - About 50% or more of the respondents used more than half of the label data with collecting locality (92%) and scientific name (82%) being used the most. This result correlated with the respondents' preferred search terms.
    - More significantly, about 50% or more of the respondents identified using *all* listed museum archive resources (Appendix 4.5.3, Page 99) with the exception of audio and video. Field notebooks and diaries (90%), collection catalogues (84%) and accession books (78%) topped the list because they contain basic information about many specimens.

USE CASES

Approximately 43 use cases were grouped according to the similarity of their objectives, and one use case per category was provided for roundtable consideration.

1. Category: *Historic reconstruction of persons, objects, collections or events* (Appendix 4.5.3, Page 101).

   This use case involves a historian investigating an object's or subject's history, that is, its provenance and circumstances of the collecting event. To achieve that, all available sources of textual and graphic information including (publications, correspondence, notes, audio

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

and video) are needed. The *inferred*[21] resource meeting these needs is an integrated humanities and natural science research infrastructure including associated archives and semantically linked data.

2. Category: *Historic reconstruction and correlation with human influence* (Appendix 4.5.3, Page 102).

   This use case involves an anthropologist/archaeologist investigating long-term human-environment (human-animal) interactions wanting to understand the human impact on the use and spatial, temporal and cultural distribution of a species. To achieve this, the researcher must quantify the species' presence across sites, create chronological context, and compare morphology across individuals. The *inferred* resource meeting these needs is an integrated humanities and natural science research infrastructure including associated archives and semantically linked data.

3. Category: *Reference: Identification of species or species comparisons* (Appendix 4.5.3, Page 103).

   This use case involves an anthropologist / ethnoecologist needing to determine the species used in the construction of organic ethnographic objects. In order to do this (it is *inferred* that), one must use a reference collection of objects, online resources and associated archives and/or expertise to identify the species. These needs could be met through the use of an integrated humanities and natural science research infrastructure including associated archives and semantically linked data (inferred).

   The anthropologist's second use case involves the need to compare similar objects between collections. In order to do this (it is *inferred* that), one must have physical or online access to similar collections including data and photos. The need could be fulfilled by the use of an integrated humanities and natural science research infrastructure including associated archives and semantically linked data (inferred).

4. Category: *Biologic / paleontologic study* (Appendix 4.5.3, Page 104).

   The final use case involves an anthropologist / archaeologist / paleontologist studying human evolution who needs to use fossil bones and lithic collections in museum and university collections. In order to do this, he must access collection catalogues, digital pictures and drawings and any other useful information such as diaries, notes and letters. The *inferred* resource meeting these needs is an integrated humanities and natural science research infrastructure including associated archives and semantically linked data.

SURVEY CONCLUSION: Of the ~43 use cases received, it was concluded that the majority of needs could be met with an integrated science and humanities research infrastructure with associated archives. There were a handful of researchers having slightly more specific needs such as georeferenced localities, mapping capability, aDNA and isotopic analysis, but these were the exception. The survey sufficiently demonstrated need by humanities researchers working at the interface of science and humanities for the use of an integrated science and humanities data

[21] The use case asked for a free text response that was often left unanswered or vaguely or partially answered. Therefore when needed, user requirements were 'inferred' based on the stated objective. In the use case tables, Appendix 3, pp. 18-21, inferred text is preceded with a boldfaced 'I" in parentheses, **(I)**.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

resource. However, the survey did not ask respondents to prioritize multiple answers or investigate the frequency of use, nor was the survey's sample size predetermined or statistically analysed thereafter.

3.7.3.3 Digital Needs of Humanities Researchers, Andreas Weber, University of Twente, The Netherlands

Andreas Weber is an assistant professor in the department of Science, Technology and Policy Studies (STePS) at the University of Twente, NL. Most of his research and teaching examines the relationship between Science, Technology and Culture (=STC) from a long-term and global perspective. This includes research into how computational technologies can be used to increase access to and learn from biodiversity heritage collections gathered in colonial Indonesia.

A. Weber introduced his recent work with the *Making Sense of Illustrated Handwritten Archives* project, an effort to interpret the rich collection of 17,000 handwritten pages of notes and illustrations from the Dutch Natural Science Commission documenting the scientific exploration of the Indonesian Archipelago (1820- 1850), using a state-of-the-art machine learning handwriting system. Often considered too historical for biologists, and too biological for historians, this resource was never fully studied or interlinked making it difficult to find the correct combination of specimen, illustration and field notes. The effort became a Netherlands Organisation for Scientific Research (NWO) project whose objective was to create a searchable digital repository of the archive for historians, biologists and the general public.

Challenges faced in the text recognition effort included:

- the use of multiple languages,
- intertwined visual and textual elements,
- different authors and styles on the same page.

Other project challenges included taxonomic (nomenclatural) uncertainty when the queried scientific name did not match the text, and the use by historians of antiquated search terms.

The semantically annotated project prototype will be launched in November 2020 and makes a good starting point to understand humanities researcher needs with respect to their use of a natural science data resources. While linkages between specimens, field notes, drawings, and publications are important to both the science and humanities domains, humanities scholars are more interested in *how* an item has been collected or the political context of the collecting event. Weber made the following observations and suggestions.

- Sufficient links between specimens, literature, field notes and diaries, and drawings are weak or missing. Specimen links to textual and visual archives are a treasure trove for cultural heritage researchers.
- Data portals of natural science collections are often a difficult starting point because search terms, e.g., species names, persons, dates, are often too narrow to yield satisfactory results of more broadly asked questions like "Which bat species were collected and drawn in Java from 1820-1833?" For this question, an advanced search combining Order, collecting date and locality, paired with an additional option to search on 'hand-written' or 'hand-drawn' category would have been helpful.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- Optimize data discovery: Free text searches, searches for domain-specific parameters (e.g., search options for handwritten, hand-drawn water-colour, print, etc.) would be helpful. In annotation, think about socio-historical context, e.g., many specimens collected by the Global North come from the Global South due to deeply linked histories of colonialism and imperialism. Humanities researchers would prefer to have an extensive list of searchable metadata but keep the fields as broad as possible.
- Seamless taxonomic mapping of queried species name to appropriate specimen and associated data would be very helpful.
- Cultural heritage researchers will find their way as soon as natural history collections are digitally searchable in a convenient, attractive way.

3.7.3.4 Humanities platforms and research infrastructures, Jane Smith, BHL Members' Council, Vice-Chair

Jane Smith has served as the Head of Library and Archives at the Natural History Museum, London since September 2012, before which she served as the Head of Library Collections and Services (2006 – 2012).

J. Smith provided an introduction to BHL's 15-year history including its mission statement to inspire discovery and improve research methodology by collaboratively making biodiversity literature freely and openly available to the world as part of the global biodiversity community. BHL supports Open Science and by extension open culture through digitization, open access, data integration, linking library and archives to specimens, and user engagement. One of the major challenges facing BHL is that much of the biodiversity literature, published, and original material is available in only a few select libraries in the developed world (the *Taxonomic Impediment)*. The lack of access to the literature has been a major impediment to the efficiency of scientific research. BHL addresses this by combining science with libraries and technology, building a critical mass of data, then unlocking and connecting it to related object and literature.

To date, BHL has 58M pages, 253,000 titles, and 252,500 volumes (30% published) whose content includes:

- species descriptions
- distribution records
- climate records
- history of scientific discovery
- information on extinct species
- scientific observations
- scientific illustrations
- ecosystem profiles.

BHL provides tools and services to support access and use via:

- API and data exports
- custom pdf downloads
- taxonomic name searching
- article indexing
- reference management tools

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- DOI assignment
- full text search
- a transcription tool to access and data mine original sources.

BHL collaboration:

- Collaborates and contributes content to: GBIF, CETAF, EOL, Europeana, TDWG, Digital Public Libraries of America (DPLA), ICSTI.
- Collaborators share a common digitization strategy regarding out-of-copyright and in-copyright material and share standards and practices. BHL respects copyright but open access is actively encouraged. Content will remain natural history focused but scope will broaden to support use by the arts and humanities. A few examples of BHL application in the arts, humanities, and social sciences include . . . its use by a glass artist, virtual reconstruction of surviving books from Charles Darwin's library, and a BHL blog about the making and communication of science in both the Victorian period and today. Feedback from interdisciplinary users is important for development.
- Collaboration with users is critical and focuses on scientists (taxonomists and systematists) but is expanding to include a wider audience in other research disciplines and the public. (User stats: 7M+ users, 118,000+ monthly users / 212,000+ monthly visits from 243 countries and territories.)

In summary, BHL expansion, collaboration across science, libraries and technology are important objectives, but the library will primarily remain focused, open, linked, global and committed to user engagement.

3.7.3.5 Humanities platforms and research infrastructures, Franco Niccolucci, Ariadne-plus, Project Director

Franco Niccolucci is the director of VAST-LAB (a research and development lab involved in the definition and implementation of new technologies for Cultural Heritage) at PIN in Prato, Italy. Prof Niccolucci has coordinated several EU-funded projects on the applications of information technology to archaeology (PARTHENOS, ARIADNE, CREATIVE CH, COINS, CHIRON and 3D-ICONS) and is currently the coordinator of ARIADNEplus, a research infrastructure on archaeological data.

Ariadne-plus (2019-2022) is the successor to Ariadne (2013-2017) whose mandate is to integrate European (+ Israel, Argentina, Japan, USA) archaeological datasets in steps, first with metadata and then by item. The 2M datasets currently available are primarily derived from emergency excavations that are catalogued in a searchable registry including images, drawings, maps, and videos. The registry includes all large existing data repositories, 60-70% of which are at a national level, plus those from other excavations or museums (2%). Cloud services are provided to Ariadne's 11,000 users and its standardized vocabulary is multilingual. One third of Europe's 60,000 computer-savvy archaeologists have used the registry at least once. Integration and search are based on a common ontology that has been extended to accommodate subdisciplines. Ariadne semantics are based on an international cultural heritage conceptual reference model, CIDOC-CRM. Partners provide standardized metadata about their datasets to the registry that is ingested and further standardized with respect to terms, periods and places. The registry is searchable by where-when-what and keywords. An update of the current portal is scheduled for the summer of 2020 and cloud-based processing services will be implemented in 2021 including a multilingual text mining tool that will

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

recognize names, entities and their key relationships. Ariadne ensures data quality through FAIRification and repository certification. It further provides DOIs, training, and support for repository creation and restructuring. Ariadne-plus is an EC H2020 funded program.

3.7.3.6 Humanities platforms and research infrastructures, Marco de Niet, Europeana Pro, Europeana Network Association (ENA) Management Board, Chair

In September 2017 Marco de Niet moved to Leiden University Libraries (UBL), to become responsible for all physical and digital services provided by Leiden University's library to the academic community and beyond. He has been involved with Europeana from its inception including formative collaborations while working at the Dutch national library until 2004.

M. de Niet provided an introduction to Europeana including its stated mission to transform the world with culture by making it easier for people to use cultural heritage for work, learning or pleasure. Its primary 2020-2025 strategy is to support the digital transformation of Europe's cultural heritage sector. Coincidentally, Europeana's largest collection is its natural history collection.

In contrast to BHL, Europeana has an unfocused approach to accommodate its many domains that are divided into topics and themes. The diverse approach ensures that Europeana finds common ground between its institutions, however, this makes developing a data model more of a challenge. The organization's data model and licensing framework requires all metadata to be published as public domain (CC0), and all digital objects must carry a copyright status. The framework further identifies tiered levels of content each corresponding to a specific use, reproduction quality, annotation and re-usability characteristics, and application benefits. Europeana has chosen to delete over 10,000 (20%+) of its current content because it does not meet criteria for its lowest content level.

Europeana has a three-pronged organizational structure (Europeana Foundation, Europeana Network Association (ENA), and the Europeana Aggregators Forum). The partitioning of functionality ensures Europeana's future when Foundation management, the operator of its platform, changes.

