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 **WP 5** / Task **5.2** Lead: MNHN
Deadline MS7

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

Evaluation of existing volunteer transcription systems

MILESTONE TITLE

MILESTONE MS26

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

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/) 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/). 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/). 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/). 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 **Erreur ! Source du renvoi introuvable.**.

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 months.

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 XX.

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 code for Zooniverse, DigiVol/DoeDat and for Les Herbonautes is 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 have 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.

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.

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 Meises 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 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 (**Erreur ! Source du renvoi introuvable.**). 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.

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

- **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.

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

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). 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 the 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 users 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*" 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.

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.