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 D5.1 / R]/ PUB; DEC
Work package WP5 / Task 5.2 Lead: MNHN
Delivery date M16

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

Working with citizens to enrich data: recommendations and source repository

Deliverable 5.1 - Recommendations for volunteer transcription systems and a source repository

Gwenaël Le Bras[1]

Simon Chagnoux[1]

Mathias Dillen[2]

1 - Muséum National d'Histoire Naturelle (Paris - France)

2 - Agentschap Plantentuin Meise (Meise - Belgium)

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

ICEDIG.EU

About this document

The initial ICEDIG proposal has attributed a single number D5.1 for the delivery of all the outcomes of task 5.2. The work done has produced four results of very different natures: report document, specification document, source-code repository and structured datasets.

The present document is just a brief introduction of Task 5.2 and the produced documents, repository and datasets.

1. Introduction

The Distributed Systems of Scientific Collections (DiSSCo) will facilitate the production of millions of natural history specimen collection images. Such a large scale digitisation of natural heritage enables new workflows where imaging precedes cataloguing and allows the general public to be involved in the documentation of natural specimens.

Today half a million of labels have already been transcribed by citizens on a dozen of websites all over the world. The aim of task 5.2 "Working with citizens to enrich data", is to explore how to foster this effort and integrate current and future transcribing platforms into the DiSSCo infrastructure.

2. Methodology

The task started by an exhaustive review of existing transcription systems. The naive idea of identifying the "best-fit" platform for DiSSCo could not resist the fact that for each platform, the main asset is not the website in itself but the community built around it.

That review resulted in a report, identified as milestone MS26 in ICEDIG project. That report had 3 different purposes:

- Help decision making for the design of the DiSSCo infrastructure
- Assist institutional curators, even if there are not technical-savvy, to make a choice among existing platforms for a transcription project
- Highlight keys for success and important features for anyone planning to design a new transcription website

That report has a special section "Recommendations for DiSSCo services" that compiles the main conclusions for DiSSCo:

1) There is no "Best platform", a web tool that outperforms all others in all features;
2) The major asset of each site is its community;
3) Different features, languages, scientific interests and gamification mechanisms attract different people across Europe;
4) So DiSSCo should not offer a "DiSSCo volunteer platform" but instead mobilize actual and future platforms to document EU collections;
5) The ICEDIG design study should focus on how to integrate the diversity of platforms in a common workflow;
6) Implementation of that workflow requires interoperability between digitization lines, collection management systems and label transcription platforms;
7) The specifications of data flows are a key to achieve that interoperability and we should pay special attention to the specifications design in ICEDIG deliverable 5.1;
8) Integrating CS activity in the future DiSSCo Dashboard could be a powerful incentive for volunteer mobilization.

This report was presented and discussed during the ICEDIG All Hands meeting. To address the recommendations 5, 6, 7 and 8 we wrote a technical specification in order to standardize dataflows:

- From digitization line to transcription platform;
- From transcription platform to collection management system;
- From transcription platform to DiSSCo dashboard;

This specification is identified as milestone MS28 in ICEDIG project.

The part of the specification that covers collection management system was a close collaboration with WP4 that addresses “Interoperability with institutional collection management systems”

In parallel we populated the DiSSCo GitHub repository with source codes of significative transcription platforms. One of them, “Les herbonautes” was not openly available before.

This specification is identified as milestone MS27 in ICEDIG project.

We tried to keep the specification easy to understand even for collection managers with limited software skills. We illustrate the specification with several examples. One of them was built with the data of “ICEDIG mission” from WP4. All illustrative datasets were published on GitHub repository, so it is included in milestone MS27 deliverable.

3. Milestone MS26: Evaluation of existing volunteer transcription systems

Report is available on GitHub:

https://github.com/DiSSCo/transcription-platforms/blob/master/Evaluation_of_2018_volunteer_transcription_systems(ICEDIG_MS26).pdf

and annexed to the present document.

4. Milestone MS28: Specification of data exchange format for transcription platforms

The document is publicly available on Zenodo repository:

https://doi.org/10.5281/zenodo.2598413

and annexed to the present document.

5. Milestone MS27: Repository of source code of transcription websites

The source code is available on Github:

https://github.com/DiSSCo/herbonauts

https://github.com/zooniverse

https://github.com/AgentschapPlantentuinMeise/volunteer-portal

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

REFERENCES:
MILESTONE MS26

Work package **WP5** / Task **5.2** Lead: **MNHN**
Deadline M7

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

EVALUATION OF EXISTING VOLUNTEER TRANSCRIPTION SYSTEMS

MILESTONE MS26

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

ICEDIG.EU

Introduction

As a result of modern natural science having been developed in Europe, numerous institutions hold and curate important collections both with regard to their age and their size. The scientific and cultural value of these collections are considerable and digitisation is a major challenge to improve access for researchers and the general public. In the last decade, the digitisation effort has started involving the "crowd". An increasing rate of digital imaging and label transcription, partly due to this recruitment, has increased uses of these collections by opening the collections to a broader audience. These uses became as well more diverse, not just for science, but as well for its cultural aspects.

European institutions holding natural history collections have made use or have developed different platforms. The first transcription platform was *Herbaria@Home* (http://herbariaunited.org/atHome/ - Figure 1) launched in 2006 by the Botanical Society of Britain and Ireland to help digitise specimens from British and Irish collections. Shortly after, in 2007, *Zooniverse* was created (https://www.zooniverse.org/ - Figure 2). Initially it was designed for astronomical and meteorological studies and has become, after a little more than a decade, the major cloud-like platform for citizen science (CS). Major Europe-based institutions have engaged projects either directly on *Zooniverse*, either on the associated platform dedicated to natural history collections transcription *Notes from Nature* (https://www.notesfromnature.org/). These institutes include the Botanical Garden and Botanical Museum of Berlin (BGBM), the Royal Botanic Gardens, Kew (RBGK), the Natural History Museum of London (NHM) and the Manchester University Museum. In 2011, under the umbrella of the Atlas for Living Australia (ALA), *DigiVol* was launched (https://volunteer.ala.org.au/ - Figure 3). Initially designed for the needs of understanding Australian biodiversity, it has become a broadly used citizen science tool, used by the NHM, RBGK and the Royal Botanical Garden of Edinburgh (RBGE), among many others. In 2017, based on *DigiVol* code, *DoeDat* (https://www.doedat.be/) was launched by the Meise Botanic Garden. Following the mass digitisation of the French National Herbarium in Paris *Les Herbonautes* was launched in 2012 (http://lesherbonautes.mnhn.fr/ - Figure 4). As *Herbaria@Home*, and unlike other systems, it was specifically designed for herbarium specimen label transcription, and now processes specimen images from herbaria from all the French network of Herbaria. Although not tested yet, the possibility of including other natural history collections is considered. In 2017, based on the code of *Les Herbonautes*, BGBM launched Die Herbonauten (https://www.herbonauten.de/). Although not launched yet, an English-speaking version has been under consideration by the RBGE. More details about these platforms can be found in Table 1, and a comparison of features of each in Table 2.

The Smithsonian Institution also uses its own platform, The Smithsonian Transcription Center (https://transcription.si.edu/), which has become a major actor of the sector. However, the use of it is reserved to this institution and doesn't directly concern the public of this report. Aside from these main platforms, different projects involving Natural Science objects were conducted such as a project on glass slides using the Dutch cultural heritage platform *Velehanden* (https://velehanden.nl/) (Heerlien et al. 2015).

Livermore and his co-workers (2015) wrote a review of the major crowdsourcing platforms mentioned above as part of the *Synthesys* project. It can be referred to for more detailed descriptions of each platform. The present report was largely based on a study made by Ellwood and her co-authors (2015), in the scope of the *iDigBio* project. It is aimed toward helping European institutions who are considering using crowdsourcing in their digitization effort. As it is generally better to adapt and improve existing solutions, rather than to start from scratch, this report presents the important issues to keep in mind when considering a CS based transcription solution.

Code for setting up such a solution has been made available through the code sharing facility GitHub. At the time of publishing this document *DigiVol* code is available (https://github.com/AtlasOfLivingAustralia/volunteer-portal), as well as its internationalized derivative *DoeDat* (https://github.com/AgentschapPlantentuinMeise) and *Zooniverse* (https://github.com/zooniverse). *Les Herbonautes* code will be shared through the *DiSSCo GitHub* account (https://github.com/DiSSCo/herbonauts) in the next few weeks.

Recommendations for DiSSCo services

Despite the specifications of future *DiSSCo* services and architecture being beyond the scope of the present deliverable, the evaluation of existing volunteer transcription systems already leads us to some conclusions regarding *DiSSCo* infrastructure:

1) There is no "Best platform", a web tool that outperforms all others in all features
2) The major asset of each site is its community
3) Different features, languages, scientific interests and gamification mechanisms attract different people across Europe
4) So *DiSSCo* should not offer a "*DiSSCo* volunteer platform" but instead mobilize actual and future platforms to document EU collections
5) The ICEDIG design study should focus on how to integrate the diversity of platforms in a common workflow
6) Implementation of that workflow requires interoperability between digitization lines, collection management systems and label transcription platforms
7) The specifications of data flows are a key to achieve that interoperability and we should pay special attention to the specifications design in ICEDIG deliverable 5.1
8) Integrating CS activity in the future *DiSSCo* Dashboard could be a powerful incentive for volunteer mobilization

Where to start from?

This report tries to give comprehensive information about CS transcription platform. A read of this document and other documentation such as *Synthesys* last report on the matter (Livermore et al. 2015) is important to get an overview.

Prior to start setting an actual website it is best trying setting projects on some existing platforms in order to get familiar with the running of such a project and get guidance from the platforms teams. There is no best solution to our opinion. Everything depends on what project designers expect. The choice of one solution rather than another has to be done depending on platform language, possibilities of annotation, data format etc. For more information about each platforms asset, cf. Table 1.

The most important part of a CS project is its community. Community management, build up and communication is the key to a successful project. We suggest it is best to use existing source codes, eventually improving them. The codes for *Zooniverse*, *DigiVol/DoeDat* and for *Les Herbonautes* are available on GitHub. *DigiVol* and *Les Herbonautes* have already been successfully adapted by several platforms.

Recruiting and keeping Volunteers

CS platforms in general have proved their ability to mobilize an efficient transcription audience. It is then of key importance to better understand which users we are going to address for the documentation of natural history collections.

CS projects have begun to become well documented (Raddick et al. 2010, Rotman et al. 2014, Zacklad and Chupin 2015, Geoghegan et al. 2016, West et al. 2016, Chupin 2017, Lee et al. 2017). Although few studies have been done on transcribing biodiversity collections tools they corroborate trends and results from those global studies. All these studies paint a similar picture of how to interpret the general features that are found in our users' communities and especially to develop effective ways of recruiting and keeping them.

Overview

Natural History Museums have several missions, which range from scientific collection management to public awareness of biodiversity. CS platforms address both these missions of conservation and outreach to the general public. For this reason, aside of being a transcription tool, our CS platforms are also a way of displaying our institutional collections and their uses. As such, these platforms should be considered as tool to display our collection richness before being seen as tool to enrich them.

A key step in setting up a CS project is to advertise it in order to build up a community. A survey study conducted in 2014 by Chupin (2017) on *Les Herbonautes*' volunteers identified and categorized the ways the platform was discovered by users. The most effective way to recruit volunteers were shown to be actions done by the project staff, such as newsletter articles shared in an existing network (i.e. *Tela Botanica*, a French well established non-professional botanist network), or oral presentations at meetings. This type of recruitment proved to reach the most people and had the longest impact, as it reached a specific public who were potentially interested. Another effective way to recruit was through press and radio probably as a result of its broad audience. On *Les Herbonautes*, an important amount of the still active major volunteers has been recruited through newspapers. Newspaper articles are an advertising medium not to neglect. Television, on the other hand, did not prove to be very effective. Another mean for recruitment explored by the study was serendipity (i.e. a thread shared on social media). In addition to being difficult to control, this medium showed mixed results in the case of *Les Herbonautes*. Most people recruited through social media just went on a tour and didn't really take part. Finally, a small number of users were recruited by word of mouth, although this is not a reliable method to count on.

Another effective way to recruit proved to be coorganized CS events , for instance *WeDigBio* (Ellwood et al. 2018). These events proved to be effective on productivity during the event, but it also boosted volunteer interest and recruitment of new users. At a smaller scale the Meise Botanic Garden organised a *transcribathon* on Thursday 17th May 2018 to get to know their user community. 17 users took part in the day, transcribing over 1000 records and having a tour behind the scenes of the herbarium and a walk in the garden afterwards. The event showed encouraging results and *DoeDat* staff at Meise Botanic Garden plan to organise this twice a year, on a 2-day event basis. A survey (personal communication) held at the end of this day confirmed most of the trends mentioned above, and that the attendees clearly mentioned they were awaiting such events.

Major trends on the CS platforms users transcription communities can be distinguish (Raddick et al. 2010, Tweddle et al. 2012, Rotman et al. 2014, Livermore et al. 2015, Zacklad and Chupin 2015, Geoghegan et al. 2016, West et al. 2016, Chupin 2017, Lee et al. 2017). Most of them fits as well for other CS users' communities.

As well as for label transcription projects as for CS in general, people taking part into projects tend to be mature (typically retired) and have an educated background. Although tested by several studies, the distribution on income level doesn't showed clear tendencies that can be extrapolated to all communities.

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

Gender distribution of the users tend to be in favour of men. However, we are not aware of studies with less than 47% women and wonder if it could be explored whether men don’t tend to respond more to survey than women.

On every CS project, most of the work is done by a small minority of participants. It is very important for a CS platform manager to keep this in mind and manage the platform in order to attract these power users and keep them engaged.

Motivations to take part in a CS project are often multiple and can change through time for a single user. It’s rather difficult to map it. However, main tendencies can be distinguished, that are common for all CS projects. Helping and contributing to sciences and biodiversity/environment knowledge is always the main motivation, alongside with an interest for the subject of the project (botany for the CS transcribing platforms tested). Learning and curiosity comes next, alongside with having fun and compete with other contributors (to have more contribution on a project).

A user-friendly interface and its responsivity play an important role in keeping the users motivated, but as much important is the support and feedback around the mission. A deficiency in one of these elements can lead to a quick participation drop-off.

Best practices and standards

- **Use different media to reach new participants**. Studies proved CS users to have been recruited by different media. It is appearing important for a new CS transcription project to use a wide range of advertisement medium.
- **Communication on site and newsfeed.** Communication with the participants is a very important tool to keep the project going. Encouraging messages sent while the project is running are very important to keep the interest of the users. The citizen scientists’ interest to the subject is also something that needs to be taken into account. Lee and his coworkers (2017) and West and her coworker (2016) are suggesting few directions to follow and take into account in CS community management.
- **Forums to enable volunteers to communicate** with one another and with project staff about specific specimens or ledgers or the general process of transcription to the project manager and each other should be provided.
- **Value scientific usage of transcribing.** A very common demand from the CS users is to get feedback over what their contribution has been used for. Feedback gives them a sense of collectivism. Although this is time consuming for the project staff, it appears to be an important trigger to ensure long term contribution. Events onsite such as *WeDigBio* and Meise’s *Transcribathon* allow easy possibility to value scientific usage of user’s activity.

- **Use gamification, but not without moderation.** Gamification is a very important leverage tool broadly used by different CS platforms to boost contributions by the community (Eveleigh et al. 2013, Greenhill et al. 2014). However, experiences on *Zooniverse* has shown that strongly enhanced competitive gamification can be really counterproductive, leading users to resign from the project (Eveleigh et al. 2013). Possibility to competition should be given, but not become the only trigger.
- **Make it easy to start.** One of the main reasons for a to-be user not to participate to the transcription, in the case of people taking the time to answer an online survey on the subject, is the impression they do not have the basic knowledge to participate (Chupin 2017). Therefore, important pedagogical effort is to take place during recruitment to emphases on the fact that no prior scientific knowledge is required other than basic web browsing skills.
- **A good training is a fun one.** Projects which require participants to undertake training, such as transcribing platforms, appear to have higher submission rates. Although the trainings seem to be taken by the user as "the non-fun part" of taking part to the projects, the presence of a training seems to lead to their engagement (the project seems more serious, and it is a way to learn, which is one of the commonly shared motivations). Gamification of the training is then a good way to reconcile these two aspects.
- **A task completion count** should provide the public participant with both progress towards the projects digitization goal and the participants overall contributions to the project.
- **Provide users with a summary page**. This page allows users to overseas their actions and eventually their rewards (Figure 5). Allowing other users to see the others user page is a good trigger for those seeking competition and allow user to better identify who their communicating with on the forum. Moreover it can allow user to scroll around their previous action, and eventually amend it.

Chupin's 2014 study (2017) on *Les Herbonautes* community led to the establishment of best practice for the platform community leading and the project *e-ReColNat* board (in French).

The *European Citizen Science Association* (ECSA) website aggregate as well an important amount of guidelines for CS projects (https://ecsa.citizen-science.net/blog/collection-citizen-science-guidelines-and-publications). Although these guidelines are broader than only transcription of natural history specimens, they are still useful when you want to set up a CS project on natural history collections.

Gaps in our knowledge and areas for improvement.

Organisation of specific events has a potential for boosting participation. However, our knowledge is limited to *WeDigBio* event and Meise first transcribathon.

WeDigBio events have had little impact on *Les Herbonautes* (Ellwood et al. 2018). This is most probably due to both a language issue, as the other platforms to take part to the event were English speaking ones, and a lack of actual physical events that took take place in France. *WeDigBio* events are set in English, and it is expected that few from *Les Herbonautes* users are English speakers or feel comfortable with it. Moreover, the platform is not accessible in English. Translation of labels into French is actually an action *Les Herbonautes* users doesn't seem to be fancy with (Chupin 2017). An area of improvement, especially crucial for European platforms, would be the organisation of such events on a multilingual scale. These events showed as well to improve boundaries between the different user communities (Ellwood et al. 2018), and an improvement in collaboration to set up these action in Europe would benefit everyone.

We are aware of some active users on *Les Herbonautes*, who are also active on *Die Herbonauten* or on *DoeDat*. However no formal studies on the relation between different platform communities have been made so far to give a complete image of the communities' bonds. This could help to better understand communities, and the possible impact of events such as *WeDigBio*.

