



Mobilise Action

Mobilising Data, Policies and Experts in Scientific Collections

Crowdsourcing for Natural Science Collections - closing the circle of data flow

WORKSHOP REPORT

An online workshop convened by WG2 of the MOBILISE Cost Action 17106
(<https://www.mobilise-action.eu>)

2nd November 2020 - 08:00-11:00 UTC

9th November 2020 - 14:00-17:00 UTC

Joaquim Santos, Arnald Marcer, Elspeth Haston



**Funded by
the European Union**

This report is based upon work from COST Action CA 17106 MOBILISE, supported by COST (European Cooperation in Science and Technology).
COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

www.cost.eu

Suggested citation: Santos J., Marcer A. and Haston E. (2023). Crowdsourcing for Natural Science Collections - closing the circle of data flow - workshop report. MOBILISE EU Cost Action CA1706.

<http://doi.org/10.5281/zenodo.7853039>

Author affiliations

Joaquim Santos, University of Coimbra, Portugal, ORCID [0000-0002-2160-4968](https://orcid.org/0000-0002-2160-4968)

Arnald Marcer, (1) CREAM, 08193 Bellaterra (Cerdanyola del Vallès), Catalonia, Spain, ORCID [0000-0002-6532-7712](https://orcid.org/0000-0002-6532-7712); (2) Universitat Autònoma de Barcelona, E 08193, Bellaterra (Cerdanyola del Vallès), Catalonia, Spain

Elsbeth Haston, Royal Botanic Garden Edinburgh, UK, ORCID [0000-0001-9144-2848](https://orcid.org/0000-0001-9144-2848)

Introduction	5
Overview	5
Session 1	6
Programme	6
Participants	6
Moderators	6
Panellists	6
Attendants	7
Talks	7
Les Herbonautes (Marc Pignal - ReColNat)	7
DigiVol (Paul Flemons - The Australian Museum)	7
DoeDat (Mathias Dillen - Meise Botanic Garden)	8
EXPLORATOR (Joaquim Santos - University of Coimbra)	8
JACQ (Heimo Rainer - University of Vienna)	8
DINA (Falko Glöckler)	8
Panel Discussion	8
Poll Result	9
Session 2	10
Programme	10
Participants	10
Moderators	10
Panellists	10
Attendants	11
Talks	11
Zooniverse (Samantha Blickhan - Zooniverse)	11
iDigBio/WeDigBio/BIOSPEX (Austin Mast - University of Florida / iDigBio)	11
Smithsonian Transcription Center (Sylvia Orli & Rebecca Snyder - Smithsonian National Museum of Natural History)	12
Die Herbonauten (Agnes Kirchhoff - Garten und Botanisches Museum Berlin)	12
Specify (Jim Beach & Norine Spears - Specify)	12
Earthcape (Evgeniy Meyke - Earthcape)	12
ARCTOS (Mariel Campbell - ARCTOS)	12
BG-BASE (Kerry Walter & Mahir Balik - BG-BASE)	12
Panel Discussion	13
Poll Result	14

Introduction

The current drive to digitise the estimated 1.5 billion objects in Natural Science Collections across Europe has resulted in massive numbers of digitised specimens becoming accessible online. However, very little of the collection data associated with these objects is currently available online given the vast amount of work involved in transcribing the specimen labels. There have been several approaches taken to increase the number of transcribed specimens, including outsourcing to external companies and taking the task to the general public for their help through citizen science and crowdsourcing platforms. These platforms have proved to be highly successful, both in the rate and quality of transcription and in engaging broader and more diverse audiences with the Natural Science Collections. However, there are still significant hurdles to overcome to achieve seamless movement of data between the institutional collection management systems (CMS) and the crowdsourcing platforms, with many transcribed data being unable to be re-ingested into the institutional database.

