Quality issues in georeferencing:

From physical collections to digital data repositories for ecological research

WORKSHOP REPORT

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

Hosted at the Biological And Chemical Research Centre, University Of Warsaw, Warsaw, Poland

10-13 February 2020

Authors: Marcer A., Haston E., Groom Q., Ariño A., Chapman A.D., Bakken T., Braun P., Dillen M., Ernst M., Escobar A., Fichtmüller D., Livermore L., Nicolson N., Paragamian K., Paul D., Petterson L. B., Phillips S., Plummer J., Rainer H., Rey I., Robertson T., Röpert D., Santos J., Uribe F., Waller J. and Wieczorek J.R.

Corresponding author

Arnald Marcer (arnald.marcer@uab.cat), CREAF, 081

https://doi.org/10.5281/zenodo.3734848

This report is based upon work from COST Action CA 17106 MOBILISE, supported by COST (European Cooperation in Science and Technology).

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

www.cost.eu

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

Author affiliations

Arnald Marcer, CREAF, 08193 Bellaterra (Cerdanyola del Vallès), Catalonia, Spain, ORCID 0000-0002-6532-7712
Elspeth Haston, Royal Botanic Garden Edinburgh, UK, ORCID 0000-0001-9144-2848
Quentin Groom, Botanic Garden Meise, Belgium, ORCID 0000-0002-0596-5376
Arturo Ariño, Universidad de Navarra, Spain, ORCID 0000-0003-4620-6445
Arthur D. Chapman, Australian Biodiversity Information Services, Australia, ORCID 0000-0003-1700-6962
Torkild Bakken, NTNU University Museum, Trondheim, Norway, ORCID 0000-0002-5188-7305
Paul Braun, Musée National d'Histoire Naturelle, Luxembourg, ORCID 0000-0002-3620-6188
Mathias Dillen, Botanic Garden Meise, Belgium, ORCID 0000-0002-3973-1252
Marcus Ernst, Botanischer Garten und Botanisches Museum Berlin, Germany
Agustí Escobar, CREAF, 08193 Bellaterra (Cerdanyola del Vallès), Catalonia, Spain, ORCID 0000-0002-6856-0480
David Fichtmüller, Botanischer Garten und Botanisches Museum Berlin, Germany, ORCID 0000-0002-0829-5849
Laurence Livermore, Natural History Museum London, UK, ORCID 0000-0002-7341-1842
Nicky Nicolson, Royal Botanic Gardens, Kew UK, ORCID 0000-0003-3700-4884
Kaloust Paragamian, Hellenic Institute of Speleological Research, Crete, Greece, ORCID 0000-0001-7372-733X
Deborah Paul, Florida State University, USA, ORCID 0000-0003-2639-7520
Lars B. Petterson, Biodiversity Unit, Department Biology, Lund University, Sweden, ORCID 0000-0001-5745-508X
Sarah Phillips,Royal Botanic Gardens, Kew, UK, ORCID 0000-0002-9155-8573
Jack Plummer, Royal Botanic Gardens, Kew, UK, ORCID 0000-0002-1575-5241
Heimo Rainer, Naturhistorisches Museum Wien, Austria, ORCID 0000-0002-5963-349X
Isabel Rey, Museu Nacional de Ciencias Naturales (CSIC), Madrid, ORCID 0000-0002-2122-5124
Tim Robertson, GBIF, Copenhaghen, Denmark, ORCID 0000-0001-6215-3617
Dominik Röpert, Botanischer Garten und Botanisches Museum Berlin, Germany, ORCID 0000-0001-6565-8450
Joaquim Santos, University of Coimbra, Portugal, ORCID 0000-0002-2160-4968
Francesc Uribe, Museu de Ciències Naturals, Catalonia, Spain, ORCID 0000-0002-0832-6561
John Waller, GBIF, Denmark, ORCID 0000-0002-7302-5976
John R. Wieczorek, University of California, Berkeley, USA, ORCID 0000-0003-1144-0290

Overview and objectives

The overall goal of the workshop was to uncover the reasons behind the insufficient quality of georeferencing data within Natural History Collection records in GBIF (tagged *PRESERVED_SPECIMEN* for the Darwin Core field *basisOfRecord*). The concrete objectives of the workshop were set to specifically come up with answers to the following two questions:

1. *What are the reasons why despite the existence of quality guidelines, protocols and tools and the investment of resources on georeferencing, georeferencing data on final public repositories, mainly GBIF, is not of sufficient quality for research purposes?*
2. *What actions can be taken to solve this situation?*

The workshop consisted of 4 morning and afternoon sessions spread over 3 consecutive days (see planned agenda). The first three sessions were organized as an informal set of 13 presentations intermingled with questions and discussions. The last session was devoted to wrapping up the main ideas and brainstorming possible answers to the two questions above that defined the workshop objectives. Each presentation lasted around 20 minutes and covered the role of natural history collections in ecological research, the quality of data at public repositories, the protocols, guidelines and tools available for the georeferencing community and, several cases of georeferencing practices at six different institutions across Europe. Presentations are available by clicking on the corresponding link in the Summary of the workshop presentations section. In the final session, a set of reasons were collectively collated as possible answers to the first question and, then, a set of actions were devised which could potentially represent a solution to them. Finally, proposed actions were prioritised and discussed not only upon their importance but also about their feasibility in the short or mid-term.

This document is structured in a way of increasing detail. The next section, Workshop wrap-up and recommendations, gives a concise view of the workshop's results and recommendations, which together with this overview can be used as an executive summary of the workshop. Readers who want to have a detailed account of the workshop can go on and read the section Summary of the workshop discussions which describes the discussions held during the session, section Summary of the workshop presentations for a brief account of each presentation with links to the presentation files handed by the presenters, and Appendices 1, 2 and 3 for further detail.

Workshop participants

Organizers: Arnald Marcer (CREAF, Spain), Quentin Groom (Botanic Garden Meise, Belgium) and Elspeth Haston (Royal Botanic Garden Edinburgh, UK)

Session chairs: Arturo Ariño (University of Navarra, Spain), Arnald Marcer (CREAF, Spain), Elspeth Haston (Royal Botanic Garden Edinburgh, UK) and Arthur D. Chapman (Australian Biodiversity Information Services, Australia)

Attendees: Arturo Ariño, Torkild Bakken, Paul Braun, Arthur D. Chapman, Mathias Dillen, David Fichtmüller, Elspeth Haston, Laurence Livermore, Arnald Marcer, Nicky Nicolson, Deborah Paul, Kaloust Paragamian, Lars B. Petterson, Sarah Phillips, Jack Plummer, Heimo Rainer, Isabel Rey, Joaquim Santos, John Waller, John Wieczorek (remotely).

Non-attending collaborators: Agustí Escobar, Markus Ernst, Tim Robertson, Dominik Röppert, Francesc Uribe

Workshop wrap-up and recommendations

Following is an account of what was concluded in Session 4 of the workshop; the final wrap-up and discussion session. This session was led as a brainstorming session whose aim was to give definite answers to the two questions (see above) around which the workshop was organized. Ideas and suggestions were written on a whiteboard and collectively edited and then prioritised. First, a list of reasons which may answer question 1 (Q1) were collated. Second, a list of actions to tackle the previous reasons were determined when answering question 2 (Q2). And third, a list of immediate tentative short-term actions was devised.

Reasons to Q1 by topic

Awareness on the importance of georeferencing

- Not seeing the wider community needs before internal project needs.
- Georeferencing being done only as an adhoc component of a major project and not given sufficient consideration by itself.
- Not perceiving the need for specific resource allocation.
- Unawareness of existing protocols and guidelines.
- Limited appreciation of the importance of georeferencing and the use of uncertainty information.
- Ignorance of the minimum essential georeferencing requirements.

Collection Management Systems and Databases

- Prioritization of needs, services and functionalities decided by CMS vendors.
- Lack of georeferencing fields and access to standards.
- Lack of consensus on standardisation practices.
- Poor user friendliness and ease-of-integration with other tools.
- Lack of automated pipelines with other tools.
- Low interoperability for moving data between different systems.
- Loss of data when moving them because of different database schemas.
- Not being able to do bulk edits in an efficient way.
- Not being able to deal with the issue of *time* (collection event).
- Not being able to geographically group or cluster records.

Work load

- Duplicated efforts among institutions or even collections. Site names may be georeferenced multiple times across collections and/or institutions, with the resulting waste of resources.
- Information is lost along the geoprocessing workflows, thus incrementing the amount of work needed.
- Georeferencing is a costly process. Imaging and databasing always take precedence.
- Lack of prioritization practices which would enable that, at least, some quality georeferencing gets done within datasets; which otherwise may not be georeferenced at all.
- Crowdsourcing and work-load sharing practices are not well established.
- Insufficient institutional oversight and coordination, probably due to ownership issues.