Volunteers can valuably take part into peripheral task such as community management. The forum (Figure 6) linked to each specimen and discussions that occur around cross checking on *Les Herbonautes* and *Die Herbonauten* for example, allow the users to share their knowledge. Volunteers can as well take part in the recruitment. This helps considerably the management team.

Another important step would be having the possibility to address user samples to citizen scientists in their own language. This would however require presorting images per language and assembling them in a repository. Work package 4 is exploring this matter amongst many others. However, to keep attractivity for the users to take part, we believe the platform should avoid sorting the image through countries. One of the attractive things for users is to learn about other countries, although it is strongly suspected that they are e more efficient to geolocate a location in their own country (to be explored in task 4.2), setting projects only about their country would be less attractive to users.

Online activity 1: Transcribing specimen label and ledger text

Ellwood and her co-authors (2015) recognize two processes from Dunn and Hedges' (2013) typology in Online activity 1 : transcription (creating machine-readable text that reflects the textual content of the specimen label or ledger; sometimes called text encoding) and cataloging (the production of structured, descriptive metadata about the text). We will here discuss both of these processes as the activity of transcription, as is common in the biodiversity research collection domain.

Overview

To date, this activity is still most commonly completed by paid technicians onsite in one step: typing (or occasionally reading) the text into appropriate fields in institution's specimen data management system (Nelson et al. 2012). These steps have been as well industrialised and are sometimes done offsite in two steps: transcription offsite by professional as from an image of the specimen on a dedicated database, the second step consisting in data integration on the institution management system mostly by IT crew. In both case, the technicians have been trained to systematically catalog the often complex and variable labels and ledgers found in the concerned biodiversity research collection. CS, however, has taken more and more place in the process lately alongside with the development of semiautomated tools.

Aside to human made transcription, different semi-automated solutions using optical character recognition (OCR) have been tested and are still under testing. They will be explored further on task 4.1 (deliverable 4.1 due 31/01/2019, interim report due on 31/07/2018). OCR creates non-structured text being an imperfect transcription. However, two methods using these imperfect transcriptions can be distinguish as concerning CS. A first method is to use the bulk results as a pre-sorting tool for further uses, in particular for CS mission/expedition design. This has been made at the MNHN, using Tesseract-OCR, and is currently being used to give more possibilities on designing missions on *Les Herbonautes* (i.e. selecting images of specimens collected by a single collector as for the mission *Eugène Poilane* http://lesherbonautes.mnhn.fr/missions/5090704). A second method consists in digesting the bulk data with one or several algorithm and allow users, to structure the text (Barber et al. 2013, Ellwood et al. 2015). Although this hasn't been tested yet, to our knowledge on CS site, it is a considered evolution by teams developing it, in particular by teams working on *Zooniverse*.

As mentioned above, public participants can be expected to be most efficient and accurate at the transcription activity when they are proficient typists and can read the language in which the label was written (Ellwood et al. 2015, Chupin 2017). Personal attributes that also benefit any of these digitization activities include attention to detail, patience, dedication, and a desire to make a difference or contribution. Useful emphases in training for the task can be placed on skills relevant to the basic understanding of specimen labels such as interpreting common scientific jargon, abbreviations, label formats, and variability in dates (ordering of month–day versus day–month in different cultures), as well as standard markup for capturing annotations, deletions, and markings in the original text. Equally important is training in how to handle label information that requires further judgment such as when to type the element verbatim and when some interpretation may be used (e.g., when common words are misspelled), how to handle inconsistencies (e.g., when the city given is not found in the state given or country names that have changed over time), and identifying targeted data elements and selecting the appropriate element when multiple similar elements exist (e.g., from among the scientific names on the original label and later annotation labels). A set of specimen labels or ledger entries can vary substantially in legibility, information content, and consistency, and training examples need to adequately represent that variation.

An efficient tool to help the volunteers address these issues, alongside with training, is forum thread linked automatically to the specimen as on *Les Herbonautes.* Although this function is going to be used mostly by few users (Chupin 2017), when a reading issue occur for a specimen, a discussion will often be started, helping less experienced user.

The main platforms allowing specimen transcription have many similitude. All of them displaying the image together with some or all the fields to fill. Differences can however be observed (Table 1). Most of them are gathering the tasks into subprojects (called projects on the *Zooniverse*/*NfN*, Expeditions on *DigiVol*/*DoeDat* and missions on *Les Herbonautes*/*Die Herbonauten*), most of them uses incentives although in slightly different ways. The main differences occur in the number of fields displayed at a time on the page, the validation of the entries and the ability to discuss tasks with reference to a single specimen.

Best practices and standards

- **Make the specimen visible while typing.** Data entry fields should be accessible whilst viewing the image.
- **The image viewer should allow an easy reading of text.** The image display should produce a clear view of all relevant text at an appropriate zoom level at once or via panning.
- **Drop-down lists should be provided** when the universe of acceptable responses can be populated from controlled vocabularies and is relatively small (e.g., the 50 US states); autocomplete functionality in free text fields should be provided when the number of acceptable responses is larger and cannot be fully populated from the beginning of the project (e.g., collector names).
- **Dependencies in the acceptable values for fields should be built in** (e.g., only those counties from the state of Georgia are available in a dropdown once the state is established as Georgia).
- **The content of autocomplete lists should be maintained regularly** Proposing obsolete or erroneous value make the lists counterproductive (e.g., French regions updated after 2017 administrative changes or botanist’s names filled in with space character at the end appearing several times).

- **Readily accessible examples and directions for each field** should be available during the activity.
- **Response and loading time of images and transcription pages should be quick** as users can be located even in remote areas with low internet access. Long loading time will lead to volunteer disengagement.
- **Permit transcribers to explore the portion of the image containing the organism** or view an image of the taxon from another source (e.g., *Notes from Nature*'s Macrofungi Interface displays images of the taxon from Encyclopedia of Life).

To our knowledge, there are not best practice documents specifically targeted at engagement of the public in transcription for biodiversity research collections. However, there are best practices for specimen imaging that must occur to permit online transcription and annotation (Häuser et al. 2005). Most of institutes have their own best practice relevant to their specific databases, and there are best practices that are generally relevant to the digitization activities identified in Dunn et Hedges (2013), such as DataONE's Primer on Data Management (http://dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf) and the online *Citizen Science Central Toolkit* (http://birds.cornell.edu/citscitoolkit/toolkit/steps).

Relevant sources of standards for this activity and, to some extent, the other two include the Dublin Core Metadata Initiative (http://dublincore.org), the Darwin Core for biodiversity information (http://rs.tdwg.org/dwc; Wieczorek et al. 2012), the Audubon Core for metadata about multimedia files associated with biodiversity research collections and resources (http://tdwg.org/standards/638), and the Ecological Metadata Language project (http://knb.ecoinformatics.org). Specific to markup text in the humanities is XML-TEI markup (http://tei-c.org/index.xml), which is important in the context of transcribing ledgers. A standard recommendation for data exchange format will be address by February 2019 as a deliverable of task 5.2.

Gaps in our knowledge and areas for improvement.

Improvements to transcription tools could enhance participant enjoyment and ease of use.

As mentioned above, an improvement could be a broader use of OCR results. OCRisation of collection prior to their integration into a CS project could improve the volunteer’s experience. Aside to allow better sorting of the specimens to be transcribed in a mission/expedition, it would as well allow further functionality development as suggest by Ellwood (2015). For example, new functionality could give the contributor more control of their transcription experience, such as providing them with the ability to establish the criteria used to determine the specimens that they transcribe (e.g., on the basis of the collection supplying the specimen images or the occurrence of a word in the OCR text strings generated from images) or the ability to toggle between interfaces that show a single field at a time and multiple fields at a time. Furthermore, records could be sorted for transcription based on similarity (e.g., overall similarity of OCR text strings). OCR results, processed through a language detection tool and with collaboration between the platform based on their linguistic particularities could allow to efficiently answer the language issues.

The establishment of a structure such as Herbadrop (https://b2drop.eudat.eu/s/QqPv9epgNiosxBR#pdfviewer), linking an OCR digest to specimen eligible to CS transcription could only be benefiting the CS operations.

Improvements could also address data quality issues by providing the ability for participants to return to earlier transcription records to correct what they later learn are transcription errors. The biodiversity research collections community would also benefit from greater sharing of best practices and tools with the digital humanities community, for the comparison of multiple transcriptions of a single text, represent significant overlap in objectives between the two communities.

To date only the *Zooniverse* have been developed as an smartphone/tablet application (Livermore et al. 2015). Initially it was mainly due to an issue of readability of the labels on the image. However this has become less and less relevant with the growing importance of the tablets and the phone screen becoming bigger and bigger. Development of phone application could then give new access to volunteers and allow to reach new public.

Online activity 2: Georeferencing

Georeferencing, as applied to biodiversity research collections, is the inference of a geospatial geometry from the textual collection locality description on a label or in a ledger (Guralnick et al. 2006). It is the first basic interpretation of label information asked from CS users. As such, it need a bit more knowledge and training than transcription. This task includes coordinates imputing, but as well input of geographical controlled vocabulary, as this can be linked to a polygon on a map.

Overview

The geospatial geometry is often expressed as a single point representing latitude and longitude, usually with an associated radius allowing representation of uncertainty (Wieczorek et al. 2004 - Figure 7). However, localities could also be represented as multipoints, lines, multilines, polygons, and multipolygons to better reflect either the collection method or imprecision associated with the interpretation of a textual collection locality description. For example, sampling transects may be recorded as a line with start and stop coordinates, as is common in samples from trawlers. The expression of uncertainty is crucial to determining a data record's fitness for use (Wieczorek et al. 2004). For example, point data with an uncertainty of 10 km may be unsuitable for an analysis across 1-km-resolution environmental gradients. Georeferences as latitude and longitude coordinates and the datum on which the coordinates are based are typically lacking from terrestrial and inland aquatic specimens collected before the 1990s (marine specimens might differ). Where those are available, they can provide useful validation for textual descriptions or vice versa, because such latitude and longitude readings also have associated, and often unreported, uncertainties.

To note that the older the specimen, the more difficult the georeferencing, mostly because of lack of information, but as well because of geographical vocabulary evolution of term through the ages. This is of crucial importance as the European collection of natural history holds an important amount of old specimens, reflecting biological sciences history (Le Bras et al. 2015, 2017, Papastefanou et al. 2016, Monteiro et al. 2017, Nualart et al. 2017, Silva et al. 2018).

Public participants can be expected to be most efficient and accurate at georeferencing when they can read the language in which the label was written, can read relevant map types (e.g., topographic or nautical), and have some familiarity with the area in which the specimen was collected (i.e., experience on the ground or with locally used names). Useful emphases in training for the task can be placed on basic geographical skills such as identifying the locality information and interpreting locality types, interpreting geographic jargon, compass bearings, abbreviations, and formats, and understanding the common types of geographic projections (e.g., equal area), coordinate systems (e.g., Universal Transverse Mercator) and geodetic systems (e.g., World Geodetic System 1984). Training will also improve a participant's ability to interpret locality descriptions and uncertainties. For these skills, training emphases can be placed on finding and using relevant maps and indices of place names, and precisely describing the georeferencing method in a standard way, using known sampling biases to interpret locality descriptions (e.g., the tendency to collect near existing roads), and describing uncertainty quantitatively (e.g., as the radius of a circle) or using other geometries (e.g., a polygon). An understanding of the historical context and relevant training in interpreting the -patterns in historical aerial photographs that are relevant to predicting the community type at alternative locations (e.g., swamp versus upland) is also helpful. The extent to which the training is needed will vary depending on the locality descriptions. For example, the description “Pushepatapa Creek, 7.8 miles north of Bogalusa at Hwy 21; Washington Parish; Louisiana” requires very little expertise to pinpoint, because it is at the intersection of a bridge and a creek. However, the description “San Francisco Bay, Shag Rock, S. 58° W, Rt. Tang. Pt. Avisadero, S. 74° W., Goat Island. Lighthouse, N. 21°W.; United States” requires an understanding of compass bearings and reading navigational charts (examples from Ellwood (2015)).

Best practices and standards

- **Show a map.** While georeferencing, people often need to refer to a map. To have access to a mapping tool is of key importance.
- **Categorize precision when georeferencing a locality name.** In order to produce precision in this activity, users need clearly differentiate fields for geographical entities (e.g. country, region/state…)
- **Closed lists of geographical entities depending on upper geographical entities.** Once entered an upper level geographical name, such as a country, a controlled list of region/state should be provided in a dropdown list.

Best practice documents specific to georeferencing specimens include Guide to Best Practices for Georeferencing (Chapman et al. 2006), Principles and Methods of Data Cleaning—Primary Species and Species-Occurrence Data (Chapman 2005), and Guide to Best Practices for Generalising Sensitive Species Occurrence Data (Chapman and Grafton 2008). However, the geospatial community has produced many other best practice documents, including those related to standards (e.g., as at the Open Geospatial Consortium; http://opengeospatial.org/standards/bp) and commercial or open-source geographic information systems (e.g., as found at ESRI; http://esri.com). A useful clearinghouse for information about the process of georeferencing specimens is provided by VertNet (http://vertnet.org) at http://georeferencing.org.

We are unaware of best practice documents produced to address public participation in the generation of geospatial data. However, on the basis of the experience of developing GEOLocate and implementing tools in projects such as VertNet (http://vertnet.org), Ellwood and her co-authors (2015) address several considerations that are important to successfully engage the public in this activity. The categorization of data records into administrative unit of specimen origin (e.g., country, state, county) is useful for assigning records to public participants; a user survey can provide information regarding on-the-ground knowledge for alignment with the specimen localities. Classification of georeferencing difficulty (using, e.g., the uncertainty that GEOLocate automatically assigns) is useful for assigning records as well; a participant's performance with control localities (where accurate coordinates are known) can be used to evaluate georeferencing skill. Each locality record should be georeferenced multiple times until the points reach some clustering threshold (a predefined spatial variance) or the replicates reach a limit, at which the record is flagged for the attention of an expert. Recommendations made for transcription best practices are also relevant here, especially provision of a forum for users to discuss specific localities or general patterns with each other and project scientists, leading to greater user proficiency and understanding.

Relevant sources of standards for the generation and communication of geospatial data include the the Open Geospatial Consortium (http://opengeospatial.org), and within Darwin Core (i.e., DC-location), as well as most of those presented for transcription.

As for the transcription tasks, forum linked to the specimens proved on *Les Herbonautes* to help better consistency in the geolocation of the specimens.

Gaps in our knowledge and areas for improvement.

We do not have a satisfactory understanding of several aspects of public participation in georeferencing, including the average number of replicate georeferencing events needed to reach a sufficient level of accuracy and effective methods for balancing accuracy and precision (e.g., by removal of outliers) to produce a useful consensus georeference. In particular, we lack the understanding over the abilities for a user match georeferencing competencies with collection localities and we lack sufficient strategies for assessing a user's georeferencing competencies, initially and through time. A better understanding of how to enable collaboration and communication (e.g., by visualizing on a map the collection localities being discussed in a forum) is also needed.

Digital imaging and linking of field notes to specimens would likely provide a big benefit to georeferencing, because field notes can contain a wealth of information about collecting sites, including travel itineraries, site sketches, environmental information, and other remarks not often found on specimen labels. Although not based on CS, the Saint-Hilaire virtual herbarium (Pignal et al. 2013) have shown feasibility of linking field notes book to herbaria. CS remain based project remain for the time being to try. The biodiversity research collections community would also benefit from greater sharing of best practices and tools with other communities, including the ecological CS projects that enable mapping of species observations (e.g., National Geographic's FieldScope project, http://education.nationalgeographic.com/education/program/fieldscope, and iNaturalist, http://inaturalist.org), digital humanities projects that rectify digital images of historical maps (e.g., Map Georeferencer, http://maps.nls.uk/projects/georeferencer/about.html, which has been used in the British Library Georeferencer Project, http://bl.uk/maps), and projects to develop “framework data” (OpenStreetMap, http://openstreetmap.org).

Online activity 3: Annotating

Beyond the label data used for the transcribing activity, and interpretation the geolocation (see above online activity 1 and 2), a wealth of additional information can be derived from the image of the specimen and shared through annotations. CS transcription facilities are design to retrieve basic human readable informations from label image to machine readable ones, consequently, annotation does not consist into the main activity. However, these platforms can be efficient tools for data enrichment.

Overview

Physical annotations traditionally were associated with a physical specimen that was visited at its home collection or examined while on loan to another collection. The most common one by far are the taxonomic identification labels (*determinavit*). In online specimen annotation, a feature of interest can be described and measured from a digital image, often with an area of interest specified, linking the annotation not only to a specimen, but a region on the specimen image. Annotations can be related to taxonomic identity, phenological state or life stage, features in existence at the time of the collecting event (e.g., evidence of disease or herbivory), damage following the collecting event (e.g., from pests), entity–quality statements (e.g., the flower is red), landmarks for morphometric analysis, and many more. Annotations are not typically a focus of the initial specimen digitization (e.g., those task clusters described by Nelson et al. (2012)) unless they are legacy physical annotations associated with the specimen at the time of digitization, but they can be fundamental to the downstream research applicability of specimens.

Augmenting specimen information with useful conclusions from the specimen image encompasses a variety of strategies and techniques that can include both automation and public participation. For example, various research projects are exploring methods for automated taxonomic identification. Similar to facial recognition applications used to identify people, these methods require an accurate training data set of identified images from one or more standard angles. These applications are widely researched (Watson et al. 2004, Francoy et al. 2008, Kumar et al. 2012, Yang et al. 2015, Kho et al. 2017, Leonardo et al. 2017, Rzanny et al. 2017, Bonnet et al. 2018, Goëau et al. 2018). Public participants take part in the development of this process by building the training data sets for these automation methods as those algorithms become more successful. Two projects examples using annotion in this goal can be found in *Les Herbonautes* mission "*Rubus reloaded*" aiming at getting an image dataset useable for training a computer over Rubus recognition leaf traits recognition (https://fr.wikipedia.org/wiki/Rubus) or the "*Project Plumage*" (Figure 8) aiming at defining polygons corresponding at morphological area of the birds to allows image analyse of birds plumage in human visible spectrum and UV spectrum (https://www.zooniverse.org/projects/ghthomas/project-plumage).

Public participants can be expected to be most efficient and accurate at annotation when they have existing familiarity with the focal taxonomic group or the focal taxonomic group within a focal geographic region, the use of authoritative resources (e.g., taxonomic keys and illustrated glossaries), and the use of relevant terms (e.g., leaves and glaucous).