Overview

Two 3 hour virtual workshops were held on the 2nd and 9th of November 2020 which analysed the current state of standards and best practice for managing data migration between crowdsourcing platforms and collection management systems. The workshops brought together developers of many of the principal crowdsourcing platforms, collection management systems along with an opportunity for collection managers, curators and data managers to participate in two sessions which included short presentations and a panel discussion to have wider discussion through a Q&A forum.

The workshop was organised in 2 sessions one week apart. The timetable of the two sessions was programmed to allow live participation from different time zones:

Session 1 - 2nd November 2020 08:00 - 11:00 UTC

Session 2 - 9th November 2020 14:00 - 17:00 UTC

The recordings of the sessions were made available the day following the session over Zoom platform, and also shared on youtube MOBISLISE channel for posterity:

https://www.youtube.com/channel/UCFTWtmwUi34J_Wk3XK99I5w

Shared documents were created for interaction with the audience, where participants could post questions and get in contact with each other.

Session 1

Programme

2 November 2020

08:00 - 08:15 (UTC)

Introduction & Aims of workshop (10 mins)

08:15 - 09:20

Lightning presentations to cover the current crowdsourcing effort (10 mins each)

- Les Herbonautes (Marc Pignal)
- Die Herbonauten (Agnes Kirchhoff) - POSTPONED FOR THE NEXT SESSION
- DigiVol (Paul Flemons)
- Doedat (Mathias Dillen)
- EXPLORATOR (Joaquim Santos)

Q&A (ALL)

09:20 - 09:30

Break

09:30 - 10:15

Lightning presentations to cover the CMS activity in terms of crowdsourced data (10 mins each)

- JACQ (Heimo Rainer)
- PLUTO-F (Urmaz Kõljalg) - CANCELLED
- DINA (Falko Glöckler)

Q&A (ALL)

10:15 - 11:00

Panel Discussion

Participants

Moderators

Elspeth Haston - Royal Botanic Garden Edinburgh

Arnal Marcer - Universitat Autònoma de Barcelona

Joaquim Santos - University of Coimbra

Panellists

Bolzinger Marie-Françoise - Les Herbonautes/ReColNat (in representation of Marc Pignal)