M. de Niet discussed Europeana's Impact Assessment Playbook, a guide developed to take museums, libraries, archives and galleries through the first phase of an impact assessment (future phases to be published), and suggested it could be beneficial to use Playbook tools to assess the value of integrating natural science and cultural heritage collections. He then discussed an ENA Researcher Needs Task Force that analysed 31 researcher needs reports produced by other humanities infrastructures *and* the results of an October 2019 survey (377 respondents from 37 countries) identifying research objectives, access, problems, skills and training, and awareness about the European Open Science Cloud.

Among the task force findings most relevant to the roundtable were:

- Tools, content and research infrastructures should cater to the increasingly multi- and interdisciplinary character of research.
- Partnerships and coordination actions should be amplified to maximize the impact of DHC and co-create new innovative services as part of the Open Science Movement
- Promote FAIRness of data through guidelines and trainings.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.7.3.7 Humanities platforms and research infrastructures, Tatja Scholte, E-RIHS NL, National Coordinator

Tatja Scholte is the Dutch National Coordinator for the European Research Infrastructure for Heritage Science (E-RIHS) and currently programme manager of 'Modern and Contemporary Heritage' at the Dutch Cultural Heritage Agency (Rijksdienst Cultureel Erfgoed).

T. Scholte provided an introduction to E-RIHS including its primary mission to support research on interpretation, preservation, documentation and management of cultural heritage. E-RIHS's four domains of heritage science include:

- Collections of art, material cultural and natural heritage
- Archaeology
- Built heritage
- Archives

She described E-RIHS' predecessor initiatives, its ERIC governance structure and provided a list of E-RIHS' partners. She further explained E-RIHS NL's governance, reviewed a list of interested Dutch parties, and explained E-RIHS' four research instruments. The instruments facilitate access to facilities, instrumentation, and physical and online access to heritage collections and data. Finally, she reviewed E-RIHS' 2020 agenda and noted that E-RIHS NL had applied to be included in the Dutch National Organization for Scientific Research's (NWO's) national roadmap.

3.7.3.8 Humanities platforms and research infrastructures, Sally Chambers, Dariah EU, National Coordinators Committee, Chair

Sally Chambers is Digital Humanities Research Coordinator at Ghent University, where she coordinates the day-to-day activities of the Ghent Centre for Digital Humanities and Belgian participation in DARIAH. From 2011-2015, Sally was Secretary-General for DARIAH-EU Coordination Office and prior to that, worked for The European Library, focusing on interoperability, metadata and technical project coordination.

S. Chambers provided an introduction to Dariah.

- Dariah is a Digital Research Infrastructure for the Arts and Humanities
- Enables excellent research in the Arts and Humanities by exchanging and sustaining tools, services, data and knowledge from its member countries and facilitating the wide uptake of digital methods.

She further reviewed key elements of its latest strategic plan released in August 2019.

- Creating: Build a Marketplace to facilitate fluid exchange of tools, services, data and knowledge.
- Transforming: Build access to education and training.
- Connecting: Build working groups, hubs and other forms of transnational and transdisciplinary organization enabling researchers to work together.
- Complementing: Build bridges between research policy and communities of practice.

She identified the activities and services contributed to DARIAH by member countries (coordination, access, expertise, interoperability, hosting content, tools and software, training, summer schools, and events), and organizational elements of The Marketplace where high-quality locally produced tools are shared at an international level. Dariah is distinct from Clariah in that Dariah-BE provides a

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

portfolio of services enabling digital scholarship in the Arts and Humanities in Belgium and beyond, whereas Clariah-VL's objective is to embed high-quality tools and resources into the workflows of humanities researchers and pave the way for Flemish participation in the European Open Science Cloud.

She summarized Dariah's training and education services and resources, working groups, and presented a graphic identifying all the cross-domain entities with which Dariah collaborates including ICEDIG and DiSSCo. She introduced the social sciences and humanities contribution to the European Open Science Cloud (EOSC) called SSHOC (the Social Sciences and Humanities Open Cloud), its composition, timeline, budget, funding and dedication to FAIR principles. She reiterated from the early afternoon's discussion that through the EOSC there will be more collaboration that could be a means for future data exchange.

She then reviewed several relevant humanities initiatives: Dialogue, connects the computational, cultural heritage and digital humanities communities; Always Already Computational, an effort to share current and potential approaches to developing cultural heritage collections that support computationally-driven research and teaching; and the Heritage Data Reuse Charter that aims to make cultural heritage data easier to access, more sustainable, reproducible and citable.

Finally, she identified potential synergies between digital humanities and the cultural and natural heritage sectors including the shared trend toward Open Science (c.f. SSHOC) and opportunities for collaborative efforts in training and education related to digital technologies. She identified data level access, i.e., access to digitized content from closed digital silos, as one of the remaining challenges.

3.7.4 RT Discussions

3.7.4.1 Definition and Prioritization of Needs

The first round of discussions was intended to better define and prioritize humanities researcher needs, however, it reverted to more fundamental themes and questions regarding the need and cost-benefit of integrating natural science and cultural heritage resources. The discussion was facilitated by A. Casino and the following points were made:

NEED

- Need: Linking natural and cultural heritage data resources is key to an overall picture of describing collections. Identifying humanities researcher needs will inform DiSSCo planning and development and has the potential to similarly inform the development of other research infrastructures, e.g. BHL content is growing in cultural richness and wants to support research across disciplines.
- Need, chicken-egg: The demand for an integrated natural science and humanities data resource is presumed to be a niche market now but could expand as capabilities and services are provided.
- Need, pluck low hanging fruit: What are the respective domain objectives and what is already available to meet those needs? What is expected of the research communities, and what do researchers want to do with integrated resources? It seems prudent to start by undertaking activities that are doable and useful . . . to grab the so-called 'low hanging fruit'. Need a long-term program.

CONCERNS

- Timing: Are discussions regarding the potential collaboration of natural and cultural heritage domains premature? Most research infrastructures are developmentally emerging themselves.
- Cost / Benefit: Is the demand for integrated natural science and humanities data resources and the precedence that it could assume over core mission objectives sufficient to justify the investment? And how can the demand be quantified to justify inclusion?
- Better understanding needed: Currently there are gaps in terms of (1) information availability and thus what gets mobilized, (2) interlinking that information, and (3) services provided to make the data available and actionable. Stakeholders and services change with this value chain so some services may need to be allocated outside of the cultural and humanities research infrastructures. Need better understanding of the landscape.

DEVELOPMENT

- Collaboration: While both sides agree that collaboration is required to formulate a long-term holistic response to global challenges that meets the needs of both domains, it is unclear which domain is most appropriate to initiate and drive the collaboration forward. Because the two domains are so different in terms of language, methodology and terminology, E-RIHS or Dariah might be the appropriate intermediary, however, E-RIHS is not currently prepared to facilitate this level of discussion. Possible opportunities for collaboration could arise under the interdisciplinary emphasis of Horizon Europe or in relation to global data accessibility of the European Open Science Cloud.
- Open data standards and quality assurance: We should not worry about interlinking infrastructures at this point, however, collaborating early on (meta)data standards and interoperability will ensure the possibility of establishing needed links in the future. Developing standards will require a long-term community investment. Rather than starting with ontologies, collaboration can be initiated by first finding common ground between the two diverse domains. However, DiSSCo needs to ensure data quality in addition to agreeing upon standards.
- Humanities research approach: Humanities researcher is looking for context. They are not focused on individual specimens or objects but are interested in reconstructing *relationships* of objects in a social or historical context. In an example from a survey use case, the research objective was to determine an object's provenance, that is how it had moved through owners and locations. This information is compiled by natural science institutions as part of their collection management system, but is not relevant to natural science research and thus not included in DiSSCo. Whether this type of information should be included in DiSSCo to accommodate the humanities domain requires further investigation of demand, cost and benefit.
- Scientific names: Although the user survey identifies scientific name as being the most frequently used search term, using it in a query can generate uncertainty for the taxonomically uninitiated because the queried name may not correspond with search results. The researcher is uncertain as to whether their results reflect the correct species, whether the name has been changed, is a synonym, or is invalid, etc., requiring further

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

taxonomic investigation. In this regard, it is incumbent upon the natural science collection research infrastructure to provide a seamless user experience by translating and mapping the search query to user friendly, reliable results. However, using scientific names as an anchor to link data remains problematic due to its inherent variability and instability.

3.7.4.2 Interface and Coordination

The final round of discussion occurred after five introductory presentations by online humanities platforms (BHL, Ariadne, Europeana) and research infrastructures (E-RIHS, Dariah) and was intended to plan for or address policy, procedural or technological considerations or action items resulting from the conceptual overlap of humanities researcher needs with tools, services and the infrastructure designed to support them. The discussion was facilitated by D. Koureas and was very limited in time.

D. Koureas began by suggesting that these preliminary discussions could use better follow-up and continuity. In similar situations the tendency has been to put them on a shelf without further intervention. He further observed that while Europeana is not really a vehicle for natural science research, its volume and diversity of data types was worth building upon. The question remains . . . how? Thematic data sets of Europeana could be a vehicle for integration with natural science, however, this would require that research infrastructures provide usable data for Europeana, and it was further suggested that EOSC might be the means for that. The flaw in this proposal, however, is that the cluster projects working towards FAIR data standards and the EOSC are siloed in thematic areas whose interactions are driven by thematic labelling and not need.

The current route of natural history content to Europeana is via institutions and in the future, it might be better coming from the research infrastructures because of convening power. This was countered by the observation that research infrastructures might not cover the necessary data and are governmentally driven as opposed to the institutions that create the data and have a wider scope.

Is there a next step in the potential for integration of natural science and humanities collections data or is this a one-off discussion? What is the mechanism to continue discussions? Suggestions and observations were:

- FAIR could be common ground to continue.
- The effort requires someone driving the conversation, and also higher level executive strategic participation.
- Interested individuals could join the research community to monitor status.
- DiSSCo is very receptive to having more input from humanities experts but the desire must be reciprocated by the humanities community. DiSSCo will be provided with a list of potential consultants.
- The conversation has to continue at different levels: strategic, via Dariah, DiSSCo General Assembly, and with national coordinators. The history of science or history of natural science could be a good entrance point.

3.7.5 Conclusions and Next Steps

Examining potential synergies between the natural science collection and humanities domains is an inevitable outgrowth of the trend towards holistic interdisciplinary research and the indisputable

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

value it adds in terms of cultivating innovation through cross-fertilization of research outlooks, routines and paradigms and expanding research possibilities. So the Task objective to identify the synergies humanities researchers have with natural science collections, data and archives is the first step in realizing this potential. The survey demonstrated the need for integrated data resources in the population of researchers working at the science-humanities interface, but more fundamental questions regarding the size, significance and prioritization of this demand remained unanswered and requires further exploration before planning can occur.

Takeaways from the presentations and discussions are identified below.

Humanities Researchers

- Need: The survey concluded that the majority of identified needs from 43 humanities researcher use cases could be met with an integrated science and humanities research infrastructure with associated archives. The demand is presumed to be a niche market now but could expand as more resources are provided.
- Characterization:
    - The disciplines most likely to benefit from integrated resources are anthropologists and historians who primarily use BHL and Europeana and believe that integrated access is very important and expands their research possibilities.
    - Humanities researchers are looking for social or historical context, i.e., not focused on the individual object per se but the object's relation to other objects, people, places, events, cultures, zeitgeist, etc., from data derived from the object's label, associated documentation and graphic files.
- Interface:
    - Optimize data discovery with fuzzy, augmented, and free text searches, and by providing more domain-specific search options, e.g. 'hand-drawn', 'handwritten', etc.
    - Scientific name is the preferred search term but can lead to taxonomic confusion and require more investigation when the searched for name does not correspond with results. It is incumbent upon the research infrastructure to provide these results seamlessly in a user friendly, unambiguous way, however the inherent variability and instability of scientific names makes this potentially difficult.
    - Sufficient links between specimens, literature, field notes and diaries, and drawings are weak or missing. Specimen links to textual and visual archives are a treasure trove for cultural heritage researchers.