Useful emphases in taxa-specific training can be placed on recognizing relevant features of the focal taxonomic group, correct usage of relevant terms, use of specific resources (e.g., a key to the millipedes of Arkansas) and the protocol for describing relevant resources and methods used for reaching the conclusion of an annotation. Process- and image-specific training can include identifying typical changes that can occur in the phenotype after preservation as a specimen (e.g., common colour changes or pest damage patterns) and typical distortions introduced by an imaging technique (e.g., deviations from a rectilinear projection or chromatic aberrations).

Best practices and standards

- **Annotation is a secondary activity.** Annotation by the CS users is a data enrichment. As such, transcription of the existing data has to be made in priority, either at the same time on the platform (as done on the *Rubus Reloaded* mission), or prior to project/mission design (as done for the *Project Plumage*).
- **Imaging techniques should take into account annotation when it is planned** or can be anticipated (e.g., many beetles are only identifiable by the number of segments on the tarsus and without that part in the image, an annotation of taxonomic identity is difficult).
- **Users should have easy access to tools for zooming and panning and designating an area of interest in the image** to associate with the annotation.
- **Use should be done of controlled vocabularies**. This to allow semantic processing and reduce misspelling.

We are unaware of best practice documents that address public participation in annotations of digital specimen images. However, best practice documents related to the creation and management of somewhat analogous annotations of images do exist in the digital humanities at Europeana Connect (http://europeanaconnect.eu; e.g., as it relates to map annotations). Ellwood and her co-authors (2015), on the basis of their experience in developing Morphbank image annotation tool, suggest several considerations to successfully engage the public in this activity as we reproduce above (the three last ones). To note that recommendations made above in reference to transcription and georeferencing best practices are also relevant here, especially provision of a forum for the users to discuss annotations with each other and project scientists, leading to greater user proficiency and understanding.

Standards relevant to annotation specifically include the relevant taxonomic codes (International Commission on Zoological Nomenclature 1999, Turland et al. 2018), the Apple Core extension of the Darwin Core (for sharing botanical annotations, http://code.google.com/p/applecore), and various controlled vocabularies that have the potential to greatly extend the value of annotations for discovery.

Gaps in our knowledge and areas for improvement.

We do not have a satisfactory understanding of several aspects of public participation in annotation including the interface design that is most suitable for capturing complex data while maintaining participants' interest and furthering science literacy goals, the accuracy rate for different forms of annotation (e.g., taxonomic identification or determination of phenological state), and the most successful methods of quality control for variable CS contributions.

To our knowledge, no CS transcription-based projects have included specimen identification by the crowd. This is considered as difficult has the users have to get an good knowledge of botany, level which is difficult to assume.

The annotation activity can potentially be improved by providing more advanced image viewing tools in the public participation sites, such as side-by-side image comparisons and transparency overlays that allow direct comparison of one image on top of another (e.g., two leaf images), more complete annotation metadata that records such information as the zoom-level and frame viewed at the time of annotation, and greater flexibility in the designation of an area of interest (e.g., using multiple polygons or edge detection or selection tools).

Conclusions

As the study of Natural History was first developed in Europe, European museums and scientific institutions holds an enormous and irreplaceable amount of information and biological collections. Considerable effort has been made in recent decades to open these collections up in order fulfil their potential, but a lot remains to do. Collection digitisation is a first step to this opening both to scientific knowledge and to a public audience. Aside from professional digitisation, CS transcription platforms have proved to be a powerful and complementary tool to increase the speed of data input speed.

Several platforms have been created to engage public participation in this challenge. It appears that the most important part of a platform lies in its community. For a platform management team, the most important jobs are building this community, training it and encourage its members. That for it is important to follow the community and try to understand it, each community being different. However, similarities can be observed with all CS communities.

The user interface and its functionalities should be considered as a tool to ensure user’s efficiency in the tasks awaited, as much as their pleasant and fun experience. Special focus should be done on geolocating tools, in order the imputed data to be computer readable, and qualitatively correct. Although not the core of the transcription activity, the annotation of the digital specimens can be a valuable activity to take place on the platform.

To be able to complete their function, CS platform should be interoperable with the collections management system. Specification of exchange will be address by April 2019 (Milestone MS28). At the time of publishing of the present document, a qualitative evaluation of the output from the different CS solution is being conducted. Output of this particular study will be published as an ICEDIG output by Deliverable 4.2.

Table 1 Online tools for public participation in transcription of biodiversity specimen labels and field notebooks. Characteristics of each are applicable to the given category. Value are valid as of May 2018 (Elwood 2015 updated)

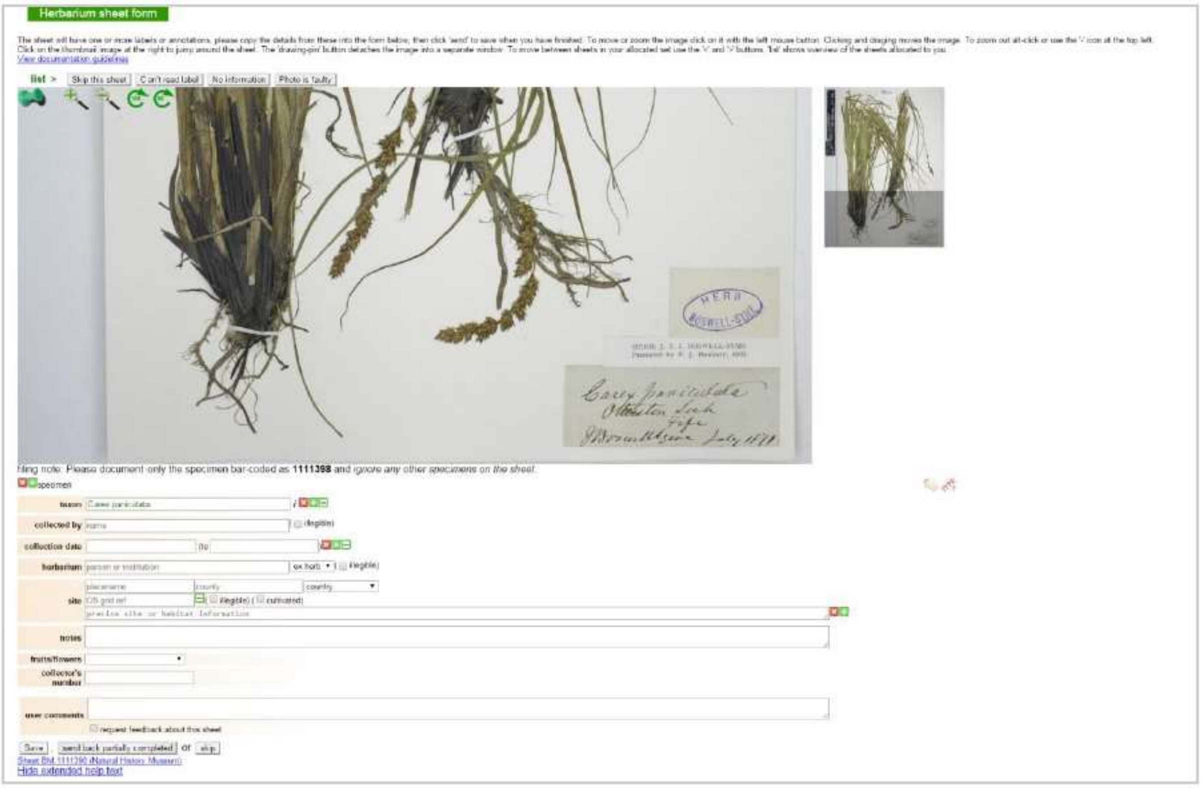
Transcription tool	taxonomic / geographic	object type focus	training	incentive	launching	contributors (single user account)	transcriptions	interface	validation process	coding language	code avalaibility
Herbaria@home	Plants / Brittish Islands	specimen labels	Online instructions and videos	None	2006	476	166 178	Zoom in on label. All fields seen at once, plant name provided, other field values provided by pull-down menu	~1% of records are cross-checked by additional participants. Data users can also make edits.	PHP	not open source. Possibility of sharing on demand.
Zooniverse / Notes from Nature	Life /global but especially USA	specimen labels and field notebooks	Onsite instruction, tutorial and forum	Badges earned upon completion of a certain number of transcription	2007	1 655 094 (all) 6 151 (NfN alone)	367 212 706 (all) 647 231 (NfN alone)	Drag box around label, label appears in window; one field shown at a time.	Four participants enter data for each specimen with postprocessing of these.	python	open on github
Digivol	Life / global but especially Australia	specimen labels and field notebooks	Onsite instruction, tutorial and forum	honour board, badges earned upon completion of a certain number of transcription, statistic board displaying all of the user action (digest in a pie chart, and raw in a table).	2011	3 152	886 658	zoom an pan in window or in separate window; all fields seen at once	each task has one transcription and one validation (proofread by an experienced transcriber).	grail	open on github
Doedat	Life but especially plants / global. Collection of Meise botanical garden mainly				2017	166	29 424		100 first contributions of a user are proofread by an experienced transcriber.		

Les Herbonautes	plants / global, collection from French herbarium network	specimen labels	Onsite instruction. Participants start with simple transcription fields (country) and are tested through a tutorial before progressing to more challenging fields	podium per mission and global (on statistic board), badges earned upon completion of a certain number of transcription, statistic board displaying all of the user action (digest on map, and raw in a table sorted per mission).	2012	3 149	3 418 857	Zoom in window; all fields seen at once.	Validation of individual fields by other participants (2 to 3), until consensus is reached. If necessary discussion is possible over the specimen.	java	to be open through ICEDIG
Die Herbonauten	plants / global but especially Europe				2017	313	262 366				

Table 2 Comparison of platform features

	Herbaria@home	Zooniverse / Notes from Nature	Digivol / Doedat	Les Herbonautes / Die Herbonauten
On site communication tools	***	**	**	***
Forum tools	***	**	***	***
Gamification	-	***	**	**
Easy starting	*	***	**	***
Training tools fun	*	***	*	**
Completion count	-	*	***	***
User page	-	*	***	**
Specimen visibility while typing	***	**	*	***
Image viewer lisibility	***	***	***	***
Drop-down list	***	***	***	***
Dependencies of drop-down lists values	***	***	*	***
Autocomplete list mantainance	***	***	***	*
Examples and directions providing for each field	**	**	***	***
Loading time	*	***	***	***
Possibility to explore the organism image	***	***	***	***
Map provided	**	-	***	**
Annotation possibilities	-	***	***	*
Versatility of CS possibilities	-	***	***	*
possibility for user to find back and correct their participation	***	*	***	**

Figure 1 Transcription interface of Herbaria@Home (image from Livermore et al. 2015)



Herbarium sheet form

The sheet will have one or more labels or annotations, please copy the details from these into the form below, then click 'send' to save when you have finished. To move or zoom the image click on it with the left mouse button. Clicking and dragging moves the image. To zoom out alt-click or use the '-' icon at the top left. Click on the thumbnail image at the right to jump around the sheet. The 'drawing-pin' button detaches the image into a separate window. To move between sheets in your allocated set use the '<' and '>' buttons. 'list' shows overview of the sheets allocated to you.

View documentation guidelines

filing note: Please document only the specimen bar-coded as **1111398** and *ignore any other specimens on the sheet.*

specimen

taxon: Carex paniculata

collected by: name □ illegible

collection date: (to)

herbarium: person or institution | ex herb ▾ | □ illegible

site: placename | county | country ▾
OS grid ref | □ illegible | □ cultivated
precise site or habitat information

notes:

fruits/flowers: ▾

collector's number:

user comments:

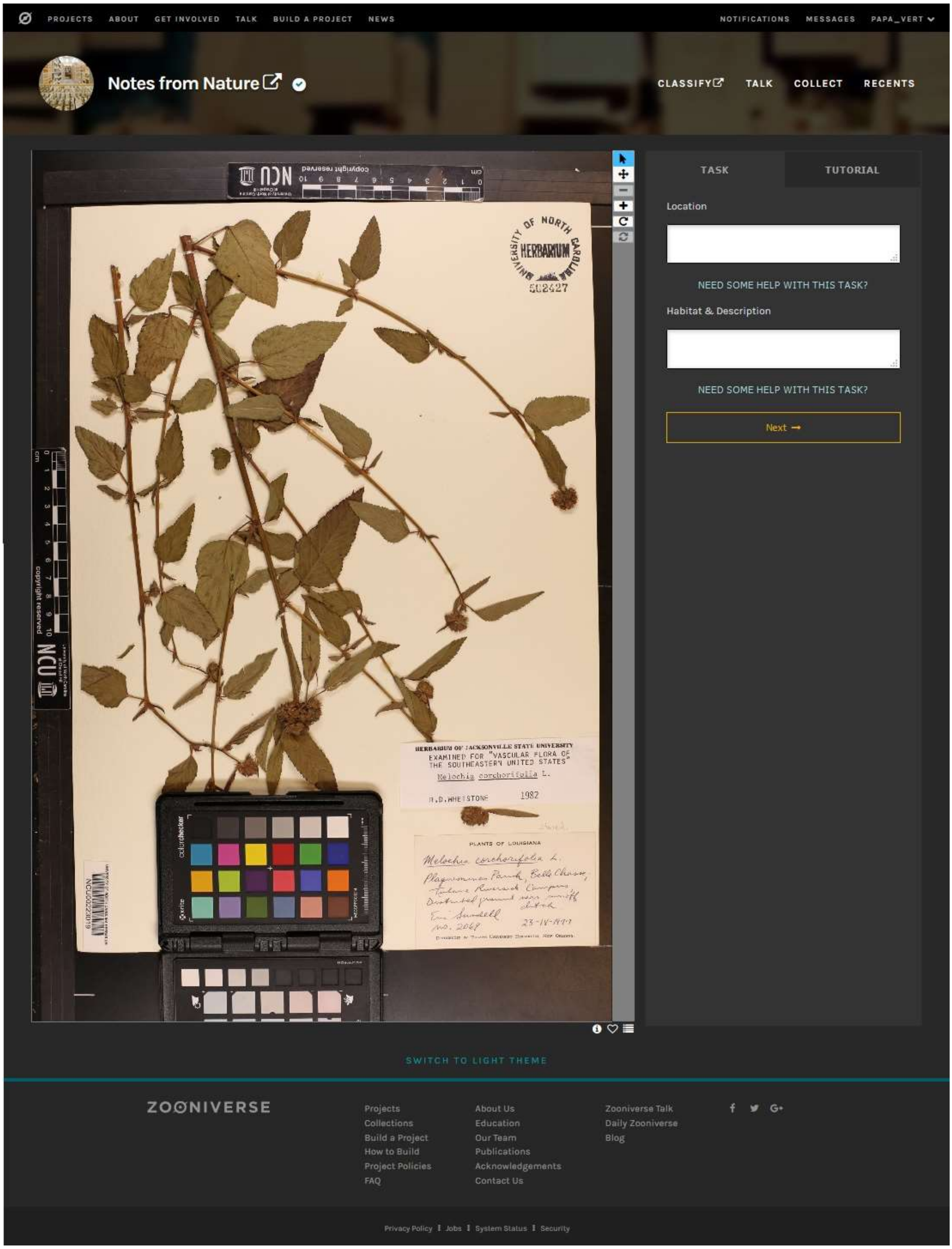
□ request feedback about this sheet

Sheet BM 111?390 (Natural History Museum)

Hide extended help text

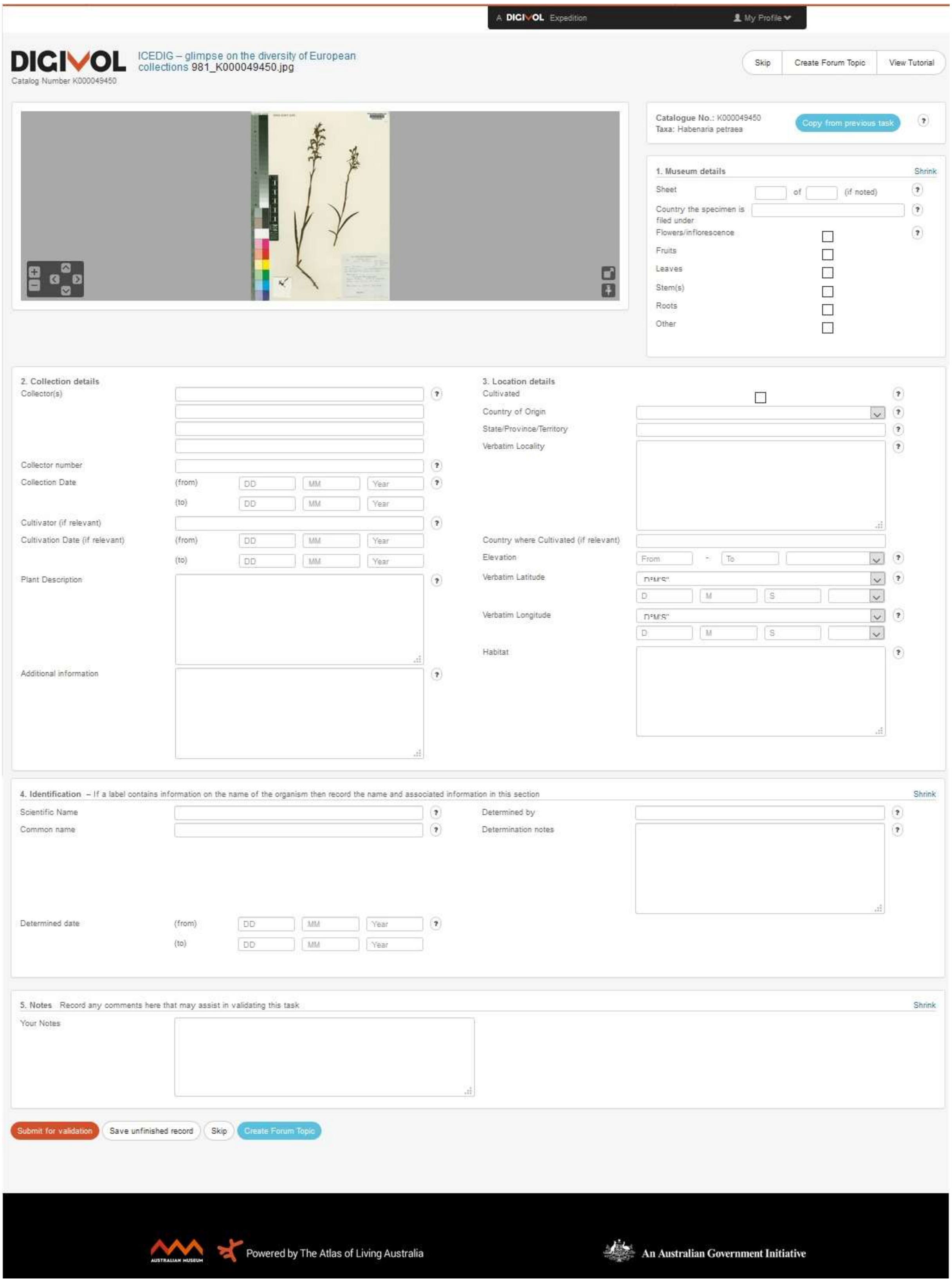
Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Figure 2 Transcription interface from Notes from Nature (part of the Zooniverse) for herbaria sheet


PROJECTS ABOUT GET INVOLVED TALK BUILD A PROJECT NEWS
NOTIFICATIONS MESSAGES PAPA_VERT
Notes from Nature
CLASSIFY TALK COLLECT RECENTS
TASK
TUTORIAL
Location
NEED SOME HELP WITH THIS TASK?
Habitat & Description
NEED SOME HELP WITH THIS TASK?
Next →
UNIVERSITY OF NORTH CAROLINA HERBARIUM
502427
HERBARIUM OF JACKSONVILLE STATE UNIVERSITY
EXAMINED FOR "VASCULAR FLORA OF THE SOUTHEASTERN UNITED STATES"
Melochia corchorifolia L.
R.D.WHETSTONE 1982
PLANTS OF LOUISIANA
Melochia corchorifolia L.
Plaquemines Parish, Belle Chasse,
Tulane Riverside Campus
Disturbed ground near runoff ditch
Eric Sundell
no. 2069 23-IV-1979
copyright reserved
NCU
colorchecker
SWITCH TO LIGHT THEME
ZOONIVERSE
Projects
Collections
Build a Project
How to Build
Project Policies
FAQ
About Us
Education
Our Team
Publications
Acknowledgements
Contact Us
Zooniverse Talk
Daily Zooniverse
Blog
Privacy Policy | Jobs | System Status | Security

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

Figure 3 Transcription interface of DigiVol (DoeDat is similar) for herbaria sheet



A DIGIVOL Expedition — My Profile

DIGIVOL ICEDIG – glimpse on the diversity of European collections 981_K000049450.jpg
Catalog Number K000049450

Skip | Create Forum Topic | View Tutorial

Catalogue No.: K000049450
Taxa: Habenaria petraea
Copy from previous task ?