Paul Flemons - The Australian Museum

Mathias Dillen - Meise Botanic Garden

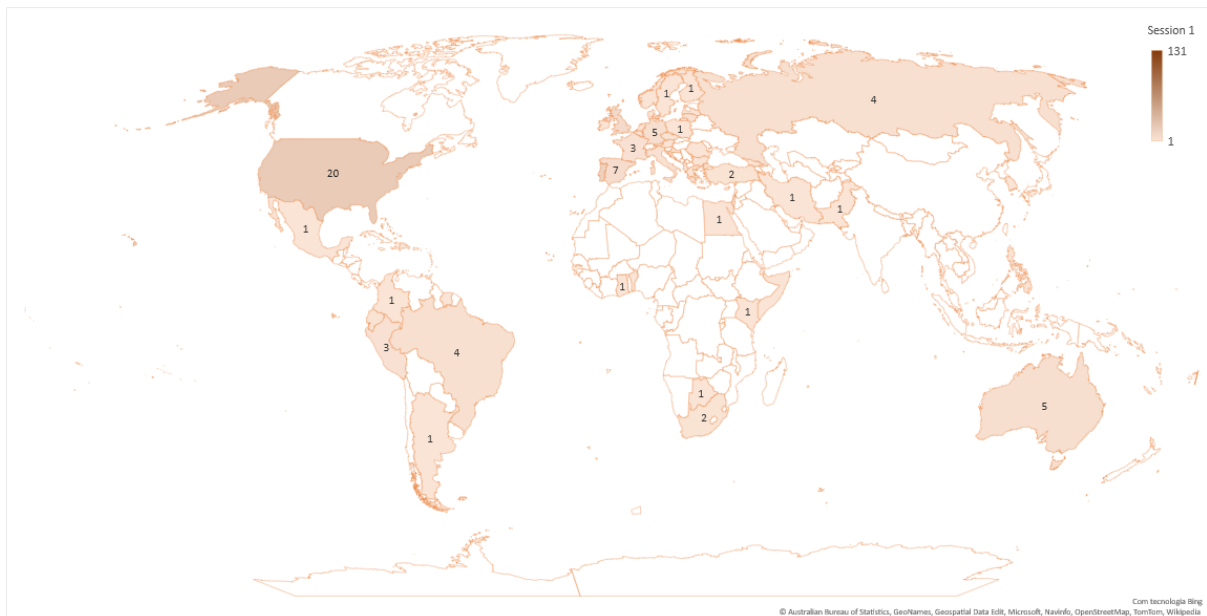
Joaquim Santos - University of Coimbra

Heimo Rainer - University of Vienna

Falko Glöcker - Museum für Naturkunde Berlin

Attendants

A total of 131 people from 50 countries registered to attend the 1st session. According to the stats provided by Zoom, 50% (66) of this number attended live.



Origin of registrants to Session 1

Talks

After a short introduction to the workshop programme and objectives, the first series of presentations took place.

Les Herbonautes (Marc Pignal - ReColNat)

The crowdsourcing platform Les Herbonautes was presented on a recorded video. A demonstration was made of the functions of the users layout, as well as the backoffice. The platform is used for citizen transcription and classification by the ReColNat infrastructure, a network of Natural History Collections in France.

DigiVol (Paul Flemons - The Australian Museum)

Explanation of DigiVol by the Australian Museum, focusing on some concepts that are behind the functioning of that platform. Special focus on some of the problems of integration with collection management systems as the standardisation of data. Also a short live demonstration of the platform.

DoeDat (Mathias Dillen - Meise Botanic Garden)

Insight to the DoeDat, explaining the origin of the platform and some figures of the utilisation. Explanation of the data flow to integrate data with the CMS database, with focus on the standardisation needed.

EXPLORATOR (Joaquim Santos - University of Coimbra)

Description of a custom crowdsourcing platform, and its live integration with SPECIFY. Focus on validation using users experience as criteria and the incremental update of data into the CMS.

JACQ (Heimo Rainer - University of Vienna)

Demonstration of JACQ software and of the workflow to import data from CSV files into JACQ using an import form. The process allows incremental updates and reconciliation with existing records.

DINA (Falko Glöckler)

Presentation of the DINA Consortium and a conceptual overview of the data model that the CMS project is building as a modular system to allow interoperability through APIs and potential integration with crowdsourcing activities.

Panel Discussion

After the two rounds of presentations, there was a discussion of some aspects involving the panelists, also based on the questions and comments on the shared document that all the participants could edit.

It started with the User's perspective by stating that it is very important for the volunteers using the crowdsourcing platform that there is communication from the collection managers and also with other volunteers, because the contribution is a learning process that can be eased by the exchange of knowledge. The transcription involves often only verbatim values, but there is always some degree of interpretation. It requires research and judgement to decide about the right values to use.

From the Collections point of view, the task of integrating the data coming from crowdsourcing platforms is time consuming, because it requires the validation and harmonisation of data. It would be desirable to reduce the amount of cleaning and human intervention between the crowdsourcing accepted values and the CMS, but there are technical and conceptual obstacles to this. The main problem is that the CMS data model is too complex to reproduce in a crowdsourcing platform in a way that could be used easily by volunteers.

The availability of crowdsourcing platforms for institutes was mentioned, and how they can be used or installed by external institutes. Some platforms are available to install with total