Humanities Platforms and Research Infrastructures

- Humanities platforms and research infrastructures expressed their organization's desire to broaden content, tools and services and become more interdisciplinary (BHL, Europeana, Dariah) while keeping core mission values intact. However, copyright issues restrict full FAIRification.
- Europeana has made concerted efforts to maintain the quality of its content and is committed to strengthening its impact by expanding partnerships and collaborations as part of Open Science and will promote FAIR data.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- Humanities research infrastructures (Dariah and E-RIHS) have potential opportunities for collaboration with natural sciences in the trend towards Open Science (SSHOC), and in training and education related to digital technologies. However, access to digitized content from closed digital silos remains a challenge.

Next Steps

- The need to quantify the demand for integrated resources and have a means for justifying its investment and prioritization was a recurring concern. A more thorough understanding of the value chain from mobilizing data to providing services is needed.
- Future collaboration is mutually desired.
  - Someone needs to drive the conversation forward although it's unclear from which domain.
  - Collaboration needs higher level strategy and should include Dariah, DiSSCo General Assembly, and national coordinators.
  - Possible opportunities for collaboration could arise under the auspices of DiSSCo's Stakeholders Forum, Horizon Europe's interdisciplinary emphasis, EOSC's global data accessibility, open data standards and FAIRification, or training and education in digital technologies. Providing for cross domain collaboration in the EU's thematically driven cluster groups would go a long way towards facilitating interoperability across disciplines.
  - Early collaboration is important in developing the (meta)data standards and interoperability that will ensure the possibility of establishing future links. Data quality must also be ensured.
  - Start small. Collaboration can be initiated by first finding common ground between the two diverse domains. The history of science or history of natural science could be a good starting point.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

4 Appendices

4.1 Appendix Roundtable Collection Digitisation Dashboards

4.1.1 Participant List of the Roundtable

General

- Ana Casino (CETAF)
- Agnes Wijers (Picturae)
- Myriam van Walsum (Picturae)
- Jeroen Bloothoofd (Picturae)
- Luc Willemse (Naturalis) – overall chair
- Emily van Egmond (Naturalis) – taking minutes
- Olaf Banki (Naturalis) – chair subgroup 1
- Wouter Addink (Naturalis) – chair subgroup 1
- Letty Stupers (Naturalis) – taking minutes

Subgroup 1- End users and parameters

- Niels Raes (Naturalis)
- Gwenaël Le Bras (NMNH)
- Jeremy Miller (Naturalis)
- Deborah Paul (IDigBio)
- Pierre-Yves gagnier
- Jaume Piera (chair working group ECSA)
- Jeroen Bloothoofd (Picturae)
- Luc Willemse (Naturalis)

Subgroup 2- Technical aspects and unifying data

- Hannu Saarenmaa (University of Helsinki)
- Andrea Hahn (GBIF)
- Elspeth Haston (RBGE)
- Dominik Röpert (BGBM)
- Robert Tiessen (Picturae)
- Simon Chagnoux (NMNH)
- Matt Woodburn (NHM)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

4.1.2 User Groups and Data needs Overview

Table 1. Overview of the user groups and their expected need for each level of data of natural history collections. It is also indicated which user groups were represented among the Roundtable participants.

	Collection level	Storage Unit level	Species level	Specimen level	User categories present at this meeting
Research				x	x
Collection		x	x		x
IT	x			x	x
Governmental	x			x	
Non-governmental	x				x
Education				x	x
Industry	x				x
Media				x	
Institution	x	x	x	x	x
Citizen science	x		x	x	x

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

4.1.3 All Collected User Stories Related to the CDD During the Roundtable

User groups	As a	I want to	So that	For this I need (data elements)	Level of digitization	Digitized/non-digitized
Media	Journalist	Link to primary source data (scientific literature, museum collections databases etc.)	My readers can learn more about the topic of an article	Collections database records	Specimen	Digitized
Governmental	Policy maker	Information on the distribution of species under the nature directives	Assess conservation status and distribution range	Detailed distribution data	Specimen	Digitized
Collection	Collection Manager	Check in which institutions certain collection categories are kept so that I can forward a collection on offer to an institute that is interested	I can forward this information to a collection holder	Details about taxonomic/geographic specialism and possibly wish lists for certain specimens	Storage/species	Digitized
Institution	Director	Hire a curator with knowledge	I can be sure they have a background that	Collection types, importance of collection	Collection	Both

		of specific groups	includes knowledge of the main collection	gauged by size, scope, and time period		
Citizen science	Citizen scientist	Know where was a certain collector on a certain day	To help transcribe a specimen	Existing transcription of specimens collected around the same time by the same collector	Specimen	Digitised
Industry	Solution provider	Build and provide solutions and related services	The keepers and scientists can work better and easier with their collections for less cost	Volumes, locations and physical sizes plus an insight on what is digitally represented and what not. Even better would be if there is an institutions priority as to what needs to be digital first	Collection and partly storage level	Both
Research	Scientist	Model South East Asian biodiversity patterns	To gain an answer to a scientific question	Detailed taxonomic and geographic information	Specimens	Which institutes hold the largest non-digitized collection
Non-governmental	Association	To gather information to have overall figures representative of partners' state-of-the-art	We can showcase the relevance of collections to policy makers and attract funds	High-level figures that feature the collections as a whole	Collections	Both, digitized and non-digitized information are valuable (to indicate the progress and the support needed, respectively)

Research	Scientist	Query when and where one or more species have been recorded, and their characteristics, and the institutions that archive specimens	I can collect more specimens, or borrow collections	Taxonomic fields, geographic coordinates, date of collection	Specimen	Digitized
IT	Software developer	Create new usages with the data and ways to add to the data, through apps or web interaction	Data is more accessible to the masses and different collections can be, for instance, cross-referenced. At the same time additional data can be added and fed back into the core databases. Geographic location will be involved as every man has GPS access today. The vantage point to access these 'big data' sources could be educational, entertaining, medical, historical and natural sciences	Scope: Collection level, details: Specimen level	Specimen	Digitized

Citizen science	Citizen scientist	Help with transcribing	I can enjoy this voluntary work	Images without transcription	Specimen	Partly digitized
Governmental	Policy maker	Know the use of the collections by other domains as a key indicator of its impact	I can distribute resources and allocate them in alignment to the strategic priorities of the government that I represent	Access to the collections, virtually and physically, from different types of users	Collection	Both, digitized (publicly available) and non-digitized (to understand the need to bridge the gap)
Education	Curious person	Learn about the species that might be in my environment	I can improve my bioliteracy	Taxonomic fields, common names, geographic coordinates, species characteristics, images	Specimen	Digitized
Citizen science	Citizen scientist	Be recognized as contributor	I can apply for funding to digitize my own collections	Contribution indicators	Could be at all levels	Digitized
Institution	Director/administrator	Know what makes our collections unique	I can effectively advertise/highlight the collections to improve usage	Collection types, with size, locality scope, time, taxonomic scope, important collectors	All levels	Both
Collection	Collection Manager	Start a digitizing project	I like to digitize a certain group of my collection, I like to do this internationally because of funding	Know where else there are collections of this group	All levels	Digitized
Citizen science	Citizen scientist	Be recognized as contributor	I can identify my contribution on	Contribution indicators (as validator)	Specimen	Digitized

			validating data from external sources			
IT	Solution provider	Tap into the vast market of digital storage solutions for digital natural collections	I can sell my services and consult	Predictable numbers on collection type, volume and progress in digitization	Collection	Both
Collection	Collection manager	Redirect a researcher to colleagues	They can examine more collections	I need to know which institute holds specific kinds of collection	Species	Digitized
Institution	Collection manager, Director, Administrator	Know the situation with collection sizes	I can plan for new space/storage needs	I need to know existing sizes of collections, and the number of new material coming in. Also, need to know status/condition (e.g. wet, dry) of existing material. Also collection health information.	Collection, species	Both
IT	Automatic identification systems developer	Which collections are available to use as a reference (training data set)	I can training my algorithms for automatic identification	Collections of target species (validated)	Collection, species	Digitized
Citizen science	CS site manager	Select a load of images	To build a CS project	Basic elements on the images	Specimens	Partly digitised (images + OCR

						results, other projects result)

4.1.4 Impressions of the Roundtable

Figure 1. All attendees of the Roundtable at the All-Hands meeting in Leiden, the Netherlands during the introductory presentation.

Figure 2. Luc Willemse giving an introduction to the Roundtable at the All-Hands meeting in Leiden, the Netherlands.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Figure 3. Olaf Banki chairing the meeting of sub group 1 on user groups and stories related to the CDD.

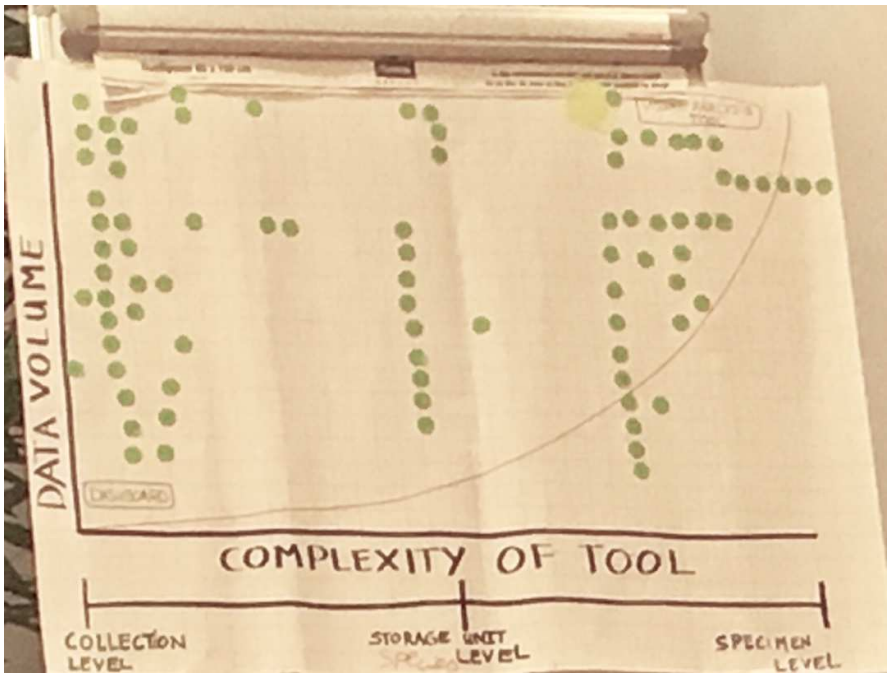


Figure 4. Detail of some of the outcomes of the meeting of subgroup 1 on user groups and the data required to fit their needs.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

4.2 Appendix Roundtable Analogue 2 Digital

4.2.1 Pictures of the Event

Figure 1. Sarah Phillips at the ICEDIG Roundtable at the at Joint Annual meeting of SPNHC and TDWG 2018