1. Museum details — Shrink

Sheet ___ of ___ (if noted) ?
Country the specimen is filed under ___ ?
Flowers/inflorescence □ ?
Fruits □
Leaves □
Stem(s) □
Roots □
Other □

2. Collection details

Collector(s) ?
Collector number ?
Collection Date (from) DD MM Year ?
(to) DD MM Year
Cultivator (if relevant) ?
Cultivation Date (if relevant) (from) DD MM Year
(to) DD MM Year
Plant Description ?
Additional information ?

3. Location details

Cultivated □ ?
Country of Origin ?
State/Province/Territory ?
Verbatim Locality ?
Country where Cultivated (if relevant)
Elevation From - To ?
Verbatim Latitude D°M'S" ?
D M S
Verbatim Longitude D°M'S" ?
D M S
Habitat ?

4. Identification – If a label contains information on the name of the organism then record the name and associated information in this section — Shrink

Scientific Name ?
Common name ?
Determined by ?
Determination notes ?
Determined date (from) DD MM Year ?
(to) DD MM Year

5. Notes Record any comments here that may assist in validating this task — Shrink

Your Notes

Submit for validation | Save unfinished record | Skip | Create Forum Topic

Powered by The Atlas of Living Australia — An Australian Government Initiative

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

Figure 4 Transcription interface of Les Herbonautes (Die Herbonauten is similar)

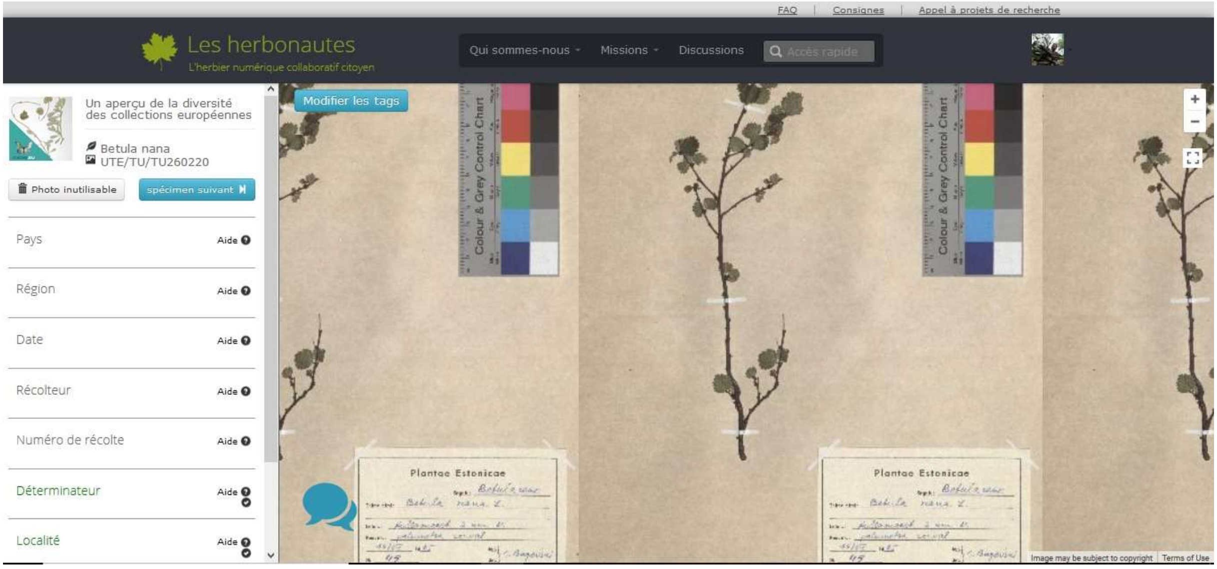
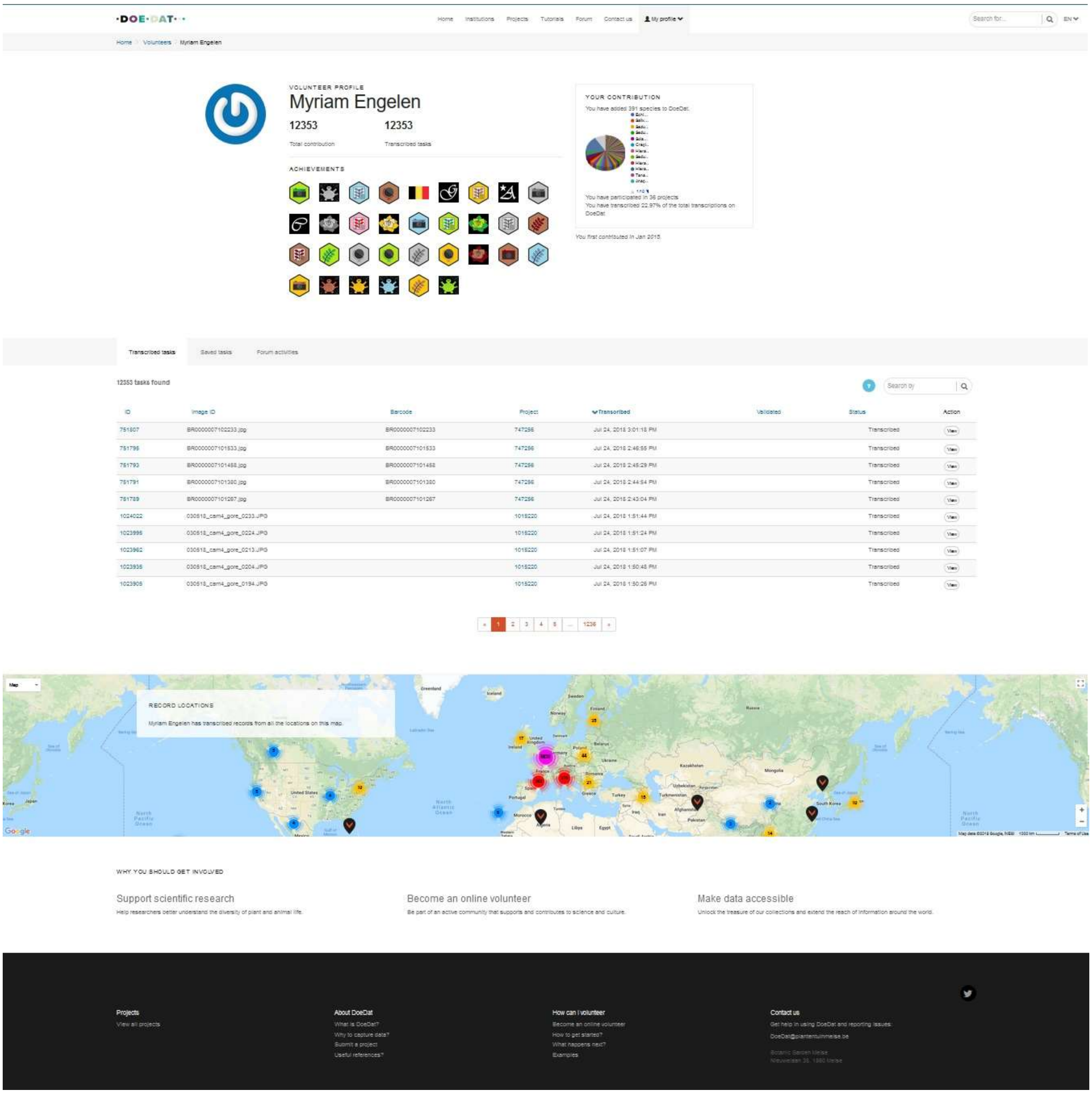


Figure 5 personal page for a user on DoeDat, displaying the rewards acquires, a digest of the data the user input in the system, and history of the users actions allowing to get back to the action, and a map of the geolocation realised



VOLUNTEER PROFILE

Myriam Engelen

12353 Total contribution

12353 Transcribed tasks

ACHIEVEMENTS

YOUR CONTRIBUTION

You have added 391 species to DoeDat.

You have participated in 36 projects

You have transcribed 22.97% of the total transcriptions on DoeDat

You first contributed in Jan 2015

12353 tasks found

ID	Image ID	Barcode	Project	Transcribed	Validated	Status	Action
751807	BR0000007102233.jpg	BR0000007102233	747256	Jul 24, 2018 3:01:18 PM		Transcribed	View
751795	BR0000007101533.jpg	BR0000007101533	747256	Jul 24, 2018 2:46:55 PM		Transcribed	View
751793	BR0000007101458.jpg	BR0000007101458	747256	Jul 24, 2018 2:45:29 PM		Transcribed	View
751791	BR0000007101380.jpg	BR0000007101380	747256	Jul 24, 2018 2:44:54 PM		Transcribed	View
751789	BR0000007101267.jpg	BR0000007101267	747256	Jul 24, 2018 2:43:04 PM		Transcribed	View
1024022	030518_cam4_gore_0233.JPG		1015220	Jul 24, 2018 1:51:44 PM		Transcribed	View
1023995	030518_cam4_gore_0224.JPG		1015220	Jul 24, 2018 1:51:24 PM		Transcribed	View
1023962	030518_cam4_gore_0213.JPG		1015220	Jul 24, 2018 1:51:07 PM		Transcribed	View
1023935	030518_cam4_gore_0204.JPG		1015220	Jul 24, 2018 1:50:48 PM		Transcribed	View
1023905	030518_cam4_gore_0194.JPG		1015220	Jul 24, 2018 1:50:26 PM		Transcribed	View

RECORD LOCATIONS

Myriam Engelen has transcribed records from all the locations on this map.

WHY YOU SHOULD GET INVOLVED

Support scientific research

Help researchers better understand the diversity of plant and animal life.

Become an online volunteer

Be part of an active community that supports and contributes to science and culture.

Make data accessible

Unlock the treasure of our collections and extend the reach of information around the world.

Projects

View all projects

About DoeDat

What is DoeDat?

Why to capture data?

Submit a project

Useful references?

How can I volunteer

Become an online volunteer

How to get started?

What happens next?

Examples

Contact us

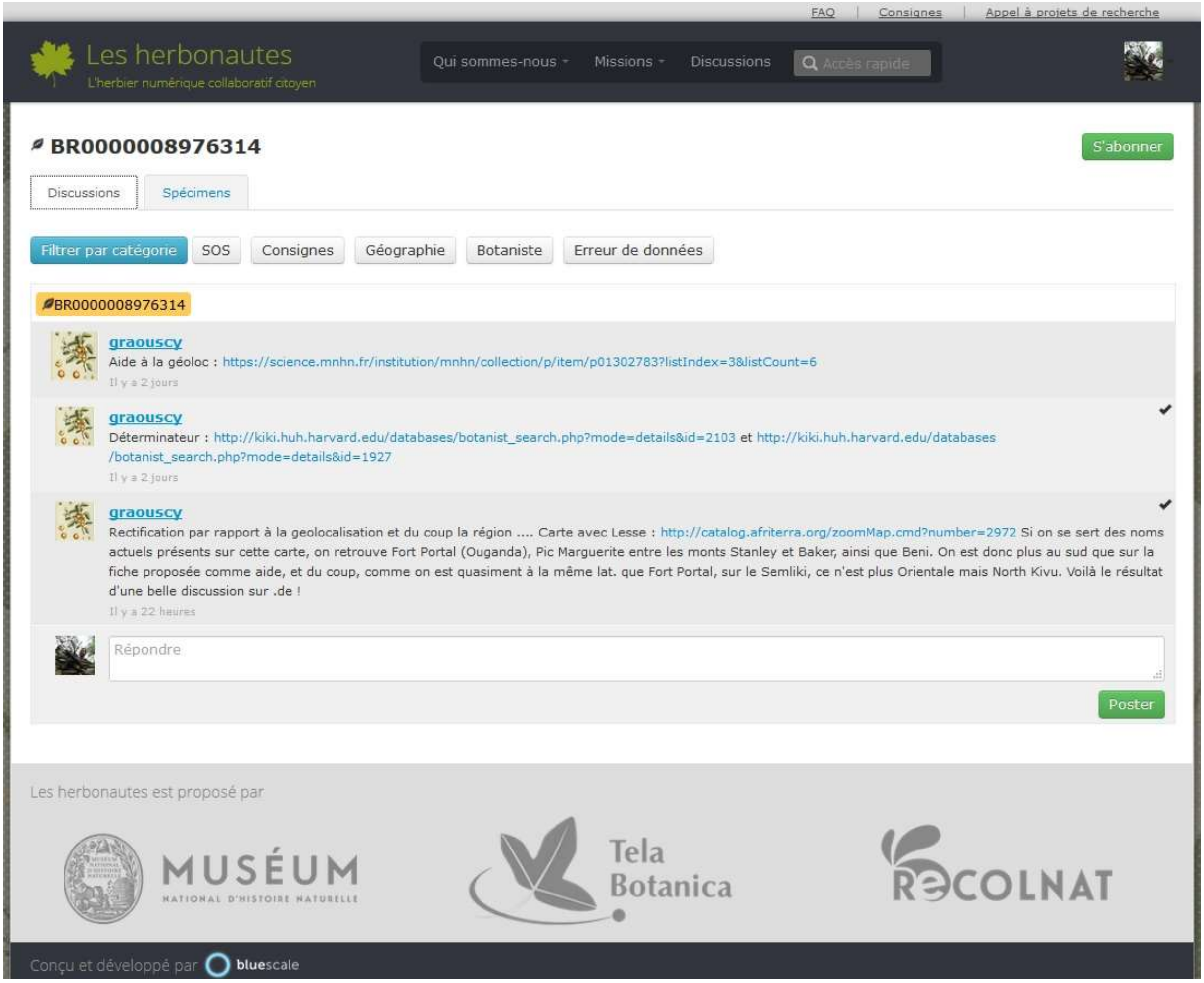
Get help in using DoeDat and reporting issues:

DoeDat@plantentuinmeise.be

Botanic Garden Meise

Nieuwelaan 38, 1860 Meise

Figure 6 Forum thread associated to the specimen BR0000008976314 on Les Herbonautes



BR0000008976314

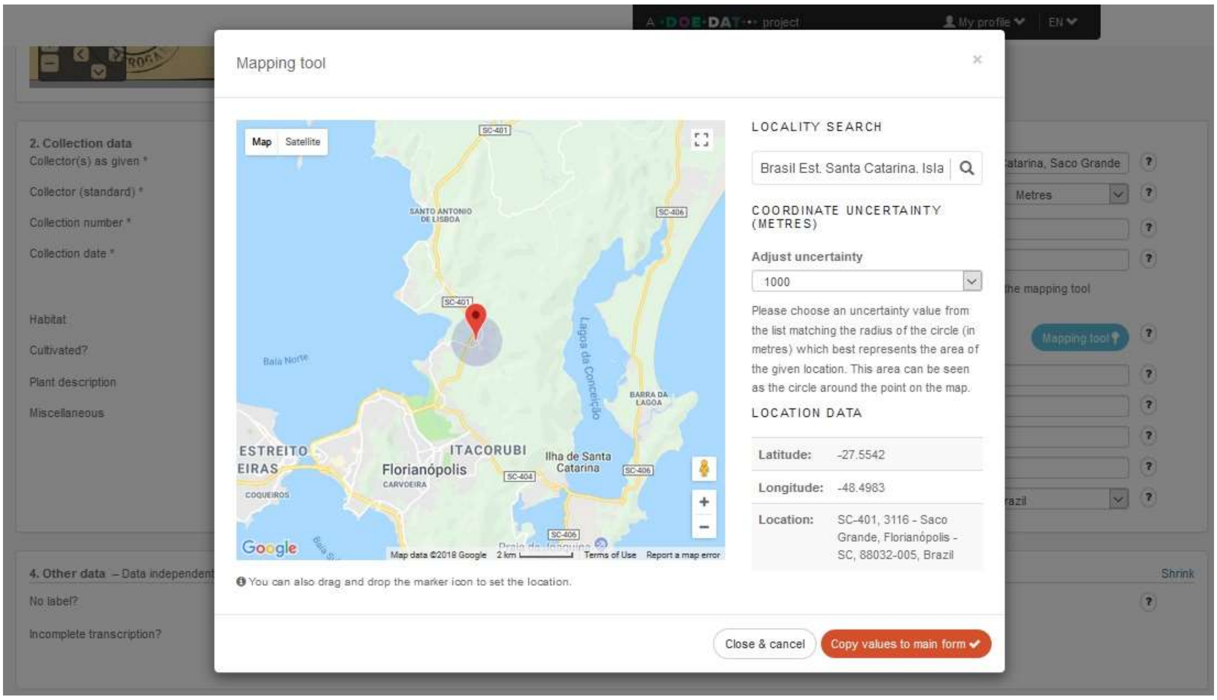
BR0000008976314

graouscy
Aide à la géoloc : https://science.mnhn.fr/institution/mnhn/collection/p/item/p01302783?listIndex=3&listCount=6
Il y a 2 jours

graouscy ✓
Déterminateur : http://kiki.huh.harvard.edu/databases/botanist_search.php?mode=details&id=2103 et http://kiki.huh.harvard.edu/databases/botanist_search.php?mode=details&id=1927
Il y a 2 jours

graouscy ✓
Rectification par rapport à la geolocalisation et du coup la région .... Carte avec Lesse : http://catalog.afriterra.org/zoomMap.cmd?number=2972 Si on se sert des noms actuels présents sur cette carte, on retrouve Fort Portal (Ouganda), Pic Marguerite entre les monts Stanley et Baker, ainsi que Beni. On est donc plus au sud que sur la fiche proposée comme aide, et du coup, comme on est quasiment à la même lat. que Fort Portal, sur le Semliki, ce n'est plus Orientale mais North Kivu. Voilà le résultat d'une belle discussion sur .de !
Il y a 22 heures

Figure 7 Mapping tool on DoeDat, displaying a map based on google maps, a locality search bar helping the location and allowing the user to adjust an uncertainty radius to the data



Mapping tool

LOCALITY SEARCH

Brasil Est. Santa Catarina. Isla

COORDINATE UNCERTAINTY (METRES)

Adjust uncertainty: 1000

Please choose an uncertainty value from the list matching the radius of the circle (in metres) which best represents the area of the given location. This area can be seen as the circle around the point on the map.