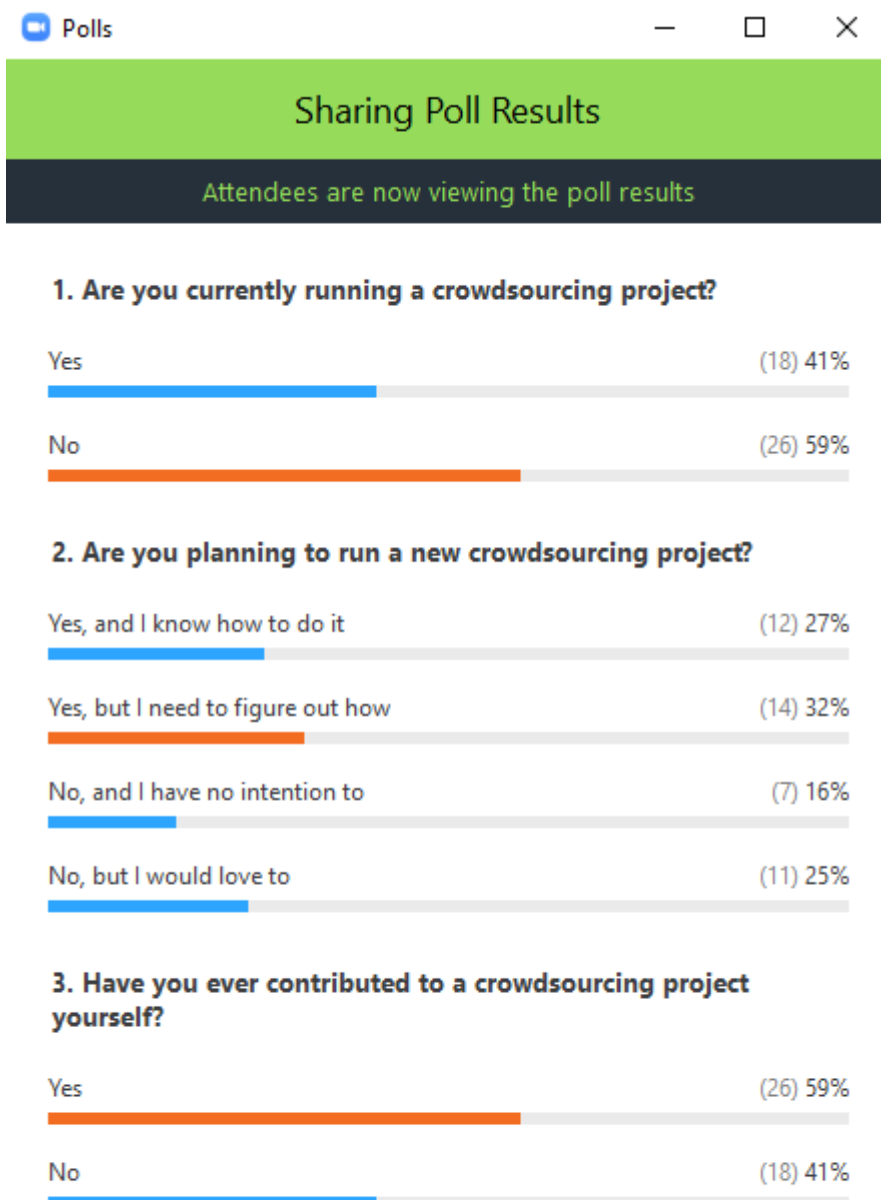
control, which is not an easy process. It is also possible to host transcribing projects/missions on platforms such as DigiVol or DoeDat.

The ethical aspect was approached briefly, referring the caution needed to protect users' privacy

The recording of the session is available here:

[Crowdsourcing for Natural Science Collections, 2nd November 2020](#)

Poll Result



Session 2

Programme

9 November 2020

14:00 - 14:15 (UTC)

Introduction (10 mins)

14:15 - 14:25

Review of presentations from 1st session (20 mins)

14:25 - 15:15

Presentations to cover the current crowdsourcing effort (10 mins each)

- Zooniverse (Samantha Blickhan)
- iDigBio/WeDigBio/BIOSPEX (Austin Mast)
- Smithsonian Transcription Center (Sylvia Orli & Rebecca Snyder)
- Die Herbonauten (Agnes Kirchhoff & Dominik Röpert)

Q&A (ALL)

15:15 - 15:25

Break

15:25 - 16:25

Presentations to cover the CMS activity in terms of crowdsourced data (10 mins each)