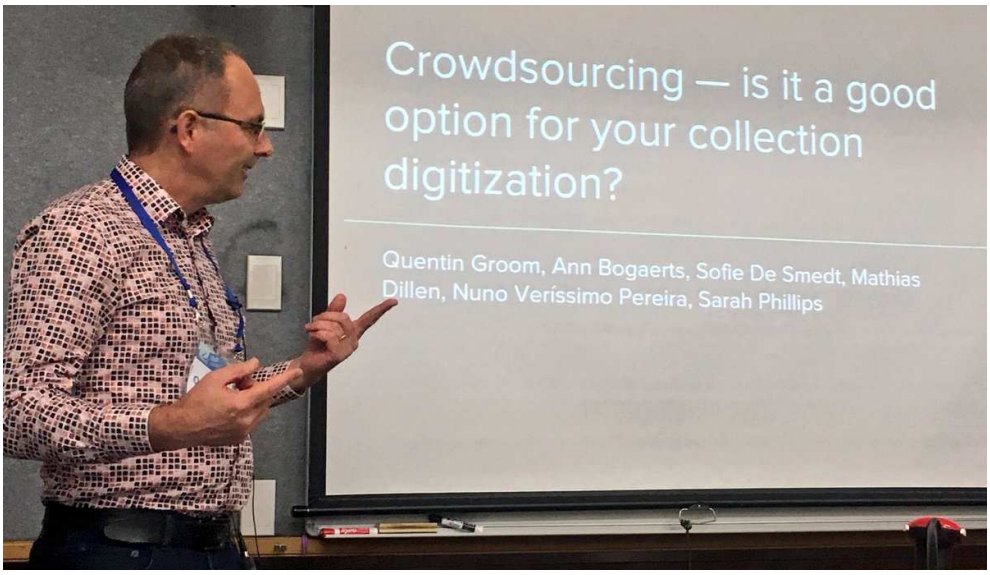


Figure 2. Quentin Groom at the ICEDIG Roundtable at the at Joint Annual meeting of SPNHC and TDWG 2018

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

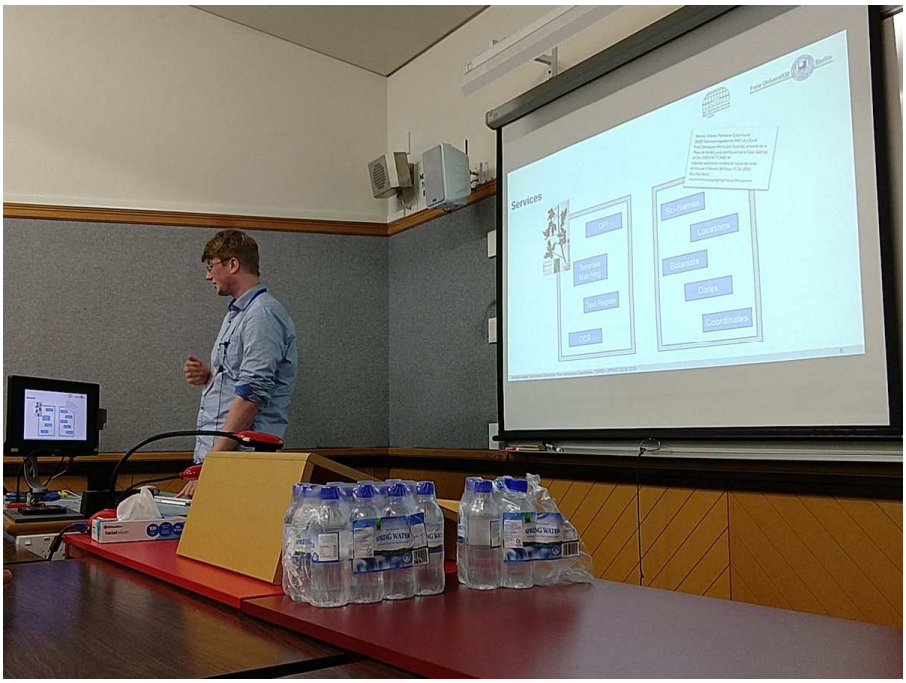


Figure 3. Fabian Reimeier at the ICEDIG Roundtable at the at Joint Annual meeting of SPNHC and TDWG 2018.

Figure 4. Henry Engledow at the ICEDIG Roundtable at the at Joint Annual meeting of SPNHC and TDWG 2018.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

4.3 Appendix Roundtable Future of warehousing and the use of robotics for NH collection

4.3.1 List of Participants

Andy Appleyard, British Library, Head of Operations North
Alison Selina, British Library, Location manager
Maarten Taborsky, Bruns, Project Director
Patrick Vermeire, Bruns, Commercial Director
Luc Willemse, Naturalis, Head entomological collections
Steven van der Mije, Naturalis, Project Leader Activities
Abraham Nieva De La Hidalga, University of Cardiff
Helen Hardy NHM Digital Collections Programme Manager (Chair)
Ben Price, NHM, Senior Curator in Charge Entomology, Insects
Mark Carine, NHM, Principle Curator in Charge Algae, Fungi and Plants
Laurence Livermore, NHM, Project manager, DCP Innovation
Clare Valentine, NHM Collections Leader
Jacqueline Mackenzie-Dodds, NHM, Molecular Collections Facility Manager
Naomi, Cocks, NHM Project Coordinator and Analyst
Giles Miller, NHM, Principal Curator, Micropalaeontology and SCIC EE
Myriam van Walsum, Picturae
Agnes Wijers, Picturae, Projectleader ICEDIG WP3

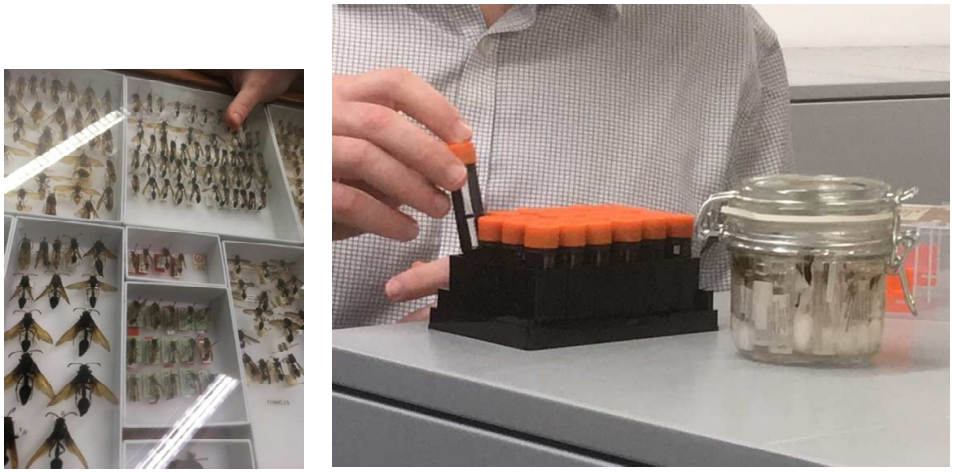
4.3.2 Impressions of Roundtable Discussions and Tours

Collection tour – Sir Hans Sloane's herbarium volumes

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Collection tour – Ben Price, senior curator in charge for Entomology

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483



Dry and wet collections in entomology

Jacqueline MacKenzie Dodds at the molecular collection

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Hamilton micro lab star

4.4 Appendix Roundtable Museum Specimen and Molecular Data Linkage

4.4.1 Participant list

1. Dimitris Koureas (DiSSCo/Naturalis)
2. Wouter Addink (DiSSCo/Naturalis)
3. Sharif Islam (DiSSCo/Naturalis)
4. Alex Hardisty (DiSSCo/Cardiff University, remote)
5. Jerry Lanfear (ELIXIR)
6. Corinne Martin (ELIXIR, remote)
7. Guy Cochrane (EMBL-EBI)
8. Quentin Groom (Bioschema/TDWG)
9. Ana Casino (CETAF, second day only)
10. Karsten Gödderz (CETAF, second day only)
11. Rob Hooft (ELIXIR-NL/DTL, first day only)
12. Dmitry Schigel (GBIF)
13. Hilary Goodson (NHM London/DiSSCo UK)
14. Donat Agosti (Plazi, second day only, remote)
15. Andy Smith (ELIXIR, second day only, remote)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

4.4.2 Pictures

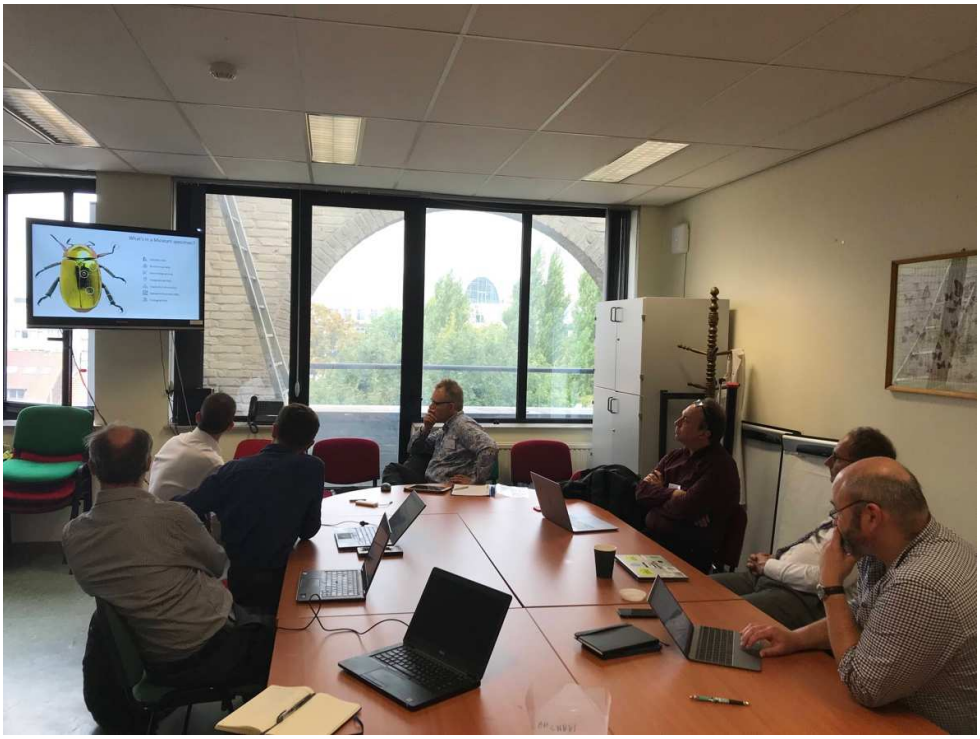


Figure 4: Dimitris Koureas presenting DiSSCo.

Figure 5: Group brainstorming on how to use DwC triplets in ENA database to generate digital objects the DiSSCo global specimen catalog.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

4.5 Appendix Roundtable Humanities Researcher Synergies with Natural Science Collections and Archives

4.5.1 Participants

Dimitris Koureas, **DiSSCo** Coordinator

Ana Casino, **DiSSCo** Deputy Coordinator for Communication and Engagement (CETAF, BE)

Sharif Islam, **DiSSCo** Data Architect

Karsten Gödderz, **ICEDIG** D9.2, CETAF Task Partner

Tina Loo, **ICEDIG** Project Officer

Agnes Wijers, **ICEDIG** D9.2, Picturae Task Partner

Luc Willemse, **Naturalis**, ICEDIG Liaison

Andreas Weber, **University of Twente**, NL, Assistant Professor, Department of Science, Technology and Policy Studies

Jane Smith, Vice-Chair, **BHL** Members' Council

Franco Niccolucci, PIN - **Ariadne-plus** Coordinator

Marco De Niet, Chair, **Europeana Pro** Network Association Management Board

Tatja Scholte, **E-RIHS** Dutch National Coordinator

Sally Chambers, **Dariah**, Belgian National Coordinator and Ghent University, Digital Humanities Research Coordinator

4.5.2 Impressions

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Jane Smith, Vice Chair of the Biodiversity Heritage Library's (BHL) Members' Council provides roundtable participants with an introduction to BHL.

Roundtable participants listen to presentation of Jane Smith, (left front to back) Karsten Gödderz (phantom photographer), Agnes Wijers, Dimitris Koureas, Ana Casino, Luc Willemse, Sharif Islam, Sally Chambers, Marco de Niet, Tatja Scholte, Andreas Weber, Franco Niccolucci, and Tina Loo.