LOCATION DATA

Latitude:	-27.5542
Longitude:	-48.4983
Location:	SC-401, 3116 - Saco Grande, Florianópolis - SC, 88032-005, Brazil

You can also drag and drop the marker icon to set the location.

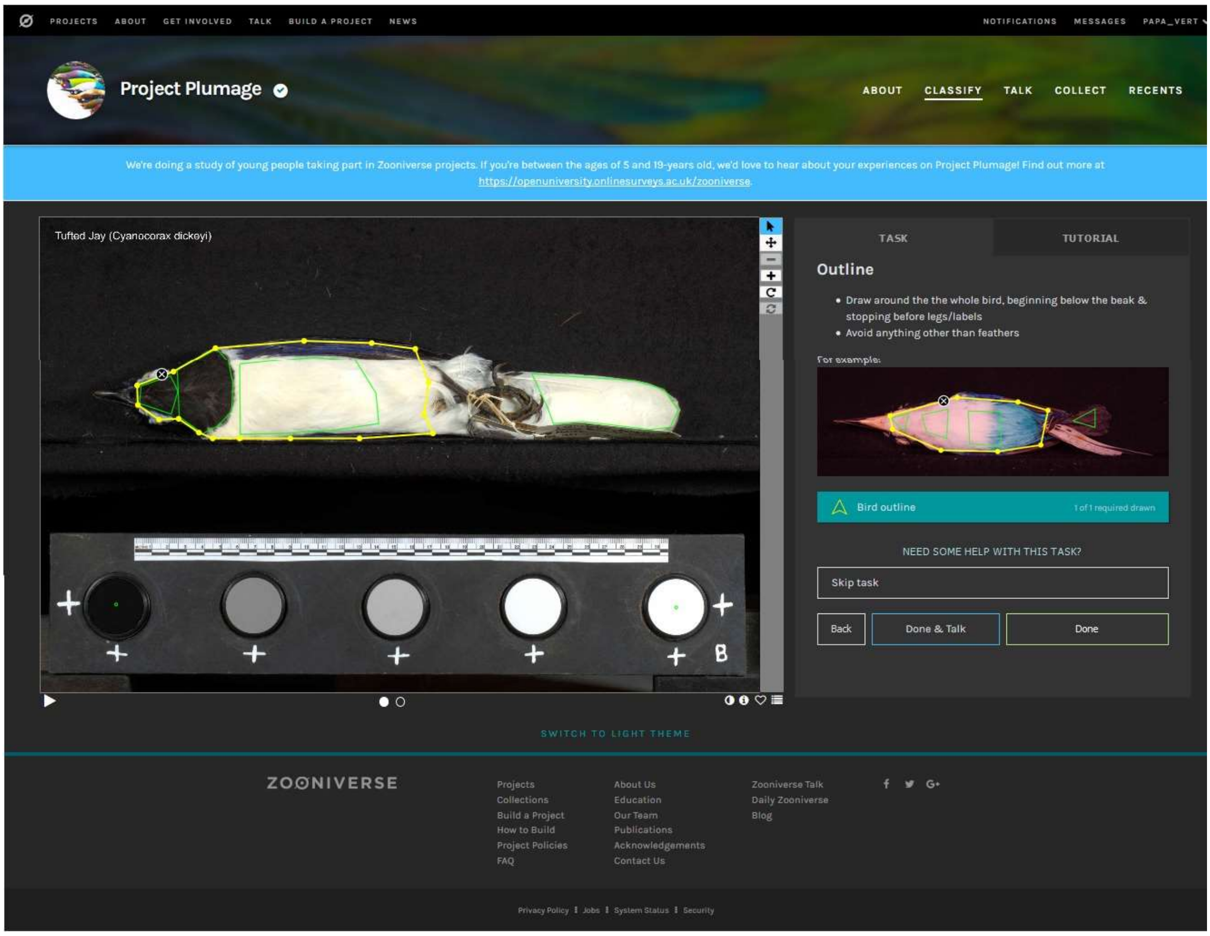
Close & cancel | Copy values to main form ✓

2. Collection data
Collector(s) as given *
Collector (standard) *
Collection number *
Collection date *
Habitat
Cultivated?
Plant description
Miscellaneous

4. Other data – Data independent
No label?
Incomplete transcription?

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

Figure 8 Annotation project Plumage on Zooniverse interface. This project is purely an annotation one. Users are asked to recognise on specimen images area and to design polygon over it for a later analyse by the project scientists.



We're doing a study of young people taking part in Zooniverse projects. If you're between the ages of 5 and 19-years old, we'd love to hear about your experiences on Project Plumage! Find out more at https://openuniversity.onlinesurveys.ac.uk/zooniverse.

Tufted Jay (Cyanocorax dickeyi)

Outline

- Draw around the the whole bird, beginning below the beak & stopping before legs/labels
- Avoid anything other than feathers

For example:

Litterature

Barber A, Lafferty D, Landrum LR (2013) The SALIX Method: A semi-automated workflow for herbarium specimen digitization. Taxon 62: 581–590. doi: 10.12705/623.16

Bonnet P, Goëau H, Hang ST, Lasseck M, Šulc M, Malécot V, Jauzein P, Melet J-C, You C, Joly A (2018) Plant Identification: Experts vs. Machines in the Era of Deep Learning. In: Joly A, Vrochidis S, Karatzas K, Karppinen A, Bonnet P (Eds), Multimedia Tools and Applications for Environmental & Biodiversity Informatics. Springer International Publishing, Cham, 131–149. doi: 10.1007/978-3-319-76445-0_8

Chapman AD (2005) Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data. Copenhagen [Denmark]. Report for the Global Biodiversity Information Facility Available from: http://www.gbif.org/document/80528.

Chapman AD, Grafton O (2008) Guide to Best Practices for Generalising Sensitive Species Occurrence Data. Global Biodiversity Information Facility, Copenhagen [Denmark], 27 pp. Available from: https://www.gbif.org/document/80512.

Chapman AD, Wieczorek J, BioGeomancer Consortium (2006) Guide to best practices for georeferencing. Global Biodiversity Information Facility, Copenhagen [Denmark.

Chupin L (2017) Enjeux communicationnels de la conception de dispositifs de médiation documentaire augmentée pour les herbiers numérisés. École doctorale Abbé Grégoire Available from: https://xupi.eu/these_lisa_chupin/these_chupin.pdf (June 4, 2018).

Dunn S, Hedges M (2013) Crowd-sourcing as a Component of Humanities Research Infrastructures. International Journal of Humanities and Arts Computing 7: 147–169. doi: 10.3366/ijhac.2013.0086

Ellwood ER, Dunckel BA, Flemons P, Guralnick R, Nelson G, Newman G, Newman S, Paul D, Riccardi G, Rios N, Seltmann KC, Mast AR (2015) Accelerating the Digitization of Biodiversity Research Specimens through Online Public Participation. BioScience 65: 383–396. doi: 10.1093/biosci/biv005

Ellwood ER, Kimberly P, Guralnick R, Flemons P, Love K, Ellis S, Allen JM, Best JH, Carter R, Chagnoux S, Costello R, Denslow MW, Dunckel BA, Ferriter MM, Gilbert EE, Goforth C, Groom Q, Krimmel ER, LaFrance R, Martinec JL, Miller AN, Minnaert-Grote J, Nash T, Oboyski P, Paul DL, Pearson KD, Pentcheff ND, Roberts MA, Seltzer CE, Soltis PS, Stephens R, Sweeney PW, von Konrat M, Wall A, Wetzer R, Zimmerman C, Mast AR (2018) Worldwide Engagement for Digitizing Biocollections (WeDigBio): The Biocollections Community's Citizen-Science Space on the Calendar. BioScience 68: 112–124. doi: 10.1093/biosci/bix143

Eveleigh A, Jennett C, Lynn S, Cox AL (2013) "I want to be a captain! I want to be a captain!": gamification in the Old Weather citizen science project. In: ACM Press, 79–82. doi: 10.1145/2583008.2583019

Francoy TM, Wittmann D, Drauschke M, Müller S, Steinhage V, Bezerra-Laure MAF, De Jong D, Gonçalves LS (2008) Identification of Africanized honey bees through wing morphometrics: two fast and efficient procedures. Apidologie 39: 488–494. doi: 10.1051/apido:2008028

Geoghegan H, Dyke A, Pateman R, West S, Everett G (2016) Understanding motivations for citizen science. Final report on behalf of UKEOF. University of Reading, Stockholm Environment Institute (University of York) and University of the West of England, 120pp.

Goëau H, Joly A, Bonnet P, Lasseck M, Šulc M, Hang ST (2018) Deep learning for plant identification: how the web can compete with human experts. Biodiversity Information Science and Standards 2: e25637. doi: 10.3897/biss.2.25637

Greenhill A, Holmes K, Lintott C, Simmons B, Masters K, Cox J, Graham G (2014) Playing with Science: Gamised Aspects of Gamification Found on the Online Citizen Science Project – Zooniverse. In: GAME-ON 2014 15th International Conference on Intelligent Games and Simulation. Dickinson, Patrick, University of Lincoln, UK, 15–24.

Guralnick RP, Wieczorek J, Beaman R, Hijmans RJ, the BioGeomancer Working Group (2006) BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data. PLoS Biology 4: e381. doi: 10.1371/journal.pbio.0040381

Häuser CL, Steiner A, Holstein J, Scoble MJ eds. (2005) Digital imaging of biological type specimens: a manual of best practice ; results from a study of the European Network for Biodiversity Information. Staatliches Museum für Naturkunde, Stuttgart, 309 pp.

Heerlien M, Van Leusen J, Schnörr S, De Jong-Kole S, Raes N, Van Hulsen K (2015) The Natural History Production Line: An Industrial Approach to the Digitization of Scientific Collections. Journal on Computing and Cultural Heritage 8: 1–11. doi: 10.1145/2644822

International Commission on Zoological Nomenclature (1999) International code of zoological nomenclature. 4th ed. Ride WDL, International Trust for Zoological Nomenclature, Natural History Museum (London, England), International Union of Biological Sciences (Eds). International Trust for Zoological Nomenclature, c/o Natural History Museum, London, 306 pp.

Kho SJ, Manickam S, Malek S, Mosleh M, Dhillon SK (2017) Automated plant identification using artificial neural network and support vector machine. Frontiers in Life Science 10: 98–107. doi: 10.1080/21553769.2017.1412361

Kumar N, Belhumeur PN, Biswas A, Jacobs DW, Kress WJ, Lopez IC, Soares JVB (2012) Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (Eds), Computer Vision – ECCV 2012. Springer Berlin Heidelberg, Berlin, Heidelberg, 502–516. doi: 10.1007/978-3-642-33709-3_36

Le Bras G, Geoffroy J-J, Albenga L, Mauriès J-P (2015) The Myriapoda and Onychophora collection (MY) of the Muséum national d'Histoire naturelle (MNHN, Paris). ZooKeys 518: 139–153. doi: 10.3897/zookeys.518.10223

Le Bras G, Pignal M, Jeanson ML, Muller S, Aupic C, Carré B, Flament G, Gaudeul M, Gonçalves C, Invernón VR, Jabbour F, Lerat E, Lowry PP, Offroy B, Pimparé EP, Poncy O, Rouhan G, Haevermans T (2017) The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. Scientific Data 4: 170016. doi: 10.1038/sdata.2017.16

Lee TK, Crowston K, Østerlund C, Miller G (2017) Recruiting Messages Matter: Message Strategies to Attract Citizen Scientists. In: ACM Press, 227–230. doi: 10.1145/3022198.3026335

Leonardo MM, Avila S, Zucchi RA, Faria FA (2017) Mid-level Image Representation for Fruit Fly Identification (Diptera: Tephritidae). In: IEEE, 202–209. doi: 10.1109/eScience.2017.33

Livermore L, Tweddle J, French L, Phillips S, Robinson L, Smith VS (2015) Making molehills out of mountains: crowdsourcing digital access to natural history collections. Synthesys Available from: http://www.synthesys.info/wp-content/uploads/2014/01/NA3-Del.-3.4-Crowdsourcing-report-Phase-2.pdf.

Monteiro M, Figueira R, Melo M, Mills MSL, Beja P, Bastos-Silveira C, Ramos M, Rodrigues D, Queirós Neves I, Consciência S, Reino L (2017) The collection of birds from Mozambique at the Instituto de Investigação Científica Tropical of the University of Lisbon (Portugal). ZooKeys 708: 139–152. doi: 10.3897/zookeys.708.13351

Nelson G, Paul D, Riccardi G, Mast A (2012) Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys 209: 19–45. doi: 10.3897/zookeys.209.3135

Nualart N, Ibáñez N, Luque P, Pedrol J, Vilar L, Guàrdia R (2017) Dataset of herbarium specimens of threatened vascular plants in Catalonia. PhytoKeys 77: 41–62. doi: 10.3897/phytokeys.77.11542

Papastefanou G, Legakis A, Shogolev I (2016) The Avian Collection of the Zoological Museum of the University of Athens (ZMUA). Biodiversity Data Journal 4: e10598. doi: 10.3897/BDJ.4.e10598

Pignal M, Romaniuc-Neto S, Souza SD, Chagnoux S, Canhos DAL (2013) Saint-Hilaire virtual herbarium, a new upgradeable tool to study Brazilian botany. Adansonia 35: 7–18. doi: 10.5252/a2013n1a1

Raddick MJ, Bracey G, Gay PL, Lintott CJ, Murray P, Schawinski K, Szalay AS, Vandenberg J (2010) Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. Astronomy Education Review 9. doi: 10.3847/AER2009036

Rotman D, Hammock J, Preece J, Hansen D, Boston C, Bowser A, He Y (2014) Motivations Affecting Initial and Long-Term Participation in Citizen Science Projects in Three Countries. In: iSchools. doi: 10.9776/14054

Rzanny M, Seeland M, Wäldchen J, Mäder P (2017) Acquiring and preprocessing leaf images for automated plant identification: understanding the tradeoff between effort and information gain. Plant Methods 13. doi: 10.1186/s13007-017-0245-8

Silva AS, Pitta Groz M, Leandro P, Assis CA, Figueira R (2018) Ichthyological collection of the Museu Oceanográfico D. Carlos I. ZooKeys 752: 137–148. doi: 10.3897/zookeys.752.20086

Turland N, Wiersema J, Barrie F, Greuter W, Hawksworth D, Herendeen P, Knapp S, Kusber W-H, Li D-Z, Marhold K, May T, McNeill J, Monro A, Prado J, Price M, Smith G eds. (2018) 159 International Code of Nomenclature for algae, fungi, and plants. Koeltz Botanical Books. doi: 10.12705/Code.2018

Tweddle J, Robinson L, Roy HE, Pocock M, UK Environmental Observation Framework, Natural History Museum (London E, Angela Marmont Centre for UK Biodiversity, Biological Records Centre (Centre for Ecology and Hydrology) (2012) Guide to citizen science: developing, implementing and evaluating citizen science to study biodiversity and the environment in the UK.

Watson AT, O’Neill MA, Kitching IJ (2004) Automated identification of live moths (Macrolepidoptera) using digital automated identification System (DAISY). Systematics and Biodiversity 1: 287–300. doi: 10.1017/S1477200003001208

West S, Pateman R, Dyke A (2016) Data Submission in Citizen Science Projects. Report for Defra (Project number PH0475). Stockholm Environment Institute, University of York

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard Sarkar IN (Ed). PLoS ONE 7: e29715. doi: 10.1371/journal.pone.0029715

Wieczorek J, Guo Q, Hijmans R (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18: 745–767. doi: 10.1080/13658810412331280211

Yang H-P, Ma C-S, Wen H, Zhan Q-B, Wang X-L (2015) A tool for developing an automatic insect identification system based on wing outlines. Scientific Reports 5. doi: 10.1038/srep12786

Zacklad M, Chupin L (2015) Le crowdsourcing scientifique et patrimonial à la croisée de modèles de coordination et de coopération hétérogènes : le cas des herbiers numérisés. Canadian Review of Information Science 39: 308–328.