- Specify (Jim Beach & Norine Spears)
- Earthcape (Evgeniy Meyke)
- ARCTOS (Mariel Campbell)
- BG-BASE (Kerry Walter & Mahir Balik)

Q&A (ALL)

16:25 - 17:00

Panel Discussion & Next Steps

Participants

Moderators

Elsbeth Haston - Royal Botanic Garden Edinburgh

Arnal Marcer - Universitat Autònoma de Barcelona

Joaquim Santos - University of Coimbra

Panellists

Samantha Blickhan - Zooniverse

Grant Miller - Zooniverse

Austin Mast - University of Florida / iDigBio

Sylvia Orli - Smithsonian National Museum of Natural History

Rebecca Snyder - Smithsonian National Museum of Natural History

Agnes Kirchhoff - Botanischer Garten und Botanisches Museum Berlin

Dominik Röpert - Botanischer Garten und Botanisches Museum Berlin

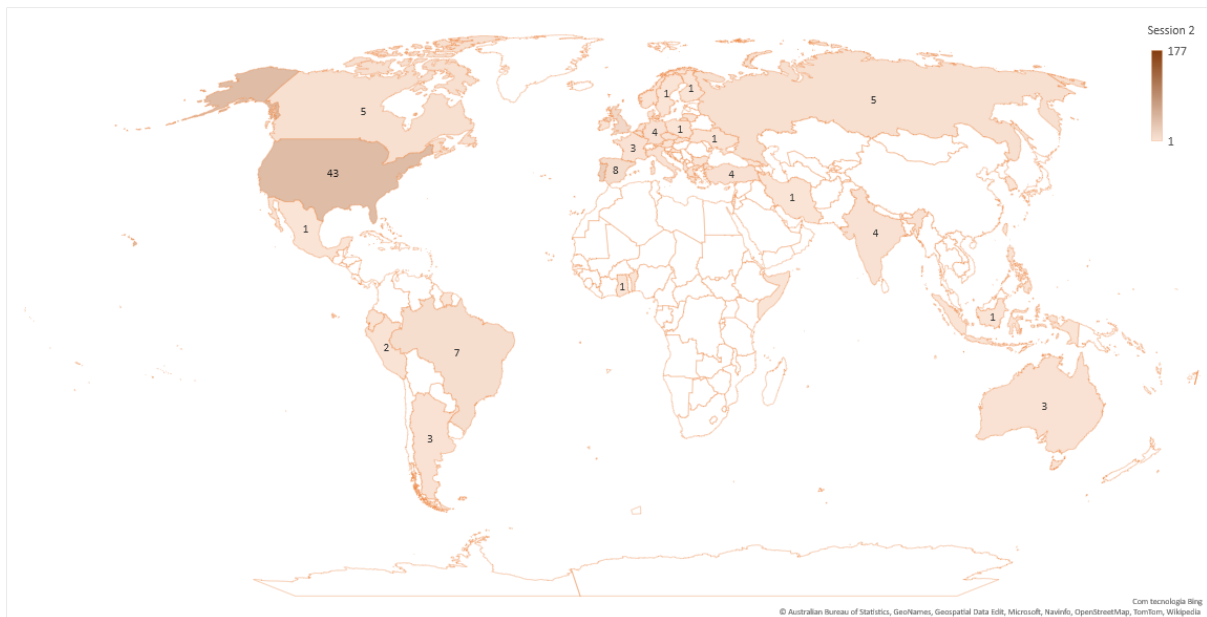
Jim Beach - Specify

Norine Spears - Specify

Evgeniy Meyke - Earthcape
Mariel Campbell - ARCTOS
Kerry Walter - BG-BASE
Mahir Balik - BG-BASE

Attendants

A total of 177 people from 47 countries registered to attend the 2nd session. According to the stats provided by Zoom, 50% (89) of this number attended live.