4.5.3 Suvery Results

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Survey

- Scope: Identify potential synergies that cultural heritage researchers working at the interface of science and humanities have with natural science museum, university or botanic garden collection objects, data and associated archives.
- "What resources you would use and how you might use them (use cases), and the added benefit you would obtain. You may assume that these resources are free, readily discoverable and digitally accessible."
- Almost all multiple choice, but always with an 'Other' option.
- Option to answer/not answer a question, and select as many answers as applicable.
- Percent results calculated relative to number of individuals responding to question.
- Selection of survey candidates was based on online research, recommendations from team members, colleagues, or forwarded on by survey recipients themselves to their colleagues.

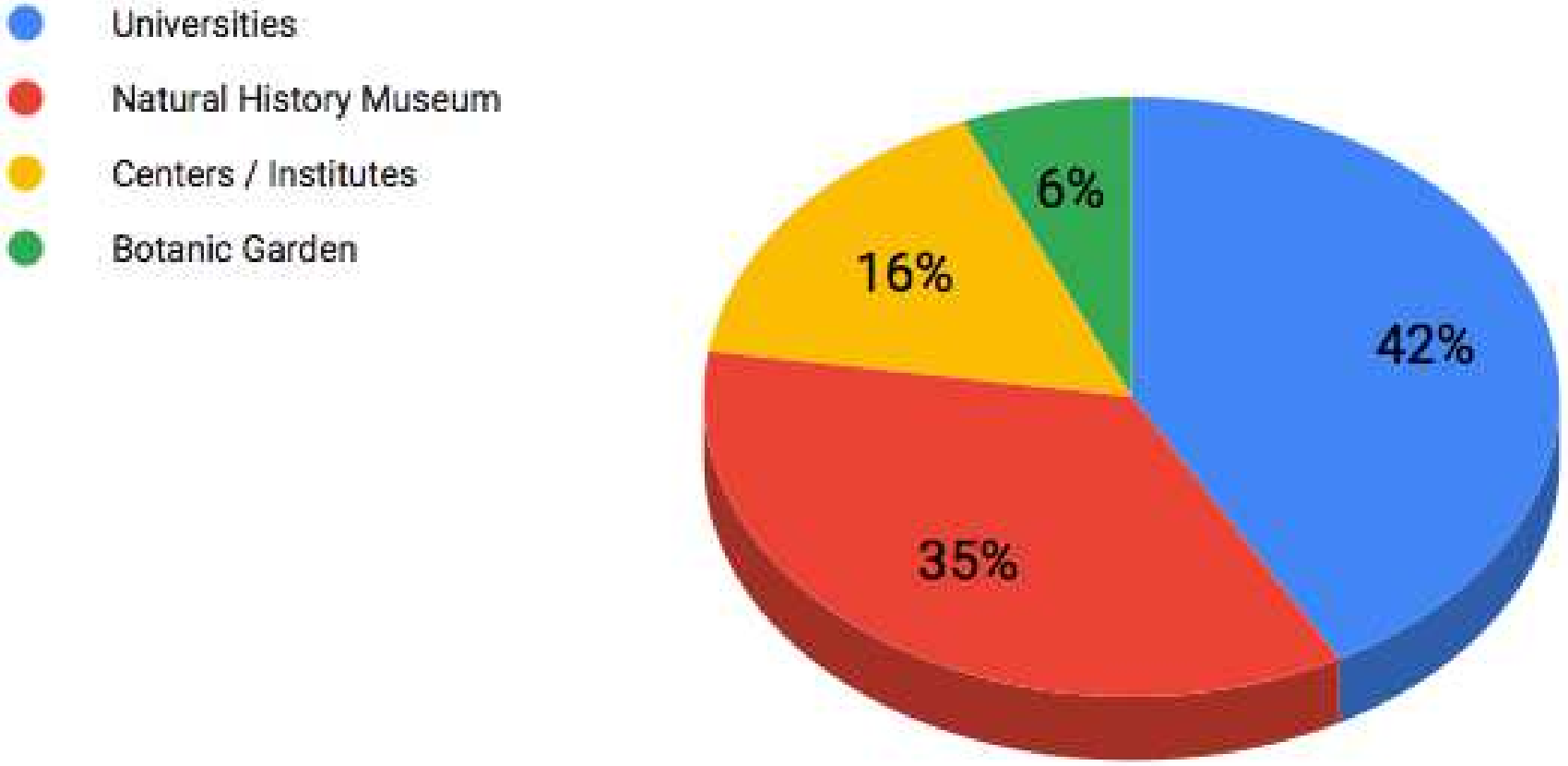
Humanities Researcher Survey

31 response
11 countries

Netherlands	19%
Germany	16%
Austria	13%
Spain	10%
UK	10%
USA	10%
France	6%
Italy	6%
Brazil	3%
Latvia	3%
Poland	3%

31 responses
11 countries

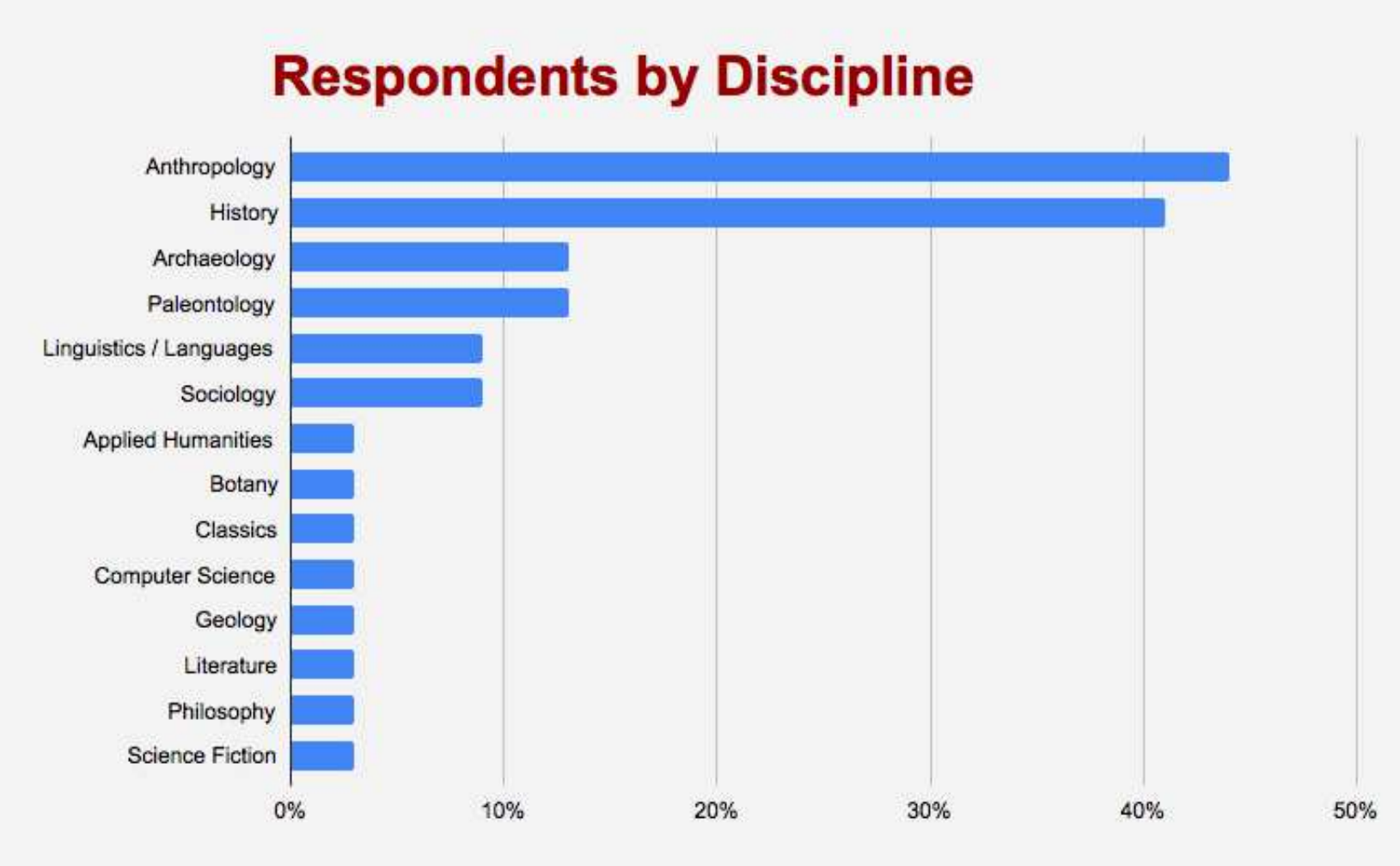
Respondents by Institution



31 responses

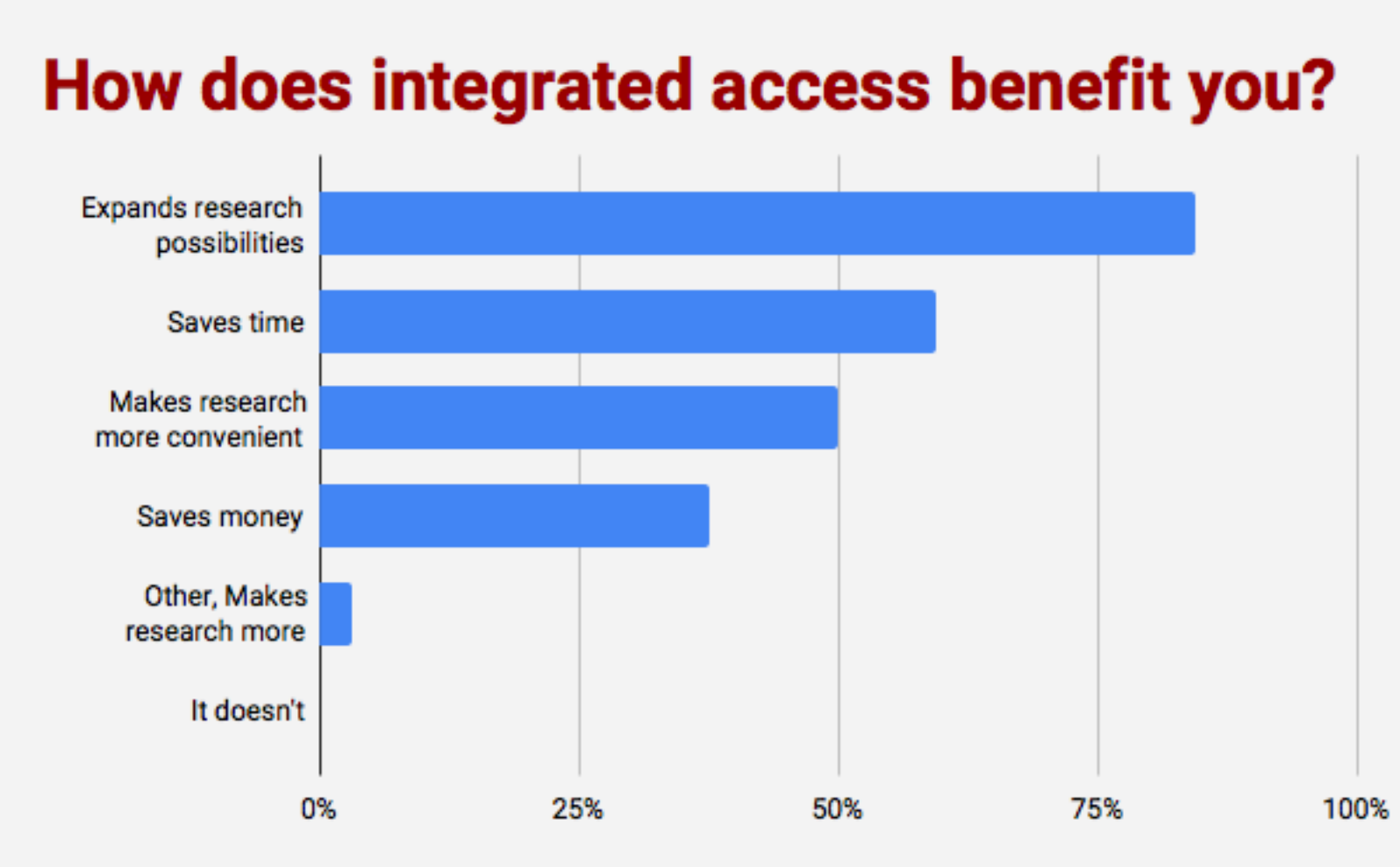
Discipline	Research Interest
Anthropology	Physical anthropology
Anthropology	Ethnoecology, tropical regions; hunter-gatherers; anthropology of food
Anthropology	Ethnoecology
Anthropology, Archaeology	Human Ecology, Human Ecophisiology
Anthropology, Archaeology	Long-term human-environment, especially human-animal interacion, during pre-Columbian and early Historic Era times periods of the circum-Caribbean. I am also actively engaged in the mobilization and digitization of zooarchaeological biological and cultural records as biodiversity specimens in the open access biodiversity network (e.g., GBIF, etc.).
Anthropology, Archaeology, History, Linguistics / Languages	Spatial Humanities, Early Modern, GIS, Corpus Linguistics, AI
Anthropology, Archaeology, Paleontology	Human Evolution, Paleolithique Archaeology, Paleoanthropology, Quaternary Sciences
Anthropology, History	History of Science, Medicine and Empire
Anthropology, History, Philosophy, Sociology, science fiction	Datafication of nature, politics of nature, history of sciences
Anthropology, Sociology	
Anthropology, ethnobotany	Medicinal and ritual plants, transatlantic slavery, historical collections, suriname, west africa
Ethnobotany	Ethnobotany; biocultural heritage
Botany	Cross Cultural Ethnobotany