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

REFERENCES:
MILESTONE MS28

Work package **WP 5** / Task **5.2** Lead: MNHN
Deadline M14

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

SPECIFICATION OF DATA EXCHANGE FORMAT FOR TRANSCRIPTION PLATFORMS

MILESTONE MS28

Authors: Gwenaël Le Bras[1], Simon Chagnoux[1] and Mathias Dillen[2]

1- Muséum National d'Histoire Naturelle (Paris - France)
2- Agentschap Plantentuin Meise (Meise - Belgium)

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

ICEDIG.EU

Content

Introduction

Cataloguing specimens has been one of the core activities in Natural History collections for centuries. In modern times, databases have progressively replaced paper books. A major change happened during the last decade, during which some institutions have completed massive industrial digitization projects, producing millions of images. The pace of this process is still increasing and the planned DiSSCo infrastructure in Europe will provide the scientific community with access to a large number of specimen images and data.

These imaging techniques are not only producing more data, they also drastically changed the workflows, allowing the general public to be involved in the documentation of natural specimens. Over half a million of labels have already been transcribed by volunteers on a dozen of platforms. The ICEDIG project reviewed those platforms (MS26: *Evaluation of existing volunteer transcription systems* available online (Le Bras and Chagnoux 2018)) and concluded that there is no “best transcription platform”: Dynamism of communities, although difficult to quantify, is more important than any specific feature of each website. So DiSSCo will not embed one single platform but will have to interoperate with a growing ecosystem of platforms with different publics, practices and languages.

However, the transcription requirements for specimens are similar enough amongst institutions to agree on a common protocol to exchange data between their databases and transcription platforms. The present document is proposing such a protocol.

We aimed at writing this report in such a way that it can be useful today for both platform administrators and less-technical collection managers, but still generic enough to build DiSSCo services upon it. We have tried not to invent anything new and stick as much as we could to Biodiversity Information Standards (https://www.tdwg.org/). We hope that within the next few years most platforms will implement this specification and become “DiSSCo compatible”

The protocol covers two data flows (Figure 1):

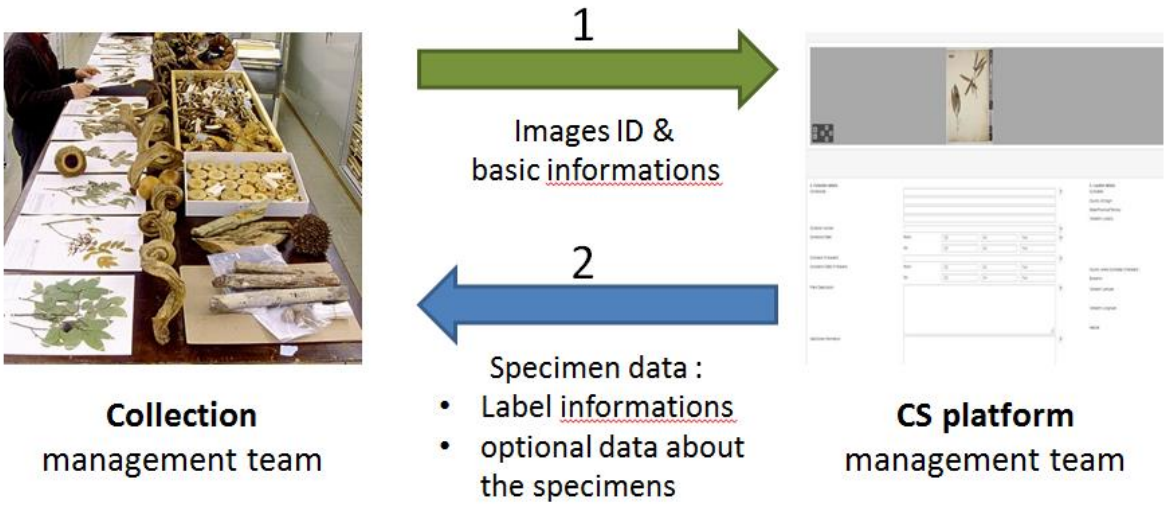


Figure 1: Basic mapping of the Data flow over a citizen science project

The first one described in the chapter “Preparing data for citizen science” offers a simple way to send a set of images and basic information on them from a collection database to a transcription platform.

The second flow delivers transcribed data back to the collections. Detailed in chapter “Structuring citizen science outputs”, it addresses issues shared with ICEDIG Task 4.3 (Interoperability with Collection Management Systems (CMS)), which we have been closely working with despite our more prescriptive approach.

Preparing data for citizen science

Requirements

Planning a citizen science (CS) project requires several steps that will condition the data sent:

- **Have a subject for the project.** One has to be able to explain clearly to the public why this project is being held. This subject can be either a scientific purpose (i.e. the study of a particular Genus, or the flora from a particular area for instance), or a mission about beautiful specimens for the pleasure of the users. It needs in any case to be explainable clearly. A transcription pilot to test data quality held on several different platforms, made us conclude that true motivations of a mission can't be hidden from the users (otherwise they start to imagine what these motivations are). Further results from this transcription pilot (pilot 2) can be found in the report online (Phillips et al. 2019)
- **Select a platform.** Each platform has their own requirements (i.e. number of specimens per project, project description files and images, image format). This information has to be decided before starting to design the project. To help select the most appropriate CS platform to collection holding institutions needs, an *evaluation of existing volunteer transcription systems* was realized in the frame of ICEDIG work package 5.2 on citizen science transcription platform. It can be found online (Le Bras and Chagnoux 2018).
- **A unique standardized way to identify the specimens (catalog number) is used for the collection, and the collection itself is identified.**
- **Every specimen to be part of the project has been imaged.** One to several images can be done for each specimen, depending on particular needs.
- **Labels are clearly readable at least on one image per specimen.** As the information to be transcribed is the target of the project, the labels are a basic requirement. Unless the specimen is of very little interest to the general audience (i.e. dry fungi), the specimen needs to be clearly visible as well on at least one image per specimen in order to keep the user's interest (Le Bras and Chagnoux 2018).
- **Images of the specimens to be included in the projects are available online at a specific Uniform Resource Identifier (URI)**. In order to alleviate the data exchange, the images will not be transferred with the archive. The receiving system will retrieve them from the network. To do so, the images have to be available through a dedicated service, and available at a distinct URI. Images should be available online at least for the duration of the project, and preferentially permanently. For the institutions not able to maintain such a permanent service, ICEDIG will provide an evaluation of Zenodo infrastructure (deliverable D6.3 of the subtask 6.3.3 dedicated to the Zenodo infrastructure and due for end of july 2019). This long-term repository offers the possibility to host online in case no institutional server is available.
- **Licencing for the images should be established** prior to the publishing of images on a third-party CS platform. Any published licence can be referenced here. We suggest to follow, as much as possible the GBIF data licencing terms (https://www.gbif.org/terms). The choice is given to the institutions to choose between the following licences of creative commons (https://creativecommons.org/):
  - CC0: https://creativecommons.org/publicdomain/zero/1.0/
  - CC BY: https://creativecommons.org/licenses/by/4.0/

  - CC BY-NC: https://creativecommons.org/licenses/by-nc/4.0/

  The citation of the licencing on the extract has to be done using the URL redirecting to the full description of the licence terms on the creative commons' website. The version we cite here is the latest to date, yet we invite each institute to check for newer versions while deciding their licencing policy.

- **A taxon or scientific name should correspond to each specimen to be part of the project.** This name can be the name the specimen is filed under. In case the specimen is not determined to species level, it is possible to indicate the lowest taxon rank it was determined to.

Within the frame of its WP6, ICEDIG is categorizing levels of minimum information standards for digital specimens (MIDS). These standards should be published soon. The requirements of our digital specimen to be able to go through a citizen science project is a MIDS level 1 with options.

Once these requirements are met, the main activity of preparing the project will be to select the images to transcribe, and to write a description of the project.
The images and data should then be structured prior to be sent. The next chapter describes that structure.

The procedure hereunder described is as simple as possible, so that any collection manager will be able to follow it. Consequently, no particular IT knowledge is required, other than being able to realise an extract from the local collection management system, and to process it with text editing software and/or spreadsheet tools.

Packing data in a Darwin Core Archive

To exchange data, we will pack them into an archive based on Darwin Core Archive.

What is a Darwin Core Archive?

A Darwin Core Archive is a biodiversity dataset using a list of standardized terms named Darwin Core (DwC). It is widely used to exchange data about species occurrence, taxon checklists, sampling events or collection specimen data. Created between 1998 and 2009 by the Taxonomic Databases Working Group (https://www.tdwg.org/), it has become a major data standard used for many of the main biodiversity science projects such as the Global Biodiversity Information Facility (https://www.gbif.org/) or the Encyclopedia of Life (https://www.eol.org/). Every institution committing data to these projects are regularly producing DwC archives.

A DwC Archive is a simple dataset easy to read on every computer. It is often a simplification of an existing complex database, and is made to share data between different databases. Created to ease biodiversity data exchange, it has become over its years of use a stable and strong data standard.

DwC Archive is the most appropriate way to exchange data in our situation.

Figure 2 depicts a common DwC archive. It is based on a collection data sharing standard, used to share data about collection specimens. More precisely, it is the scheme of DwC archive-building used by the MNHN to share data about the collection it holds to the GBIF.

A DwC archive constitutes a .zip folder containing two types of files:

- .xml files. These are the descriptor files. Descriptors show the metadata of the dataset. There are two of them:
  - meta.xml describes the structure of the dataset itself for a proper reading of the data by a computer. It mentions the encoding, the delimitation character of the values in the data files, the field order, gives links to definitions of the DwC terms used, etc.
  - eml.xml. This file gives a description for human reading of the content of the dataset, in order for users to be able to make good use of the data. It will give information about the persons in charge of managing the data, a text describing the dataset, date of constitution of the dataset. Also, in the case of a specimen collection, the area the specimen originates from, a date (range)when the specimens were collected, etc.
- text files (.txt). These are the data files. They gather the information of the dataset in a separated values file. Usually, this is a tab separated value file, but it can also be comma separated or semicolon separated.
  - A DwC archive contain at least one data file: the core. On the Figure 2 example, the core datafile is “occurrence.txt”, and it contains the basic information about the specimen itself and its collection event. Each row within this file has a unique identifier, which works as its primary key.
  - Usually, the core datafile is accompanied by one or several additional data files. On Figure 2,there are 4 of them: Reference, Identification, Identifier and Multimedia. Within these additional data files, each line refers to one specific line in the core file through this last identifier (secondary key).

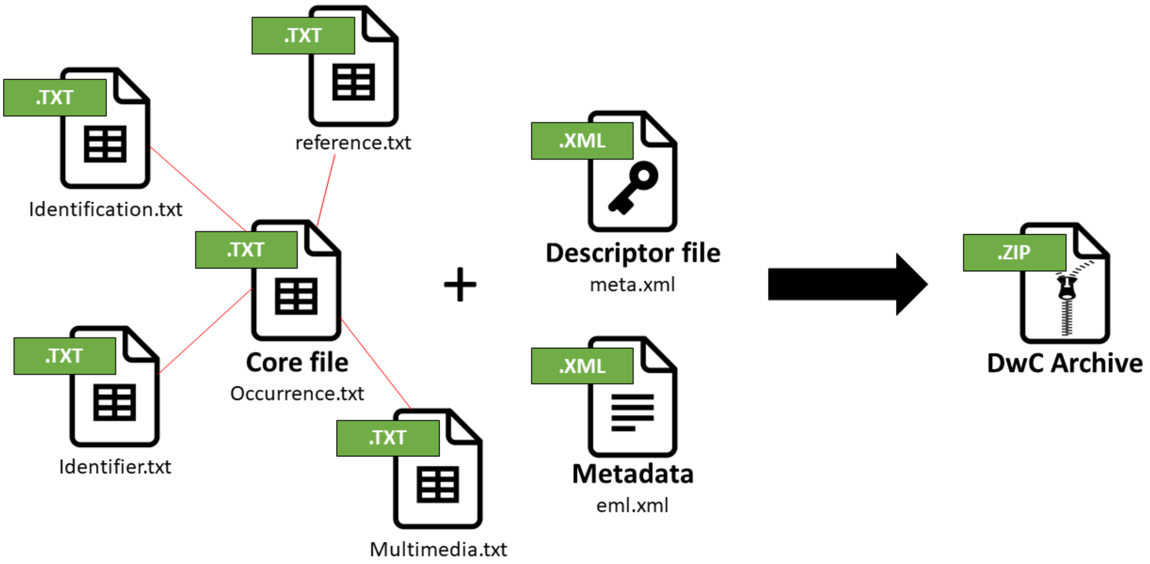


Figure 2: Common Darwin Core Archive composition

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

A simple DwC exchange archive

Data transferred to a CS platform have to be altogether easy to prepare by the collection holder, easy to process by the platform managing team and contain all the required information to set up a CS transcription project. In order to ease the creation of the mission, the data included have to be as complete as possible. We will use here a simplified DwC archive package with only two files:

- **a core datafile** based on the images information
- **a descriptor xml**

No need here to set an eml file, as it will mostly contain mission characteristics that could be sent in an easier way to the platform by mail or informal discussion. As each platform has its own sets of attributes for a transcription project, fixing arbitrary fields for a single record is an unnecessary burden.

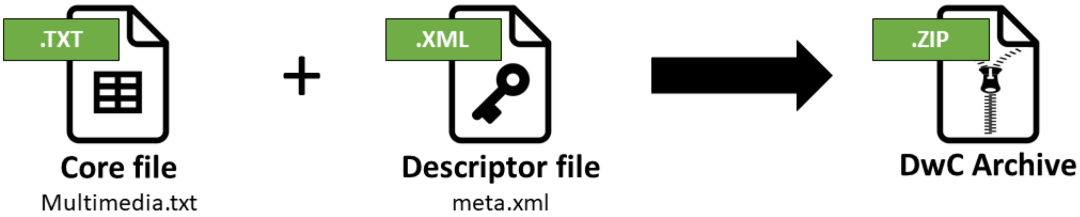


Figure 3: Simplified DwC based Archive for data input on a CS platform from a CMS

The information to include in the core file will be:

- **The unique identifier of the image/media** (*typically URI* - DwC: associatedMedia**)**. This is the URI from where the platform management can get the image from. It is preferably a permalink, but in case no permalinks are available, it should be valid for at least the duration of the CS project.
- **The unique identifier of the specimen** (DwC: occurrenceID). This should ideally be a permalink. These identifiers are to be used to link the newly produced data to the correct specimens. In the case of several images per specimen, this information will allow the CS management team to link all the images of the specimen to the correct entry. The Consortium of European Taxonomic Facilities (CETaF) has worked on unique identifiers in order to help institutions to set their own. More information about this is available on the CETAF website through a poster (Güntsch et al. n.d.) or on the dedicated wiki (https://cetafidentifiers.biowikifarm.net/wiki/Main_Page).
- **The media type** (*usually stillImage* - DwC: type). This is to be useful for the CS platform, as the media has to be handled differently depending on this type.

- **The media format** (DwC: format). CS platforms usually need the input images to be in a certain format. Describing here the format of the image stored and available on the URI will allow them to convert the images if necessary.
- **The licence under which the media should be used** (DwC: licence). This is the licencing chosen as described above.
- **The code of the institution** holding the specimen (DwC: institutionCode).
- **The collection code** the specimen is part of (DwC: collectionCode)
- **The catalog number of the specimen** within the collection the specimen is part of (DwC: catalogNumber).
- **A scientific name** corresponding to the specimen, e.g.: the one the specimen is stored under (DwC: scientificName). In case no taxon name can be linked to the specimen (i.e. a non-determined specimen), a value should be included to state clearly that the absence of data is not due to an informatic technical issue, but rather a determination issue. For instance, the MNHN chose "Insertae sedis" in our data example. The meaning of this value can be easily found by the CS user on the web.

Archive structure in details

The archive to send data to a CS platform is constituted of two files, meta.xml and multimedia.txt, as depicted above (Figure 3). To help picture this archive structure, we built up an illustrative archive (Le Bras 2019a) displaying the specimens used for the trans-institutional and trans-platform pilot project held in the frame of ICEDIG (pilot 2 on the relevant online report (Phillips et al. 2019)). It is possible to re-use the descriptor file (meta.xml) from the example by respecting the following rules:

- Encoding of the files has to be UTF8. This encoding format allows for correct coding of most world language characters, and is supported on most machines. Consequently, it is the most appropriate encoding system for our purpose.
- Files have to be delimited by tabs (\t)
- Lines have to be delimited by the line feed character (\n)
- No field enclosure characters will be used
- Headers have to be one single row
- If needed, new columns should be added at the end

For more details about the descriptor file structure, a description is provided in Appendix 1: The descriptor file in details.

multimedia.txt: This file contains the actual data from our archive. In our example, this file is a tab separated value file, created by doing a copy/paste from spreadsheet software into a text editor software. As mentioned above, the first column (id) is a copy of the second one (associatedMedia). This first column is the primary key of our data (it is then no actual data). The remaining 9 columns are containing relevant data described above. Each column corresponds to a <field/> entry in the data.xml file:

- associatedMedia
- occurrenceID

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

- type
- format
- licence
- institutionCode
- collectionCode
- catalogNumber
- scientificName

Example 1: Data sent from the CMS

Case for the simple specimen P03558024

For readability reasons, we here framed our fields with quotation marks (“). A tabular version of this example is available online (https://doi.org/10.5281/zenodo.2579686).

ID = "http://mediaphoto.mnhn.fr/media/1441365719281fbgNH3QOftOJIz09"

associatedMedia = "http://mediaphoto.mnhn.fr/media/1441365719281fbgNH3QOftOJIz09"

occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"

type = "StillImage"

format = "image/jpeg"

licence = "http://creativecommons.org/licenses/by/4.0/legalcode"

institutionCode = "MNHN"

collectionCode = "P"

catalogNumber = "P03558024"

scientificName = "Castanopsis acuminatissima (Blume) A.DC."

Example 2: Data sent from the CMS

Case for the multi-imaged vertebrate specimen MNHN-ZO-2013-152

For readability reasons, we here framed our fields with quotation marks (“). A tabular version of this example is available online (https://doi.org/10.5281/zenodo.2579738).

This specimen of razorbill has 5 images recorded in GBIF. If it was to be sent to a CS transcription platform, there would be 5 lines describing it in the relevant multimedia datafile filled as follows. This is a common case for zoological or paleontological specimens. In our file, a line in the document should correspond to each image sent, and the same information needs to be repeated for occurrenceID, institutionCode, collectionCode, catalogNumber and scientificName. This will allow the CS platform to link several images to the same specimen.