Origin of registrants to Session 2

Talks

The session started with a short introduction to the workshop programme and objectives, and also a summary of the previous session

Zooniverse (Samantha Blickhan - Zooniverse)

Introduction to the Zooniverse platform, namely the Notes from Nature project. It was an overview of the basic concept and tools available for the community.

iDigBio/WeDigBio/BIOSPEX (Austin Mast - University of Florida / iDigBio)

The presentation was focused on Biospex, a tool to allow the dataflow between CMSs and external sources as Zooniverse, and how it is being used by institutes in the iDigBio community. It consists of a visual tool that automates the data transfer mechanisms from and to CMSs

Smithsonian Transcription Center (Sylvia Orli & Rebecca Snyder - Smithsonian National Museum of Natural History)

Overview of the platform used by the Smithsonian National Museum of Natural History to transcribe labels with the help of volunteers. A quick explanation on the criteria for the institute to add specimens to the transcription center and the workflow they follow. The crowdsourcing is totally integrated with the EMu CMS used by them.

Die Herbonauten (Agnes Kirchhoff - Garten und Botanisches Museum Berlin)

Explanation on how the german version of Les Hebonautes is used by the Berlin Herbarium, mostly on the user interaction processes, such as the wiki and the conflict resolution communication. Integration of crowdsourced data is imported into a separate table of the CMS and is displayed on the herbarium catalogue as a separate resource.

Specify (Jim Beach & Norine Spears - Specify)

Presentation of Specify Software, specially the way of how data can be ingested into the database. The option to transcribe from an image using the form interface on Specify6 was shown.

Earthcape (Evgeniy Meyke - Earthcape)

Demonstration of the software capabilities, namely the possibility to configure forms, which enables the possibility to have specific forms to run crowdsourcing projects with the software interface. Validation of values can happen on form interface, but also on spreadsheet view.

ARCTOS (Mariel Campbell - ARCTOS)

Enumeration of Arctos platform properties and its versatility for exchange of data. Exemplification of annotation to records using the ready available form in Arctos, which can be used to enrich specimen records.

BG-BASE (Kerry Walter & Mahir Balik - BG-BASE)

Presentation on digitisation principles for collections, such as verbatim, derived data, atomised fields, table interactions, etc. Demonstration of some software screens. Import and export processes were addressed.

Panel Discussion


From the Q&A and discussion sections, several topics were addressed. Some of the most relevant were a bit more on the technical side, such as mapping of fields on the ingestion process or using clustering techniques on the crowdsourced data.

It was clear from the presentations and comments that CMSs and crowdsourcing platforms need to be flexible to allow data flow. On the other hand, customisation of fields makes the task of integration more difficult, since the mapping relation is not obvious. When getting data into CMS from external sources, it might be useful to be able to merge with existing data and keep a log history of the changes made.

The use of universal identifiers was also mentioned as an advantage when dealing with information from different collections in order to ease the integration and cleaning of data.

The ethical issues are a concern in people's minds. This is tightly connected with the relationship between institutes and volunteers. It was stated that crowdsourcing projects need to be clear about how the data will be used, as it can go against the users' expectations. It is also recommended to acknowledge volunteers on the publication of crowdsourced data.

The recording of the session is available here:

 [Crowdsourcing for Natural Science Collections, 9th November 2020](#)

Poll Result

🗳️ Polls

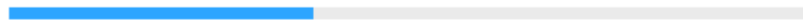


Sharing Poll Results

Attendees are now viewing the poll results

1. Are you currently running a crowdsourcing project?

Yes (29) 38%

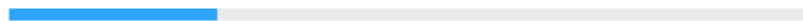


No (47) 62%



2. Are you planning to run a new crowdsourcing project?

Yes, and I know how to do it (20) 26%



Yes, but I need to figure out how (26) 34%



No, and I have no intention to (8) 11%



No, but I would love to (22) 29%



3. Have you ever contributed to a crowdsourcing project yourself?

Yes (52) 68%



No (24) 32%