Discipline	Research Interest
Cultural Heritage	Biology / taxonomy/ 19th century natural history
Applied Humanities	Open innovation in science; applied humanities; knowledge for development; participatory methods; knowledge design; AI; visualization
Classics, Linguistics / Languages, Literature	Digital Philology
Linguistics / Languages	Translation Studies, digital humanities
Computer Science	History of cartography, history of science
History	Provenance research on objects from colonial contexts
History	Various
History	The history of natural history
History	Social and cultural history of botany / natural history
History	History of Entomology
History	History of Sciences, Scientific Illustrations, History of Photography
History	Mediterranean environmental history
Historical ecology, environmental history	Historical ecology, environmental history
History	Geospatial Science, Labour History, Historical demography, Commodity frontiers, Middle Ages, Early Modern
Paleontology	Evolution and systematics of ungulates, biostratigraphy, biogeography, Mio-Plio-Pleistocene, insular ecology and evolution
Paleontology, Earth & Life Sciences	Echinodermata; Phylogeny; Evolution; Systematics; Taxonomy; Nomenclature
Paleontology, Geology	Taphonomy, paleoecology
Sociology	Cultural policy, post-soviet studies, sociology of music, environmental sociology, subcultures



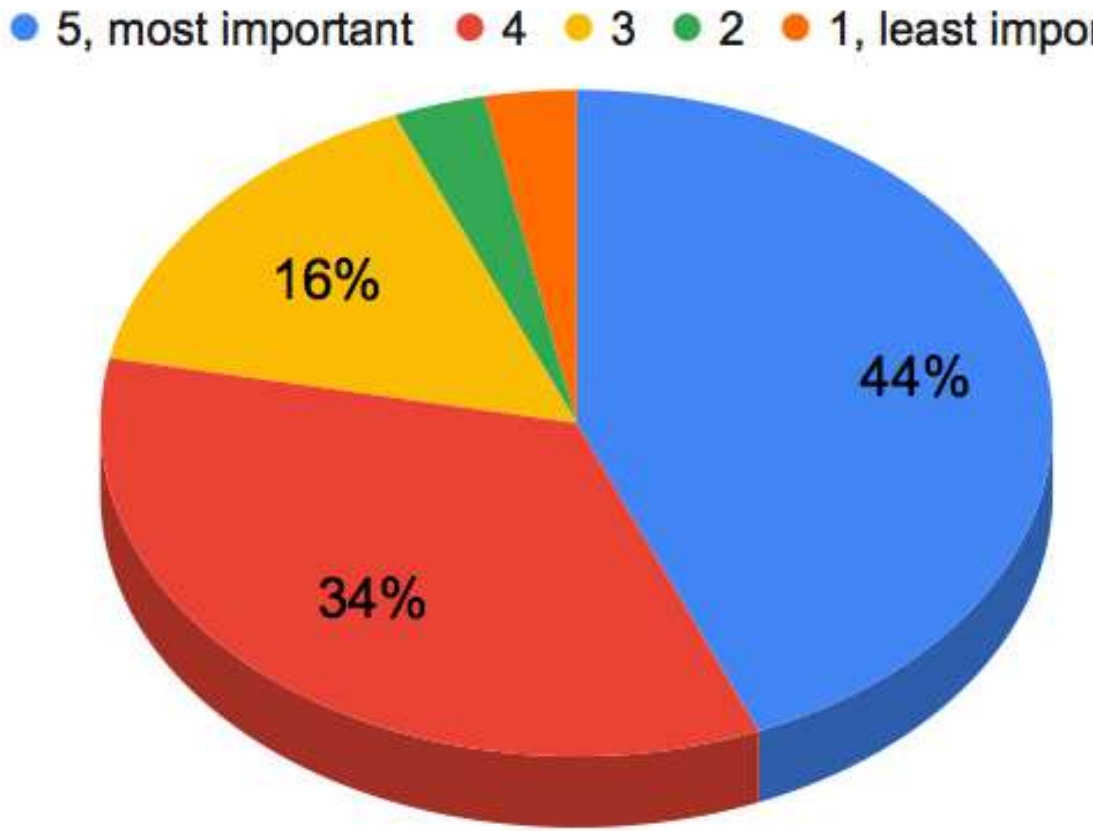
32 responses
14 disciplines

Most Frequently Used RIs	
	Percent
BHL	64%
Europeana	32%
GBIF/iDigBio	12%
Home Grown	12%
Ariadne	4%
Other	48%
Access to physical collections in museums and university laboratories	
Animalbase	
Archives (national, regional, academic, etc.), Wikipedia, Delpher	
Internet Archives; Matricula Online	
CLARIN	
Google services (Books, Scholar)	
Local statistical bureau	
Open Context	
Perseus Digital Library, Musisque Deoque, Memorata Poetis	
ROAD, NQMDB, EVOBREATH, PANTHERIA	
Scopus	
Various sources for cartography, scientific instruments	
25 responses	



32 responses
75 answers

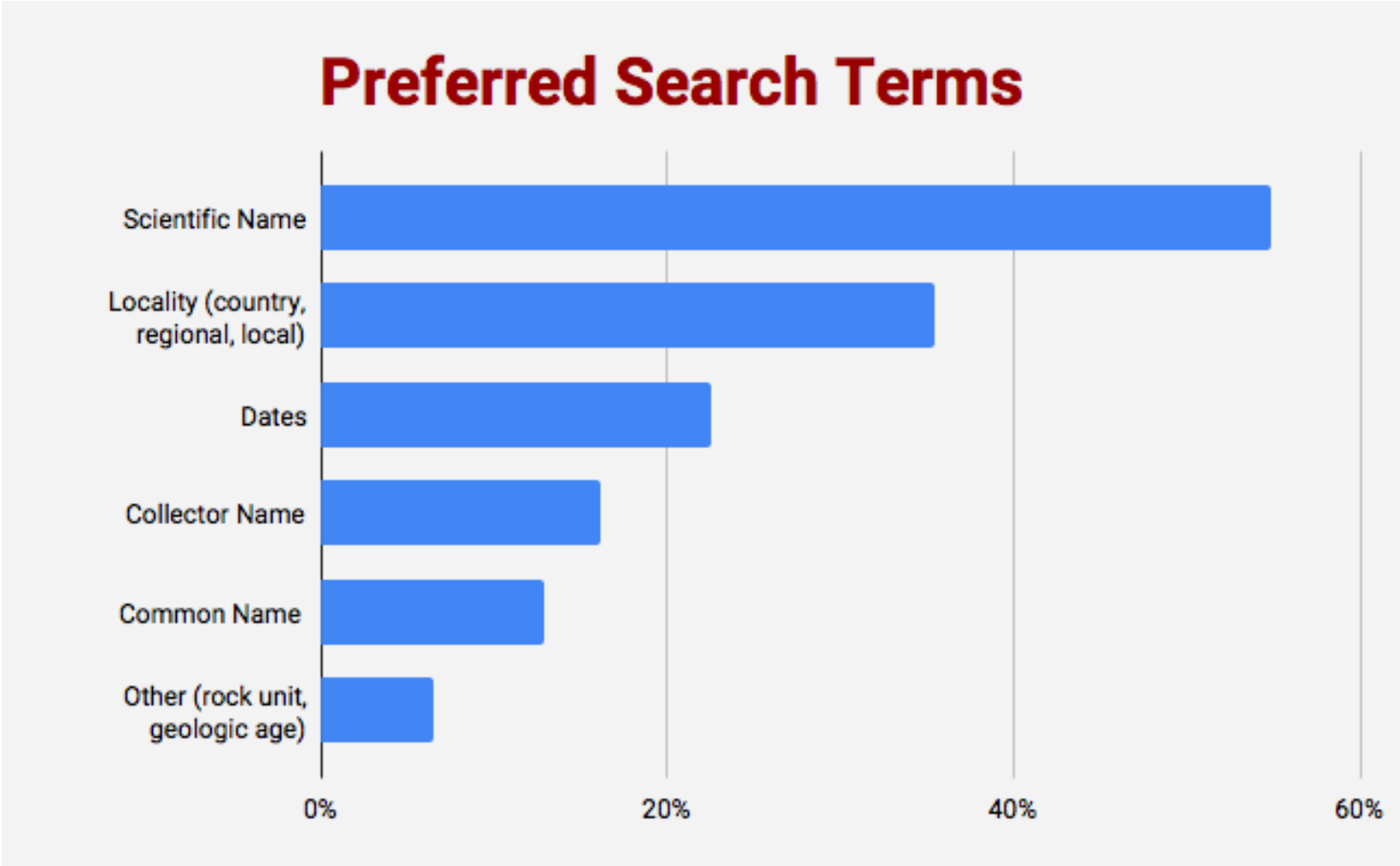
How important is integrated access to natural science and archive data?