ID = "http://mediaphoto.mnhn.fr/media/1432022935007Ijp7LVEZylb7BUyF"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1432022935007Ijp7LVEZylb7BUyF"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"
scientificName = "Alca torda Linnaeus, 1758"

ID = "http://mediaphoto.mnhn.fr/media/1432022936311Lia8CCKdSuOY52v7"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1432022936311Lia8CCKdSuOY52v7"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"
scentificName = "Alca torda Linnaeus, 1758"

ID = "http://mediaphoto.mnhn.fr/media/1432022937102nn5mviWGcYEw5eln"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1432022937102nn5mviWGcYEw5eln"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"

```
scentificName = "Alca torda Linnaeus, 1758"

ID = "http://mediaphoto.mnhn.fr/media/14320229378493SF5trxl8WGLFJG1"
associatedMedia = "http://mediaphoto.mnhn.fr/media/14320229378493SF5trxl8WGLFJG1"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"
scentificName = "Alca torda Linnaeus, 1758"

ID = "http://mediaphoto.mnhn.fr/media/143202296244433EpMw3CYLIHKp9l"
associatedMedia = "http://mediaphoto.mnhn.fr/media/143202296244433EpMw3CYLIHKp9l"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"
scentificName = "Alca torda Linnaeus, 1758"
```

In case several specimens are present on the same image

This case can occur for paleontology specimens or herbarium sheets with small specimens attached together on the same sheet. In this case, it is important the specimen can be easily identifiable, and that each one can be clearly linked to its catalog number. The image has then to be duplicated for each specimen recognised on it. One image URI can correspond only to one specimen. Consequently if 3 specimens are present on one image, the image has to be duplicated into three images, each getting a distinct URI which will be linked to a distinct specimen.

Although not impossible, this is a tricky situation even for professional digitising teams, so we suggest to avoid as much as possible such complicated cases in CS transcription project.

An alternative to image duplication could have been to handle the id with distinct value from the associatedMedia.

Structuring citizen science outputs

Once the transcription project is completed, the data needs to be sent back to the curating institution facility. To do so, we will use again a DwC archive. This time our archive will be centred on the specimen information (occurrence for DwC) as depicted in Figure 4. As more data are linked to the specimen, this part is a bit more technical, and requires a better understanding of the DwC archive used and biodiversity databases. However, it is important for both sides of the exchange to understand how it is constituted. The descriptions below address both the collection manager and platform operator. For the collection manager, this document will help him understand how it has been built, in order for him to understand how to treat the archive. For the platform operator, this document offers precision on the DwC terms used and the format of the data stored under DwC terms that needs to be followed.

This step is usually done by the platform operator. He should prepare an export that is conform to the present specification. Of course, automatization of that export by a compiling script or even as a feature of the web application will be the ideal target. The later integration of the data into a collection management system usually necessitates formatting the data to the schema of the systems. The format hereunder was developed in order for this part to be as easy as possible, but it still requires databasing skills. The ideal solution will be to have an import feature in most CMSs based on the described protocol. But at an initial stage, semi-automated methods such as SQL scripts or Open refine (http://openrefine.org/) manipulation can be used.

The format we describe below applies for most of the information included in actual transcription projects. We did have to make compromises in some cases to match platform and DwC characteristics, which are detailed in the methodology chapter.

Packing data in a DwC Archive

Prior to packing the data into an archive, the data should be formatted to fit a common standard define below. We categorised the field/terms used for describing a specimen in function of their use. We defined 3 categories as follow (cf. Methodology):

1. **The basic information**: these are the terms giving the most important information about a specimen. They are answers to the questions where/when/what/by who. They constitute the information that will be most commonly used to describe the specimen, cite it in literature and find it through search engines. These information fields are the firsts ones asked for from citizen scientists. Consequently, these values have to be transcribed if available on the labels. A digitisation cannot be considered complete if one of these fields are left blank (cf. below for the cases with no information available).
2. **The common additional data:** These are the information fields precising the previous ones. As such, they will be used as a complement for basic queries on a search engine.
3. **The optional additional data**: These are the information fields used for specific research projects or fields.

We used here the DwC terms to present them. All these terms can be found online at http://rs.tdwg.org/dwc/ .

Basic information

These data are the very basic ones describing a specimen. They are the ones most commonly searched for in a database, and the ones used to describe the specimen in literature. As such these fields are mandatory in a CS project output (note that the collection number is mandatory only for botany).

- **institutionCode:** the code of the institution holding the specimen.
- **collectionCode:** the code of the collection the specimen is part of.
- **catalogNumber:** the catalogue number of the specimen within the collection the specimen is part of.
- **recordedBy:** the name of the individual(s) collecting/capturing the specimen. The name should be in the format "Name, I.". This format is the most commonly used in existing databases (cf. Methodology). In case several people are mentioned as collector/capturer of the specimen, the names should be transcribed in the same order as on the specimen label, and separated by a ";". Example: "Bonpland, A.J.A.; von Humboldt, F.W.H.A.". There is a work in progress to propose standards for assigning unique IDs to people, but more time will be needed before platform implementation and validation by TDWG.
- **fieldNumber:** the field number attribute to the specimen collection event. It is the number given to the collecting event. This field doesn't apply to all collection specimen. It is a particularity to botany. However, in botany, it is a mandatory field (collection number). In this field, do not use space characters.
- **eventDate:** The date format will follow the norm ISO 8601 (https://en.wikipedia.org/wiki/ISO_8601) to format the dates, in order for most systems to

correctly interpret the given information. The date precision will be to the day (YYYY-MM-DD). In case the collection date is a date range, the information will be put in as described by the ISO norm, with a starting date and an ending date. In case the collection/capture date mentioned is not precise to the day (just a month/year, or just a year), it will be coded as a range. Examples:

  - A specimen collected/captured on the 5th of december 2018: eventDate= “2018-12-05”
  - A specimen collected/captured in December 2018 (no more precision): eventDate: “2018-12-01/2018-12-31”
  - A specimen collected/captured during a mission between the 5th of march 1865 and the 23rd march 1865 (no more precision): eventDate= “1865-03-05/1865-03-23”.

- **countryCode**: the code of the country where the collection took place. Use will be made of the norm ISO 3166-1 alpha-2 (https://en.wikipedia.org/wiki/ISO_3166-1), which is broadly used and interpretable by most systems.
- **country:** The name of the country in full, for easy human reading of the dataset. **scientificName**: The taxon name linked to the specimen. On some platforms, other names can be added to the specimen data. It should then fit into this same field in a different identification row (see below).
- **verbatimLocality:** a verbatim text name of the locality.

Common additional data

- **stateProvince:** the first level of administrative layer below the country one. If possible in a formatted way. Use will be made of the norm ISO 3166-2 (https://en.wikipedia.org/wiki/ISO_3166-2), as it is the most complete referential available.
- **county:** the level of administrative layer below the stateProvince. If possible, in a formatted way.
- **decimalLatitude**/**decimalLongitude**: The latitude and longitude coordinate where the collection/capture took place, in decimal format. This field is the one that should be used to store mapping application outcomes.
- **coordinatePrecision**: The coordinate precision when the mapping application allows to produce an incertitude.
- **verbatimCoordinates**: Full text data can be included in this field. This field will contain the information manually transcribed by the CS user (i.e. transcription of coordinates present on a sheet).
- **minimumElevationInMeters**/**maximumElevationInMeters**: The minimum/maximum elevation where the collection/capture took place, in meters (conversion into this unit accepted by the International System of Units has to be made in the case the sheet mentions elevation in feet).
- **verbatimElevation**: For the case no elevation in meters are available, or the relevant elevation fields in the CS database are not formatted to receive information only in meter, it is possible to include this field in order not to lose the information.

- **establishmentMeans**: The process by which the biological individual(s) represented in the occurrence became established at the location. Use needs to be made of controlled vocabulary here (managed, native, invasive, introduced).
- **identifiedBy:** the name of the scientist who named the specimen. This field will be formatted like recordedBy.
- **dateIdentified:** The date the specimen was identified. As for the eventDate, use will be made of the norm ISO 8601 (https://en.wikipedia.org/wiki/ISO_8601) to format the dates. In case the date precision is not to the day, but to the month or to the year, no interval will be used here, but respectively the YYYY-MM or the YYYY format.
- **modified:** This field records when the data was last modified (automatically implemented by the platform application). As for dateIdentified and eventDate, use will be made of the norm ISO 8601 (https://en.wikipedia.org/wiki/ISO_8601).

Optional additional data

- **habitat:** a verbatim rendition of the habitat the specimen was collected/captured from.
- **occurrenceRemarks:** a verbatim field to gather all information relative to the specimen that cannot fit in the previous fields, such as conservation state. This non-format field can contain loads of different information and be difficult to exploit.
- **organismRemarks:** a verbatim field to gather all information relative to the organism that cannot fit in the previous fields, such as a morphologic particularity. This non-format field can contain loads of different information and be difficult to exploit.
- **taxonRemarks:** a verbatim field to gather all information relative to the taxon that cannot fit in the previous fields, such as the use of this taxon for traditional medical purpose, or food in general. This non-format field can contain loads of different information and be difficult to exploit.
- **Multifield scientific name:** For some uses, it is necessary to have the scientific name split into taxonomical ranks. Although the name should be concatenated into scientific name, it is possible to use the following split. Although not recommended here, if agreed upon between the collection management team and the CS platform, it is also possible to include the rank level above the genus one.
  - **genus**
  - **specificEpithet**
  - **taxonRank:** in case of infraspecificEpithet, this field has to be used to precise the taxon rank of the infraspecific names given in infraspecificEpithet (example “subsp.” or “var.”).
  - **infraspecificEpithet:** this field has to be linked to the presence of data in taxonRank.
  - **scientificNameAuthorship:** scientific authority that described the considered taxon, formatted following the relevant code (botanical (Turland et al. 2018) or zoological (International Commission on Zoological Nomenclature 1999)). For botany, this should follow the IPNI abbreviations based on Brummit works (http://www.ipni.org/ipni/authorsearchpage.do).
- **Geological context:** a paleontological specimen CS project would need to include terms about the geological context the specimen was collected from. Depending on the questions asked

from the citizen scientists, a range of specific terms are available on the TDWG wiki (http://rs.tdwg.org/dwc/terms/GeologicalContext).

- **vernacularName:** a vernacular name of the taxon if available.
- **otherCatalogNumbers:** any other catalogue number than the official one, or number associated to a subcollection the specimen is part of, for example.

Specific cases: no information or uncertain information

The absence of information in a field can mean several things:

- nobody transcribed data relevant to this specific field
- no relevant information is available on the specimen itself

These two cases should be different in their resolution. To resolve the first one, the specimen can be included in a new mission with this specific question asked. In the second case however, putting the specimen through another CS project is pointless. Transcription platforms often include a check-box in case no data are available on the labels. When no data are present on the label (that’s to say, when a CS volunteer checked the “no data” box), use of the code *n/a* has to be made in the relevant field of the DwC archive.

Some CS platforms give their users the possibility to quote a checkbox “I’m not sure” in order to express their uncertainty on some transcription made. To record this in the data archive, a question mark in square brackets will be added to the end of the relevant field character chain (" [?]").

Example:

on the platform:

collector = "von Humboldt, F.W.H.A."
not_sure_checkbox = true

in the DwC archive:

recordedBy = "von Humboldt, F.W.H.A. [?]"

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

Archive structure

As the data the output archive contains describes a specimen, and not only the images associated to it, it will be anchored on the occurrence and this time have several associated files:

- **identification.txt**: the CS project may include the transcription of other determinations than the scientific name from the input file. Consequently, the identification data have to be in a separated data file.
- **multimedia.txt:** this data file could be produced even if the initial data were already including them. This to facilitate the work of troubleshooting in case of issues with the data.
- in case of specific projects, such as measurements of a specimen or segmentation of an image together with transcription of the labels on it, it is possible to include tables as *measurements.txt* or *segmentation.txt* to compile relevant data. These cases are not developed further here, as we concentrated on the transcription process. To make use of them, both sides exchanging data will have to agree on terms. Terms for measurements can be found on the relevant page of the TDWG website (https://terms.tdwg.org/wiki/Darwin_Core_Measurement_or_Fact). DwC allows for the creation of new terms if needed, after agreement by both exchanging sides, such as those that one can imagine in *segmentation.txt* to transmit information about an image segmentation project.

The described archive here will be constituted of a core data file: occurrence.txt, two data tables called identification.txt and multimedia.txt, and meta.xml containing the machine-readable metadata.

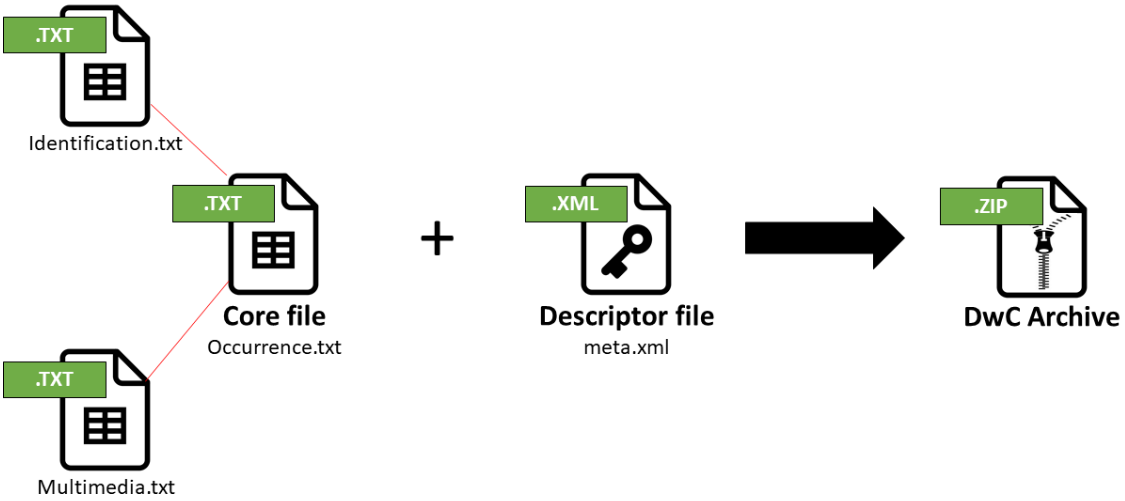


Figure 4: Simple DwC based Archive for data output from a CS platform to a CMS

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

meta.xml: the metadata of the archive. This file's function is to technically describe the archive to allow machines to process it. It is an .xml file. The file uses the following tags:

- <archive> see Appendix 1: The descriptor file in details
- <core> *within the <archive> tag*. see Appendix 1: The descriptor file in details
- <extension> *within the <archive> tag.* This tag functions as the <core> one, but stores secundary data (data linked to the <core> one through secondary keys).
- <file> *within the <core> and <extension> tag.* see Appendix 1: The descriptor file in details
- <location> *within the <file> tag*. see Appendix 1: The descriptor file in details. In this case the core is named "occurrence.txt" and the extensions "identification.txt" and "multimedia.txt". All three are located in the same folder as the meta.xml (the archive folder).
- <id/> *within the <core> tag*. This non-including tag defines the very first column of the core datafile. It contains the primary key of our data
- <coreid/> *within the <extension> tag*. This non-including tag defines the very first column of the core datafile. It contains the secundary key of our data, that refers to the primary keys (id) from the core.
- <field/> *within the <file> tag.* see Appendix 1: The descriptor file in details.

occurrence.txt: This datafile is the core of our document. As such, the id's in the first column are the unique identifier of our data. As there should exist an institutional unique identifier for each natural history collection specimen, the ID of occurrence will be the occurrenceID. The core datafile then contains all the information concerning the specimen, organism and collection/capture event. The file has to contain at least the mandatory ones and the basic informations:

- **id** (technically mandatory): As said above this is a column for machine reading. The choice has to be made of a unique identifier per specimen. We chose occurrenceID in our example.
- **modified** (optional): as the data have been modified during the CS project duration, it could be useful to give back that information to the collection.
- **occurrenceID** (mandatory): This column is for human reading. Even if it bears the exact same information as ID column, it has to be repeated.
- **institutionCode** (mandatory)
- **collectionCode** (mandatory)
- **catalogNumber** (mandatory)
- **recordedBy** (basis of record)
- **recordNumber** (basis of record)
- **eventDate** (basis of record)
- **countryCode** (basis of record)
- **verbatimLocality** (basis of record)
- **stateProvince** (optional common)
- **county** (optional common)
- **decimalLatitude** (optional common)

- **decimalLongitude** (optional common)
- **coordinatePrecision** (optional common)
- **verbatimCoordinates** (optional common)
- **minimumElevationInMeters** (optional common)
- **maximumElevationInMeters** (optional common)
- **verbatimElevation** (optional common)
- **establishmentMeans** (optional common)
- **habitat** (optional)
- **occurrenceRemarks** (optional)
- **organismRemarks** (optional)
- **otherCatalogNumbers** (optional)

identification.txt: This extension data file contains the data pertaining to the identification and the taxon the specimens have been identified as. Each specimen must have at least one identification (= correspond to a line in this datafile). The separate data file for the identification allows us to get several identifications for the same specimen. Consequently, it is possible to get several lines with the same coreID in the identification file, each of these lines corresponding to a single identification of the specimen. However, each identification line should correspond to one and only one line in the occurrence file.

- **coreID** (technically mandatory): This is the secundary key linking the data from this data file to the occurrence data file.
- **modified** (strongly suggested)
- **scientificName** (basis of record)
- **identifiedBy** (optional)
- **dateIdentified** (optional)
- **taxonRemarks** (optional)
- **genus** (optional)
- **specificEpithet** (optional)
- **taxonRank** (optional)
- **infraspecificEpithet** (optional)
- **scientificNameAuthorship** (optional)
- **vernacularName** (optional)

multimedia.txt: this data file contains the data pertaining to the images used for the CS project. Same as for identification.txt, it is possible to have several images for the same specimen (=several lines in the datafile with the same coreID/occurrenceID, but with a different associatedMedia). This data file is presented here to allow troubleshooting in case of mismatch during the production or the data exchange.