Tool friendliness

- Georeferencing tools need to improve their friendliness both to end users and developers.
- There is a need for the different existing georeferencing tools to be interoperable in order to share data more easily.

Geographic features

- Lack of gazetteers, *i.e.* geographical dictionaries or directories used in conjunction with maps, with the capacity to contain shapes and to be open to collaboration.
- Not enough geographic features which are both georeferenced and widely available.

- The temporal dimension is not present in the existing gazetteers. These should be able to deal with historical site names.
- Gazetteers should be able to be queried at a global and local scale, probably with different degrees of precision.
- Gazetteers would need to allow personal configuration so that users can have their *My features* section.

Actions to tackle Q2 by topic

The group first came up with a list of categories or types of actions to be considered. Then, since not all of them could be explored due to time constraints, a subset of them were prioritised and actions were outlined. The non-prioritised list of topics which were recognised as needing effective actions dealt with: resource availability, homogenization of tools, centralised support, visibility of georeferenced data, development of automatic tools, feedback mechanisms, better databases, software development, user stories compilation, tool flexibility, setting minimum requirements for georeferencing practices and building a one-stop shop for georeferencing. Out of these, the following types were prioritised and a set of actions were devised for each type:

Resource availability

- Creation of regional, national and global gazetteers.
- Promotion of proposal submission to SYNTHESYS+ project Virtual Access calls for georeferencing.
- Design and development of crowdsourcing platforms and projects.
- Creation of a volunteer program within institutions, when possible.
- Finding a mechanism that harnesses the finding that georeferencing quality improves when georeferencers are closer to site names; *i.e.* the more they know the geography the better the georeferences are.
- Exploration of the possibility of funding via sponsorship programs from corporate partners and, via crowdfunding.
- Including georeferencing actions within other project proposals submitted for funding.
- Establish criteria for prioritising specimens for georeferencing in order to make the process more efficient.

Centralised support

- Organization of institutional support programs for georeferencers.

- Pool existing sites into a single global hub for georeferencing information, including iDigBio resources, georeferencing.org, existing wikis and the Darwin Core Hour for Georeferencing.

Development of automation tools

- Develop a tool capable of converting text strings with georeference information into georeferenced localities with uncertainty information.
- Revitalize the BioGeomancer project (link 1, link 2) and include parsing of georeferencing clauses. Pool efforts with the Naturalis initiative.
- Develop tools for bulk processing.
- Add the following functionalities to georeferencing software tools:
    - Operation from local to global scale.
    - Handling of temporal, *a.k.a.* historical, site names.
    - Conversion of gridded data into points with uncertainty.
    - Checking of entries against a geographic backbone.
    - Integration of auxiliary resources, *e.g.* maps, literature, field books, etc.
- Evaluate and develop the list of tools proposed by John R. Wieczorek.

Better databases

- Conduct a survey on georeferencing capabilities across CMSs within DISSCO participants.
- Create a document on specifications and requirements for databases and CMSs to accommodate georeferencing. This document can then be used to tie a contract with service providers as in TDWG tenders. An important point on this action will be to first decide how to coordinate it between DISSCO, iDigBio and ALA.
- Solve the heterogeneity of databases among institutions by standardizing them and making them interoperable.
- Make clear the different strengths of CMSs in relation to different users and tasks.
- Create user groups to interchange experiences on the use of CMSs in relation to georeferencing.
- Establish tight collaboration between informatics departments in NHC institutions and CMS developers and vendors. Due to pressing needs, it would be desirable as a middle step to first urge for improving the basic requirements before deeper refinements are started. Tool improvement should include helping documentation for data entry fields, warnings via pop-ups for detected errors, and data validation practices on a per-field basis.

User stories compilation

- Document and disseminate successful (and unsuccessful) user stories with their lessons learned in order to encourage adoption of better georeferencing practices.
- Make these user stories available at the GBIF website to reach a maximum audience.
- Compile a document outlining the necessity of using uncertainty data when conducting research and listing successful use cases.

Tentative list of short-term next actions

This is a list of the most immediate actions that could potentially be executed.