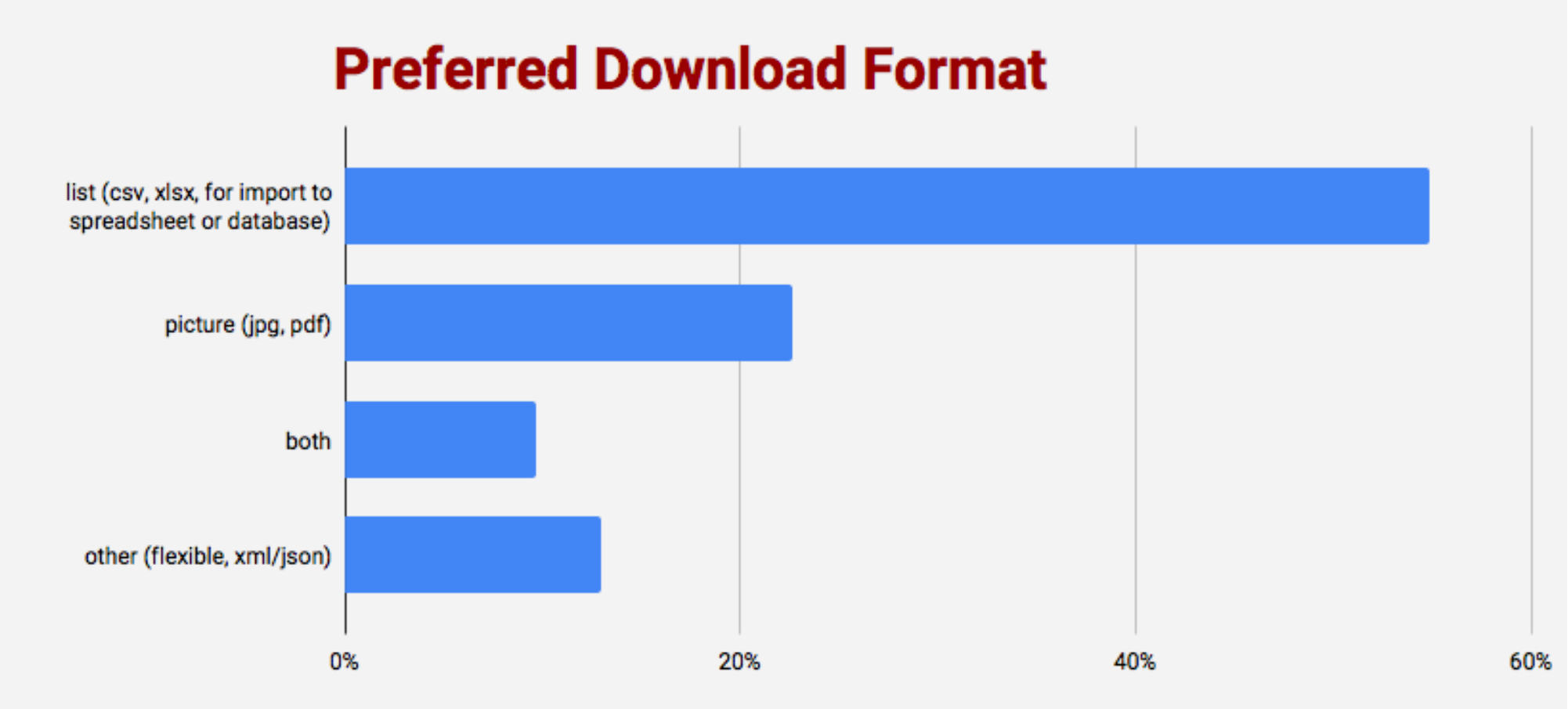
32 responses

Preferred access to natural science collection and archival data

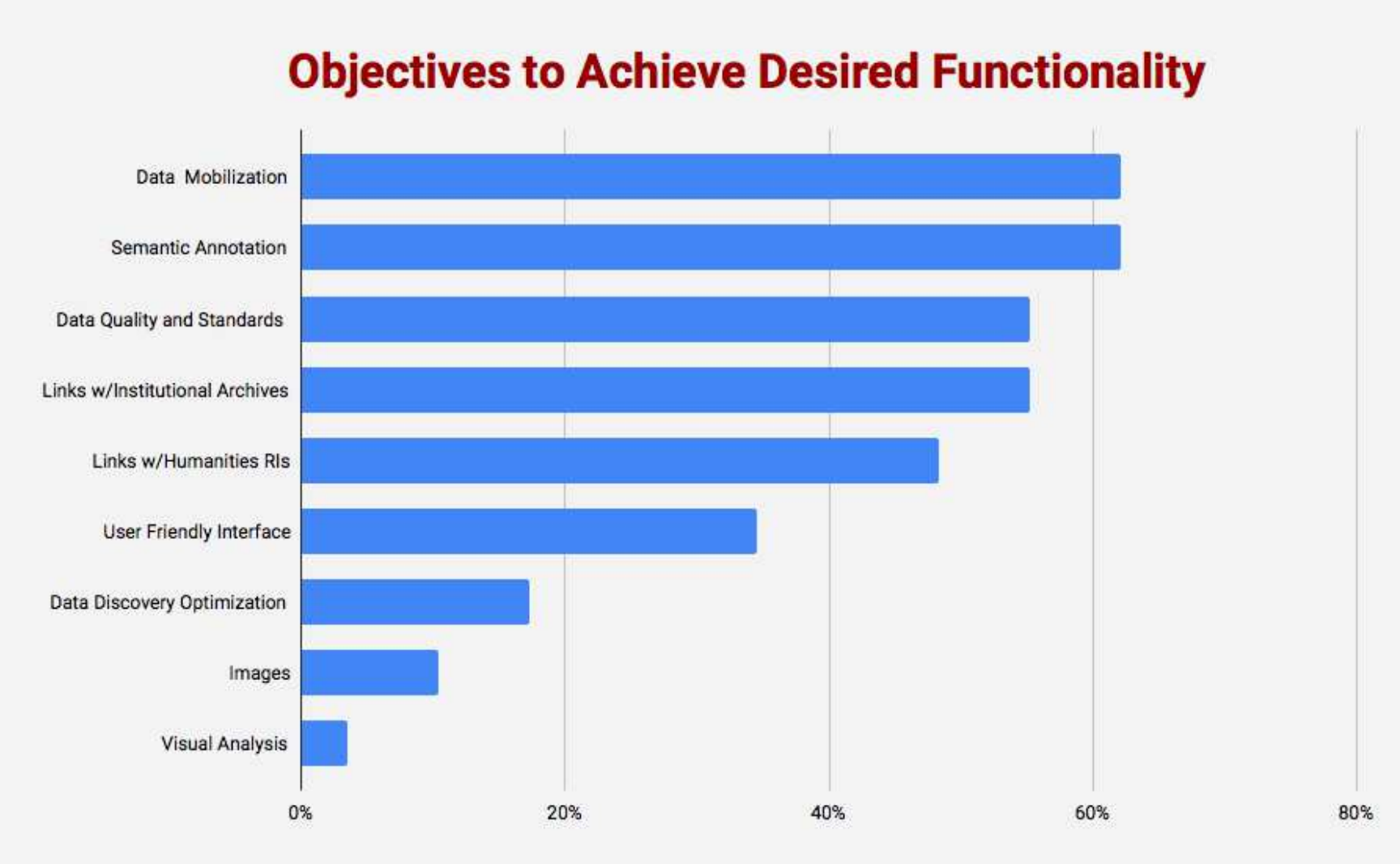
		Percent
Either		44%
Via an online cultural heritage collections resource portal (e.g., BHL, Europeana, Ariadne)		25%
Via a natural science collections data resource portal (e.g., GBIF)		16%
Portal combining both		13%
Physical access		9%
Other		13%
	FAIR data, e.g. Zenodo; project specific/thematic platforms	
	European Open Science Cloud (EOSC)	
	A complete online fossil catalogue	
	Google	
32	responses	



31 responses
46 answers



31 responses
31 answers



29 responses
87 answers

Section 3 of 4

Use Case #1

The following three questions apply to your first 'use case'.

Please identify the following types of natural science collections data you might use in your research. This data is typically found on the object's label and reflects the collecting event. Check all that apply. (Use Case #1)

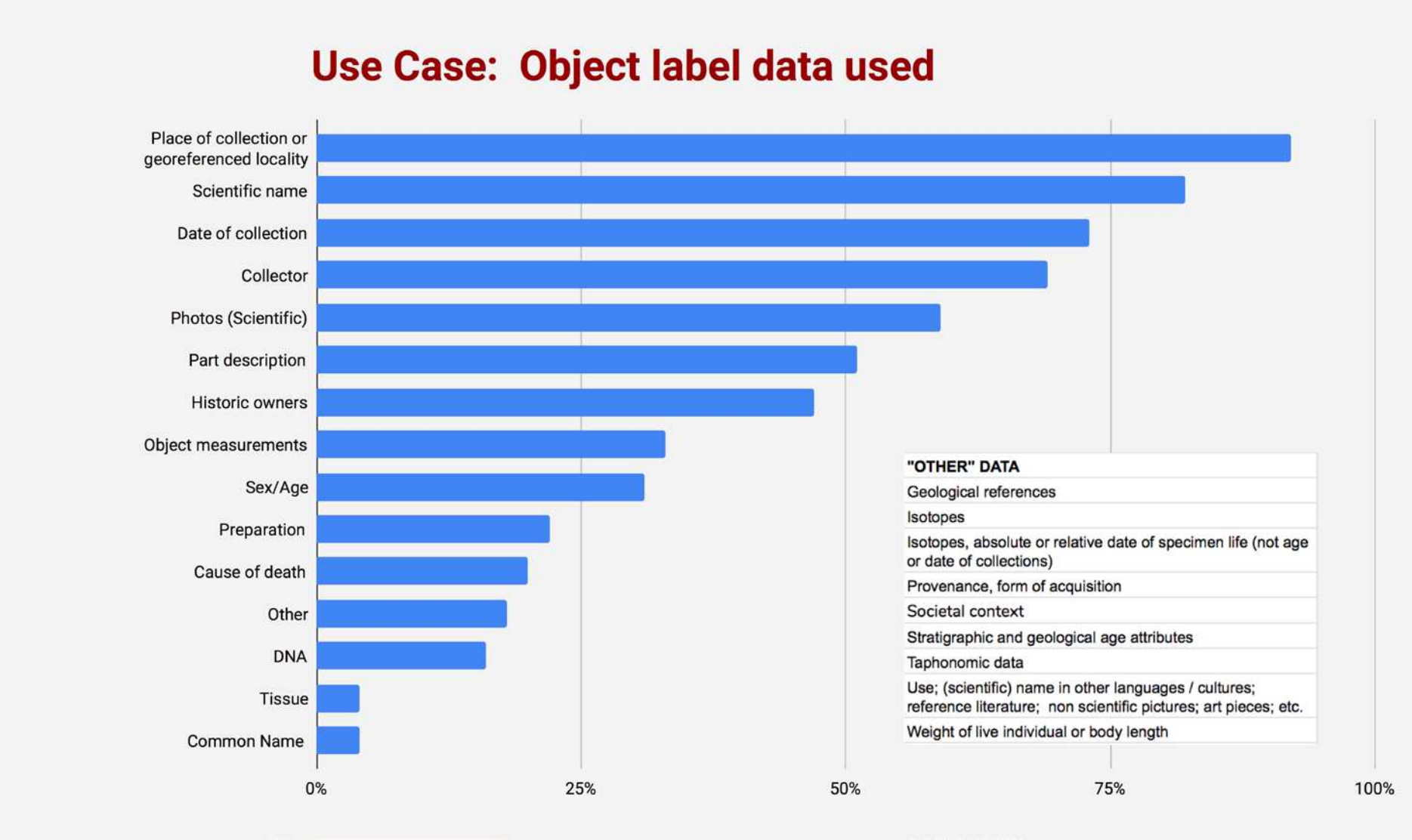
- [ ] Scientific Name
- [ ] Sex / Age
- [ ] Part Description
- [ ] Place of Collection or Georeferenced Locality
- [ ] Date of Collection
- [ ] Collector
- [ ] Preparation
- [ ] Object Measurements
- [ ] Photos (scientific)
- [ ] Historic Owners
- [ ] Cause of Death
- [ ] DNA
- [ ] Tissue Sample
- [ ] Other...