- **coreID** (technically mandatory): same as for identification.txt.
- **associatedMedia** (mandatory)
- **type** (mandatory)
- **format** (mandatory)

- **licence** (mandatory)

As for the data sent to the CS platform, we created an illustrative archive (Le Bras 2019b) with CS project data. We used here the data produced on Les Herbonautes for the pilot project held within ICEDIG WP4.2 (pilot project 2 on the report (Phillips et al. 2019)). As for the input archive, it is possible to re-use the meta.xml, however a modification of the terms in the mission result may have to be done. The formating rules to follow to use the archive as working base are the same as above.

Example 3: Data sent from the CS platform

Case for the simple specimen P03558024

A tabular version of this example is available online, in which the two data files are made into separate sheets into the same spreadsheet (https://doi.org/10.5281/zenodo.2579753).

Remarks: In this example, empty fields are specified by null. The field number of this specimen was not transcribed although it is clearly visible on the image (13488). The information reflects it was not transcribed by leaving the field empty. Same applies for dateIdentified, which is as well clearly visible on the label (1965-05-03). On the other hand, no elevation was specified on the label, and the mention n/a indicates someone did notice the lack of information on the label and reported it in the dataset. Same applies for the name of the identifier (in facts, specialists knowing the collection will know M. Debray is here himself the identifier, but that is deduction we can hardly ask from volunteers).

occurrence.txt (1 line)

```
id = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
modified = "2015-09-04"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
institutionCode = "MNHN"
collectionCode = "P"
catalogNumber = "P03558024"
recordedBy = "Debray, M."
fieldNumber = null
eventDate = "1964-07-28"
countryCode = "FR"
verbatimLocality = "Côtes-du-Nord : Perros-Guirec à Ploumanac'h"
stateProvince = "FR-E"
county = "FR-22"
decimalLatitude = 48.83698
decimalLongitude = -3.4831
coordinatePrecision = null
verbatimCoordinates = "48° 50' 13.128'' N ; 3° 28' 59.16'' O"
minimumElevationInMeters = "n/a"
```

```
maximumElevationInMeters = "n/a"
verbatimElevation = "n/a"
```

identification.txt (1 line)

```
coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
scientificName = "Thymus polytrichus A.Kern. ex Borbás"
identifiedBy = "n/a"
dateIdentified = null
```

multimedia.txt (1 line)

```
coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1441365719281fbgNH3QOftOJIz09"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
```

Example 4: Data sent from the CS platform

Case for the multi-determined specimen P01978557

A tabular version of this example is available online, in which the two data files are made into separate sheets into the same spreadsheet (https://doi.org/10.5281/zenodo.2579768).

Remarks: In this example, three lines in the identification documents refer to a specimen with three different identifications (action of identification). The date of collection, as often in old specimens, was not specified to the day, so the information was treated as a range (between the 1st of February 1889 and the 28th of February 1889. On the other hand, Kok made his identification in June 2018, and the month is treated as such following ISO 8601.

occurrence.txt (1 line)

```
id = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
modified = "2018-07-17"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
institutionCode = "MNHN"
collectionCode = "P"
catalogNumber = "P01978557"
recordedBy = "Balansa, B."
fieldNumber = "2414"
eventDate = "1889-02-01/1889-02-28"
countryCode = "VN"
verbatimLocality = "Hanoï, dans les jardins"
```

stateProvince = "VN-HN"
county = null
decimalLatitude = 21.02776
decimalLongitude = 105.83416
coordinatePrecision = null
verbatimCoordinates = null
minimumElevationInMeters = "n/a"
maximumElevationInMeters = "n/a"
verbatimElevation = "n/a"

identification.txt (3 lines)

- coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
  scientificName = "Cinnamomum tonkinense (Lecomte) A.Chev."
  identifiedBy = "Kok"
  dateIdentified = "2018-06"

- coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
  scientificName = "Cinnamomum tonkinense (Lecomte) A.Chev."
  identifiedBy = "Kostermans"
  dateIdentified = "n/a"

- coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
  scientificName = "Cinnamomum albiflorum Nees var. tonkinensis Lecomte"
  identifiedBy = "Kok"
  dateIdentified = "2018-06"

multimedia.txt (1 line)

coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
associatedMedia = "http://mediaphoto.mnhn.fr/media/14413029065481L6TwhTj5Lagqxn4"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"

Giving feedback on transcriber activity

During the transcription process, the present document specifies no data exchanges between transcription platform and institutional CMS or future DiSSCo infrastructure.

A dashboard displaying the digitization progress of European collections is under study. As a first step, during the ICEDIG project, the design of such a DiSSCo Dashboard is focusing on reflecting the state of member's collections and digital catalogues. Citizen science was out of scope. However, having the results of their work displayed outside of the platform is a major incentive for transcriber's communities.

A global dashboard for transcription platforms has been set up temporary for the WeDigBio event. On February 2019 it was still available on https://wedigbio.org/. IDigBio, the United States of America program to facilitate national collection digitization, launched WeDigBio in 2014. This event is annual, lasts 4 days, and aims to highlight and encourage biodiversity collections transcription by citizens worldwide. A dashboard has been developed to display on the event website the worldwide activity linked to the project.

Some interoperability was needed during WeDigBio event. A draft protocol is documented on https://github.com/iDigBio/wedigbio-dashboard. This event-driven dashboard can serve as a starting point for designing a more general protocol.

Implementation of such a protocol in platforms and a DiSSCo Dashboard will allow to measure the role of volunteers in transcription and foster participation.

Methodology

During the process of design of exchange protocols, our major concern was to deliver a simple solution. As simplicity was not always simple when facing the diversity of Natural History collections, practices and transcription platforms we had to make several trade-offs.

This chapter details each step of the design process.

1- Inventory of transcription platforms

Prior to work on this document, we made an inventory of the existing major transcription systems dedicated to natural history collections. This *evaluation of existing volunteer transcription systems* is available online (Le Bras and Chagnoux 2018).

2- Choice of a basic data model

The first step in order to write this specification was to choose the biodiversity standard to start from. We wanted it to fit the following requirements:

- Remain simple
- Being a commonly used solution. This is especially important because:
    - There are more chance the exchanging parties will know about the data model already
    - The platform technical teams might already be able to create automated imports/exports using the standard (so the data model can be applied straight ahead)
    - Documentation about the data model exists
    - People already worked on and thought about the data model, its limitations, and how to improve it
    - The model is then stronger
- Being adapted to specific needs of biodiversity collections databases
- Being open source
- Being costless to use, and having no particular software involved in its use
- Using light files to transfer the information

Darwin Core Archive appeared to us as being the best possible solution, as it met all of our requirements.

It was then proposed during ICEDIG all-hands meeting in Meise on the 5th of december 2018, and our proposition was validated by the 5.2 working group.

3- Adaptation of the DwC model structure to our specific requirements

In order to facilitate the use of DwC we decided to get simplified structures for our archive. This included:

- The abandoning of the EML descriptive file. Indeed, the project description requirements significantly differ from one platform to another. It seemed to us way more complicated to fit these different requirements in a standard. We then left it to the two parties to exchange data on this subject as they already do.
- The data sent from the CMS to the platform was to be limited and focusing on the images.
- The data from the platform to the CMS was to be focused on the occurrence.

These basic schemes were then proposed during ICEDIG all-hands meeting in Meise on the 5th of December 2018, and our propositions were validated by the 5.2 working group.

4- Inventory of terms currently used on the major CS platforms and DwC correspondence

The first step was to inventory the existing fields on the major CS platforms (Herbonautes/Herbonauten, Doedat/Digivol, Zooniverse). As could be expected, for each platform dealing with the same types of data, the vast majorities of the fields were common from one platform to the other, with only slight differences.

A correspondence in DwC terminology was then sought for each CS platform when possible.

The first issues were then identified with particular fields on some platforms, such as the Belgium national geographical grid cells systems (IFBL) codes been asked for on Doedat, for instance. If such concepts cannot fit in existing fields like verbatimCoordinates, the decision was made to leave this issue to be discussed between the CS platform and the collection management team on a case to case basis.

5- Categorization of DwC terms following their use in collection

The list of terms was then categorized into three categories:

- **Mandatory for setting a CS mission**. This list was buildt by listing the minimal information about a specimen that appears on a CS platform prior to its transcription, and also based on the experience of what is needed to build a mission on les Herbonautes.
- **Basic information**. This list was built by considering the information given in a summary of results from search engines used by institutional collection platforms, GBIF, Jstor and others. This list was then compared to information used to cite a specimen in literature.
- **Additional information.** This list contained all the information about the specimen that can be found on a CS platform and that fits our list of terms. As it contains a lot of terms, we later divided it in two subparts, mainly in order to ease readability of the document. This division is suggestive, based on authors' experiences of collection databases, but exists purely for an easier understanding, and it doesn't affect the specifications themselves. These parts are:
  - **Common additional data.** The additional data that can be used for sorting the specimens in a database.

    - **Optional additional data.** The fields in which the data will be usually input in a way to specify the information, but which are difficult to query by a search engine.

The following list was made:

- mandatory for setting a CS mission
    - associatedMedia: the unique identifier of the image/media
    - occurrenceID: the unique identifier of the specimen corresponding to the image
    - type: the type of the media (usually stillImage)
    - format: the format of the media
    - licence: the licence under which the media should be used
    - institutionCode: the code of the institution holding the specimen
    - collectionCode: the collection code the specimen is part of
    - catalogNumber: the catalog number of the specimen
    - scientificName: a scientific name corresponding to the specimen.
- basic information
    - recordedBy
    - recordNumber
    - eventDate
    - countryCode
    - verbatimLocality
- optional fields
    - stateProvince
    - county
    - decimalLatitude
    - decimalLongitude
    - coordinatePrecision
    - verbatimCoordinates
    - verbatimElevation
    - minimumElevationInMeters
    - maximumElevationInMeters
    - establishmentMeans
    - habitat
    - occurrenceRemarks
    - organismRemarks
    - otherCatalogNumbers

This list was later crossed with the minimum information standard for digital specimens working document (version 0.5) set in the frame of work package 6 of ICEDIG.

6- Realisation of an illustrative archive of data import to a CS platform

In order to test the archive realization step by step, we created one based on the pilot conducted by ICEDIG for WP4.2. This mission constituted of specimens from 7 institutions within Europe, with different languages involved and should represent a wider case study. More information on the data used in this pilot can be found in the data paper (Dillen et al. 2019). As is the case for most transcription projects however, this project is constituted only of herbaria specimens. Indeed, due to greater technical difficulty, zoological and even more paleontological collections are imaged to a much lower extent than botanical ones. At the time of writing this document, no CS paleontological project was being run. Consequently, there was no reference we could build on. Issues from different collection types were hypothesized based on authors' experience, and DwC plasticity can be expected to allow relatively easy troubleshooting in case of issues due to zoological or paleontological particularities.

7- Realisation of a notional archive of data exportation from a CS platform

We then build a notional archive of export based on the pilot project data produced on les Herbonautes (mission held from 22/06/2018 to 10/10/2018). This raised several issues:

- How to indicate the absence of data on the specimen (to differentiate it from "no data has been produced"). On les Herbonautes, there is a "no information" checkbox to be checked by transcribers in the case there is no data available. This type of checkbox is as well present on other CS systems. We then decided to use "n/a" to distinguish it from null.
- Although never quoted in the final data export from our mission, the systems give the possibility to the citizen scientist to express their uncertainty by quoting a checkbox for "I'm not sure". This sort of checkbox exists on most of the CS platforms. We decided to add " [?]" to the end of the relevant field content in case the checkbox is quoted.
- The collector/capturer names (RecordedBy) are very often spread between several columns, especially in the case of multiple collectors (first collector in a column, others in a second column). Although it is basic information about the specimen, it is very difficult to have it standardized. Before deciding a format, we compared data from different CS platforms and on GBIF. We then chose the name format which was most common: the family name first immediately followed by a comma, a space, and the initial(s) each followed by a point. (Name, F.). In order to ease the separation of the names and to distinguish between the comma that separates the initials from the surname, a semicolon will be used between the names of different persons.
- Several formats of dates are in use on the eventDate. The ISO 8601 https://en.wikipedia.org/wiki/ISO_8601 norm provides a solution to all the issues met (range, lack of precision). The other date format columns (dateIdentified and Modified) should be formatted in ISO 8601 normally as exposed above.
- Sometimes, no scientific name can be linked to the specimen (i.e. a non-determined specimen). An indication should then be included to state clearly that the absence of data is

not due to an informatic technical issue, but rather a determination issue. We chose here "Insertae sedis", as the meaning of this value can be easily found by the CS user on the web.

Part of these different solutions were discussed on 5th December 2018 at the second ICEDIG All-Hands meeting by the WP5.2 group.

8- Crossing information with data quality working group

The ICEDIG Report on new methods for data quality assurance, verification and enrichment (Phillips et al. 2019) did compare data quality from different crowdsourcing platforms and other resources. It allowed us to confirm the choices made above about standardisation for recordedBy, eventDate, and fieldNumber. It helped us understand the most common issues met within the data in order to confirm the solutions we proposed. Data standard solutions described here are therefore considered to facilitate data interoperability as well as data quality.

9- Redaction of the specification document

We then gathered all the information here and structured the first version of this document. This document was later completed with remarks from the ICEDIG community and some external partners (WeDigBio and BGBM).

Conclusion

The protocol described in this document aims at facilitating data exchanges and, as such, at facilitating natural history collections digitization through citizen science platforms. It was elaborated with the concern of keeping things as simple as possible.

The proposed protocol seems simple enough to be implemented soon by major platforms. We hope that the document is clear enough to allow collection managers to prepare the images for citizen science. We also hope that the document is precise enough for implementation by platform administrators.

Independently of implementation, the transcription by volunteers will continue in the next coming months and years. Having several protocol compatible platforms in the European landscape will allow better interoperability, which will both open citizen science to institutions with no transcription platform, and leverage transcription capacity by taking advantage of the specific strength of each community, the most obvious one being language proficiency.

With these advances in interoperability, the integration of transcription platform into DiSSCo should be relatively straightforward, when the infrastructure will be up and running in a few years time.

Appendix 1: The descriptor file in details

From one archive to another, meta.xml keeps the same structure. Its function is to technically describe the archive to allow machines to read and process it. It is an .xml file. This file uses the following tags:

- <archive> This tag includes an attribute describing the whole archive format:
  - xmlns (XML NameSpace) indicate which "xml language" your archive is written. In our example, it is a DwC archive, basic terms of which are available on "http://rs.tdwg.org/dwc/text/"
- <core> *within the <archive> tag*. this refers to the central file of our archive (here multimedia.txt). As we only have one datafile, it is the one being described here. This tag includes attributes describing the core format:
  - encoding: the character encoding of the file (in our example it is UTF8).
  - fieldsTerminatedBy: which character are used to separate fields within your archive. By default, if saved in common spreadsheet software ";", if copy/paste from a spreadsheet software onto a notepad solution "\t". In our example "\t".
  - linesTerminatedBy: which character are used to go on the next line within the core of your archive. By default, in common spreadsheet software generated files "\n". In our example "\n".
  - fieldsEnclosedBy: which character are used to frame your field content (depending on your procedure on spreadsheet software can be framed by " or not). In our example, there are none.
  - ignoreHeaderLines: This define the size (in row) of the headers in the datafile. In our example, one single line describes the contents of each columns. To note that the headers in the datafile are only for human reading. The machine takes the information about the columns from what will be given in the tags <id/> and <field/> described here under.
  - rowType: where to find the information about the core format. In our example, it is the DwC extension for simple multimedia file (http://rs.gbif.org/terms/1.0/Multimedia)
- <file> *within the <core> tag.* This tag refers to the datafile basic informations
- <location> *within the <file> tag*. this mention the relative location of the core file (relatively to the meta.xml one). In our case the core is named “multimedia.txt” and located in the same folder as the meta.xml.
- <id/> *within the <core> tag*. This non-including tag define the very first column of our datafile. Its attribute is
  - index: locate the content in the row. For <id/> the index is 0 as it is not properly data for the computer: the computer considers it as the primary key of our core data (the unique identifier of each row). In our example, we used the associatedMedia content, as it is a unique identifier for our images.
- <field/> *within the <core> tag.* This describe the proper content of each column of our datafile. Each column containing data in our file should correspond to a <field/> entry. Its attributes are:

- index: defining the position of the column after the ID. The column with index=1 is then the first column after the id (that’s to say, for a human eye, the actual second column, or the one spreadsheet software usually defines as the column “B”). <field/> should be ordered by growing index.
- term: gives the URL of a descriptive of the content.

References

Dillen M, Groom Q, Chagnoux S, Güntsch A, Hardisty A, Haston E, Livermore L, Runnel V, Schulman L, Willemse L, Wu Z, Phillips S (2019) A benchmark dataset of herbarium specimen images with label data. Biodiversity Data Journal 7. doi: 10.3897/BDJ.7.e31817

Güntsch A, Hagedorn G, Hyam R, Röpert D (CETAF stable identifiers for specimens. CETAF-ISTC Available from: http://cetaf.org/sites/default/files/cetaf-istc_stable_identifiers_poster50x70.pdf.

International Commission on Zoological Nomenclature (1999) International code of zoological nomenclature. 4th ed. Ride WDL, International Trust for Zoological Nomenclature, Natural History Museum (London, England), International Union of Biological Sciences (Eds). International Trust for Zoological Nomenclature, c/o Natural History Museum, London, 306 pp.

Le Bras G (2019a) Illustrative Darwin core archive to input data on a citizen science platform from a collection management system. doi: 10.5281/zenodo.2579778

Le Bras G (2019b) Illustrative Darwin core archive to output data from a citizen science platform to a collection management system. doi: 10.5281/zenodo.2579782

Le Bras G, Chagnoux S (2018) Evaluation of Existing Volunteer Transcription Systems. Zenodo doi: 10.5281/zenodo.2578938

Phillips S, Dillen M, Groom Q, Green L, Weech M-H, Wijkamp N (2019) Report on New Methods for Data Quality Assurance, Verification and Enrichment. ICEDIG/DiSSCo. Deliverable 4.2 Available from: https://icedig.eu/sites/default/files/deliverable_d4.2_icedig_data_quality_in_transcription.pdf.

Turland N, Wiersema J, Barrie F, Greuter W, Hawksworth D, Herendeen P, Knapp S, Kusber W-H, Li D-Z, Marhold K, May T, McNeill J, Monro A, Prado J, Price M, Smith G eds. (2018) 159 International Code of Nomenclature for algae, fungi, and plants. Koeltz Botanical Books. doi: 10.12705/Code.2018