Please identify the following types of natural science archival resources you might use in your research. Check all that apply. (Use Case #1)

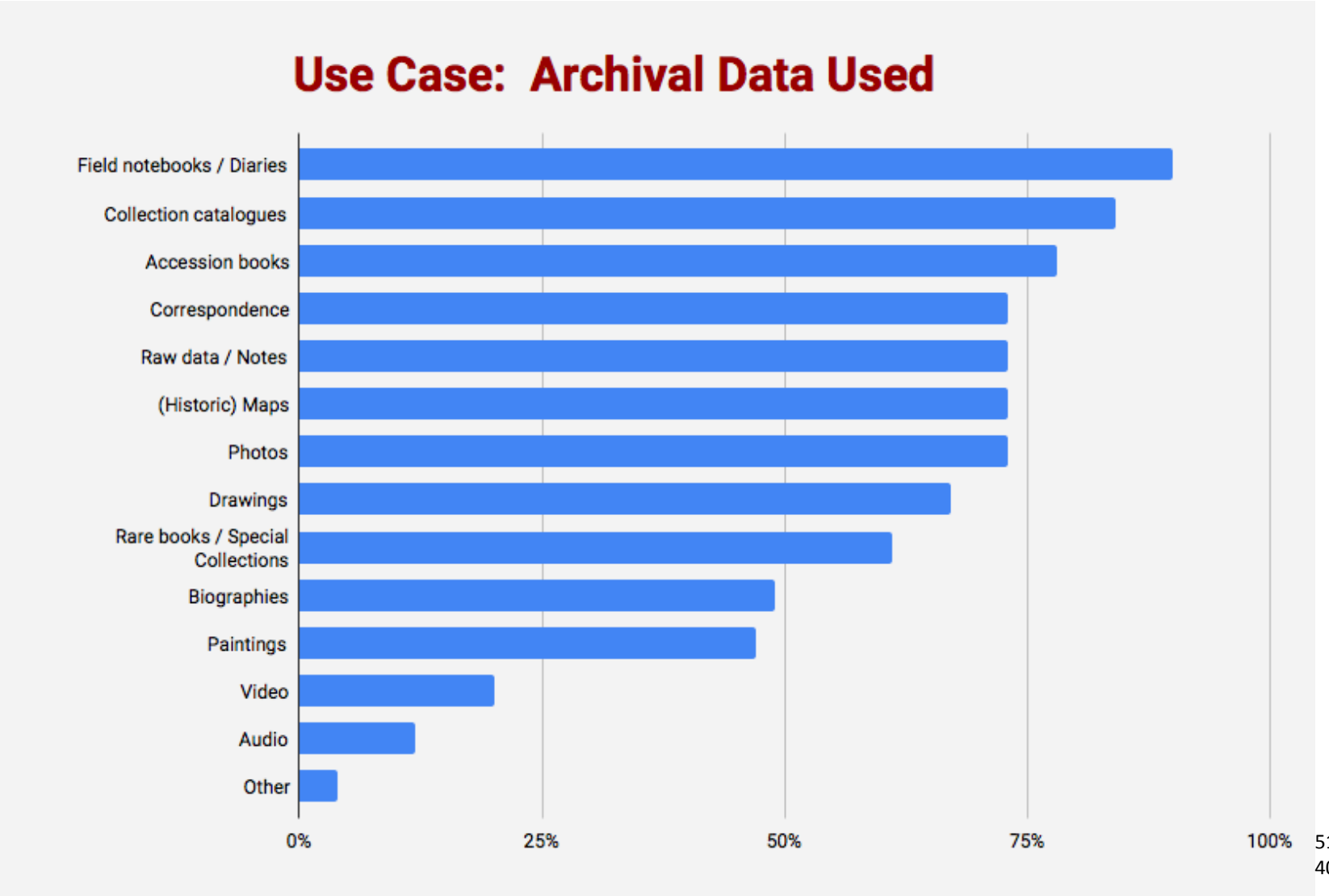
- [ ] Accession Books
- [ ] Collection Catalogues
- [ ] Field Notebooks / Diaries
- [ ] Correspondence
- [ ] Raw Data / Notes
- [ ] Biographies
- [ ] (Historic) Maps
- [ ] Drawings
- [ ] Photos
- [ ] Paintings
- [ ] Audio
- [ ] Video
- [ ] Rare Books / Special Collections
- [ ] Other...

In a few short sentences, please state your research objective for Use Case #1 and briefly describe how the resources check-marked in the two questions above would be used. For example, 'IN ORDER TO provide temporal and social context to a painting I NEED TO identify the species of wood of the painting's frame USING the museum's wood reference collection and associated data.'

Long-answer text



51 responses
316 answers


Use Case: Archival Data Used
Field notebooks / Diaries
Collection catalogues
Accession books
Correspondence
Raw data / Notes
(Historic) Maps
Photos
Drawings
Rare books / Special Collections
Biographies
Paintings
Video
Audio
Other
0%
25%
50%
75%
100%
51 responses
409 answers

USE CASE
GENERAL CATEGORIZATIONS

- Historic reconstruction of persons, objects, collections, events
- Historic reconstruction to correlate with human influence
- Reference: Identification of species or species comparisons
- Biologic / paleontologic study

USE CASE: Historic reconstruction of persons, objects, collections, events

Discipline / Research Interest	Objective	One Must	Requirement (Inferred)	Data Requirement	Archive Requirement	Survey Number and Use Case Original Text
Historian, History of Sciences, Scientific Illustrations, History of Photography	To fully understand/trace an object or subject's history (how an object came to the institution, provenance, circumstances of collecting event)	Use all available sources of information, images, text (publications, correspondence, notes, etc.) audio, or moving image.	• Integrated humanities and natural science RIs including associated archives. • Semantically linked data	Scientific Name Sex / Age Part Description Place of Collection or Georeferenced Locality Date of Collection Collector Preparation Object Measurements Photos (scientific) Historic Owners Cause of Death Provenance form of acquisition,	Accession Books Collection Catalogues Field Notebooks / Diaries Correspondence Raw Data / Notes Biographies (Historic) Maps Drawings Photos Paintings Audio Video Rare Books / Special Collections	**30** UC1: In order to fully understand/trace an objects/subjects history (how an object came to the institution, provenance, circumstances how it was collected) I have to use all available sources be it an image, text (publications, correspondence, notes, etc...), audio or moving image
	To put a collection in socio-historical context	Use all available resources, for example, a specimen's sex, age and date of collection is relevant for identifying a possible bias of the collectors/collection during a certain period of time.		Scientific Name Sex / Age Part Description Place of Collection or Georeferenced Locality Date of Collection Collector Preparation Object Measurements Photos (scientific) Historic Owners Cause of Death	Accession Books Collection Catalogues Field Notebooks / Diaries Correspondence Raw Data / Notes Biographies (Historic) Maps Drawings Photos Paintings Audio Video Rare Books / Special Collections	UC2: In order to put a collection in a socio-historical context I have to use all available sources. For example a specimens sex, age and date of collection is relevant ford identifying a possible bias of the collectors/collection during a certain period in time.

USE CASE: Historic reconstruction to correlate human influence

Discipline / Research Interest	Objective	One Must	Requirement (Inferred)	Data Requirement	Archive Requirement	Survey Number and Use Case Original Text
Anthropologist, Archaeologist Long-term human-environment, especially human-animal interaction, during pre-Columbian and early Historic Era times	To understand the human impact on the spatial, temporal and cultural distribution and use of a species	Quantify the species' presence across sites, create chronological context, and conduct morphological comparisons across individuals	• Integrated humanities and natural science RIs including associated archives. • Semantically linked data	Scientific Name Part Description Place of Collection or Georeferenced Locality Object Measurements DNA isotopes absolute or relative date of specimen life (not age or date of collections)	Accession Books Collection Catalogues Field Notebooks / Diaries Correspondence Raw Data / Notes (Historic) Maps	**31** Long-term human-environment, especially human-animal interacion, during pre-Columbian and early Historic Era times periods of the circum-Caribbean. I am also actively engaged in the mobilization and digitization of zooarchaeological biological and cultural records as biodiversity specimens in the open access biodiversity network (e.g., GBIF, etc.). UC1: In order to understand the human impact on the spatial, temporal, and cultural distribution and use of a species, I need to be able to quantify their presence across sites, create chronological context, and conduct morphological comparisons across individuals.
	To identify patterns of animal management or incipient domestication	• Quantify target taxa across a site(s); • determine age and sex to create mortality and demographic profiles of the taxa, • conduct aDNA to assess impacts of human influence on population genetic diversity, and • use isotopic analysis to assess diet.		Scientific Name Sex / Age DNA isotopes	Field Notebooks / Diaries Raw Data / Notes	UC2: In order to identify patterns of animal management or incipient domestication, I would need to be able to quantify target taxa across a site(s), determine age and sex to create mortality and demographic profiles of the taxa, conduct aDNA to assess impacts of human influence on population genetic diversity, and use isotopic analysis to assess diet.

USE CASE: Reference: Identification of species or species comparisons

Discipline / Research Interest	Objective	One Must	Requirement (Inferred)	Data Requirement	Archive Requirement	Survey Number and Use Case Original Text
Anthropologist, Ethnoecology	To determine the natural species used in the construction of objects	**(I)** Use reference collection objects, online resources and associated archives to identify a species	• Integrated humanities and natural science RIs including associated archives. • Semantically linked data	Scientific Name Place of Collection or Georeferenced Locality Date of Collection Collector Photos (scientific)	Collection Catalogues Field Notebooks / Diaries Correspondence Biographies (Historic) Maps Photos	**19** UC1: to determine the natural species used in the construction of objects and to compare objects between our collections and the similar collections in other institutions
	To compare objects between our collections and similar collections in other institutions	**(I)** Have physical or online access to similar collections including data and photos				
	To evaluate changes in the making of objects	Understand object's temporal context		Object Measurements Historic Owners	Accession Books Raw Data / Notes Drawings Paintings	UC2: temporal informations to evaluate the changes in the making of the objects

USE CASE: Biologic / paleontologic study

Discipline / Research Interest	Objective	One Must	Requirement (Inferred)	Data Requirement	Archive Requirement	Survey Number and Use Case Original Text
Anthropologist, Archaeologist, Paleontologist, Human Evolution	To use fossil bones and lithic collections in museum and university collections upon which my research relies	Access their catalogues, digital pictures and drawings (for old collections) as well as any other useful information for my research such as diaries, notes, and letters	• Integrated humanities and natural science RIs including associated archives. • Semantically linked data	Part Description Place of Collection or Georeferenced Locality Date of Collection Object Measurements Photos (scientific)	Collection Catalogues Field Notebooks / Diaries Correspondence Raw Data / Notes Drawings Photos Rare Books / Special Collections	**24** UC1: In my research I frequently rely on fossil bone and lithic collections stored in museums and university laboratories. To use the collections, it is important to access their catalogues, digital pictures and drawings (for old collections) as well as any other useful information for my research such as diaries, notes, and letters.
	To research fossils and prehistoric stone tools	Compile information such as scientific name, dating, geographic information on the localities where they were found, basic quantitative and qualitative description. Other data needed is field notes, illustration (drawings and photos), and any other useful information about them that was published or unpublished.		Scientific Name Sex / Age Part Description Place of Collection or Georeferenced Locality Date of Collection Collector Object Measurements Photos (scientific)	Collection Catalogues Field Notebooks / Diaries Correspondence Raw Data / Notes (Historic) Maps Drawings Photos Rare Books / Special Collections	UC2: For my research on fossils and prehistoric stone tools, I would need information such as scientific name, dating, geographic information on the localities where they were found, basic quantitative and qualitative description. Other data needed is field notes, illustration (drawings and photos), and any other useful information about them that was published or unpublished.

RESPONDENT PRIMARILY:

- west European
  - from universities and natural history museums
- anthropologists and historians

RESPONDENT:

- primarily uses BHL and Europeana
- feels that *integrated* access
  *expands their research possibilities
  *saves time
  *is convenient
  *is very important

RESPONDENT

- can access science RI via *both* science and humanities portals
- searches using scientific name, then locality
- prefers to download in list format

USE CASE: Object label data

50% or more of the respondents use:

- Place of collection or georeferenced locality
- Scientific name
- Date of collection
- Collector
- Photos (scientific)
- Part description
- Historic owner

USE CASE: Archives

50% or more of the respondents use almost *all* the archival resources.

- Field notebooks / diaries
- Collection catalogues
- Accession books
- Correspondence
- Raw data / notes
- (Historic) maps
- Photos
- Drawings
- Rare books / special collections
- Biographies
- Paintings

USE CASE TYPES

- Historical reconstructions
- Historical reconstructions correlated with human influence
- Reference: species ID and comparison
- Scientific studies

IN CONCLUSION

Survey establishes:

There is sufficient demonstrated need by humanities researchers working at the interface of science and humanities
for the use of an *integrated* science and humanities data resource.