- Explore the possibility to create a gazetteer with geonames with Globally Unique Identifiers (GUIDs) and shapes in wikidata format (link 1, link 2).
- Evaluate the convenience and feasibility of creating a reference wiki on resources for georeferencing, including the webinars of Darwin Core Hour series.
- Promote the adoption of the point-radius method in the georeferencing protocols of NHC institutions, including data validation in the uncertainty fields.
- Perform a baseline study on uncertainty in GBIF records and measure progress on the quality of georeferencing across time.
- Organize virtual update meetings on the progress of agreed actions.
- Promote SYNTHESYS+ Virtual Access calls.
- Submit, by members of the group, proposals to SYNTHESYS+ Virtual Access calls.
- Conduct a survey on georeferencing user stories and functional requirements aimed at users of Collection Management Systems.
- Draw up a georeferencing requirements doc for Collection Management Systems.
- Incorporate georeferencing in the Specimens Data Refinery work package within SYNTHESYS+.

Summary of the workshop discussions

Discussion pivoted on the concept of uncertainty of georeferenced locations, which determines georeferencing quality, the main theme of the workshop. There was unanimous consensus on the importance of having georeferenced locations documented as precisely and completely as possible and on how fundamental this is for research in areas such as ecology, conservation or evolution and all their derivatives. However, it was also recognised that, currently, there is a quite large gap between the best possible practices and practices currently being carried out in most institutions. The main cause for this stems from the fact that quality georeferencing is a very costly process and full-automation is still beyond reach.

Fully automated workflows are still not in sight for the near future and manual curation of data is nevertheless necessary to generate quality datasets. Georeferencing was referred to as the most expensive data enhancement protocol within the whole digitization process; it has been quantified in some instances to take as much as three times longer than imaging and databasing combined. It was broadly agreed that the main culprits for this kind of situation were: a general lack of resources in terms of trained personnel, available funding and time; a degree of unawareness and lack of appreciation on the importance of recording coordinate uncertainty among the community; and, software tools not meeting the user-friendliness and quality that the task at hand demands.

With respect to the first point of lack of resources there is a simple solution: allocate the necessary resources! Yet, in order to allocate these resources, the only solution is to increase funds bound for georeferencing – crowd sourcing and volunteer programs apart, though they also need some basic funding. In this respect, two existing opportunities for funding were worth mentioning: the Virtual Access calls within the SYNTHESYS+ project which offer *à la carte* georeferencing for projects funded at either the european or national levels and; the availability of funds from GBIF to enhance the georeferencing of records (*e.g.* GBIF-Norway). Funding is a problem on another level of discussion, far beyond the scope of this workshop. Frustratingly, and to rub salt into the wound, the presentation on the NTNU University Museum in Trondheim, Norway, actually confirmed that when resources are allocated, results can be obtained. Norway is the country with the highest proportion of both georeferenced records and of records with uncertainty information, as was revealed in the presentation of a survey on georeferencing at Natural History Collections across the world. Initiatives attempting to pool georeferencing efforts across institutions would be very welcome and would mean a higher efficiency in the spending of available resources. An alleviating action would be to establish some form of record prioritisation which could redound in even greater efficiency on the use of resources in relation to the needs of the community at any given time. Often georeferencing is completed in a research project to meet a specific aim where a particular set of specimens are often selected taxonomically whereas georeferencing may be more efficient when grouping all specimens from a specific locality or collector together. However, institutions should look at putting mechanisms in place to leverage more from the work completed within these research projects.

With respect to technical knowledge on georeferencing, there was a recognition for the need of offering more training courses for georeferencers and more training for the trainers themselves. The possibility of incorporating this kind of training in undergraduate and postgraduate courses was also pointed out.

Crowdsourcing has its pros and cons. On one hand, there is the problem of being able to attract the target groups which can meet institutional target policies. On the other hand, there are some experiences which have been successful at varying degrees, *e.g.* DigiVol (link 1, link 2) from the Atlas of Living Australia and BioExplora from the *Museu de Ciències Naturals de Barcelona.* Two yet-unsolved issues in using crowdsourcing for georeferencing are finding a way to record how the process is done and, devising a tagging system which ranks the quality of georeferences – similar to the iNaturalist platform.

The reporting, by the above mentioned survey on georeferencing of NHC collections, of the existence of a fair degree of unawareness on the importance of recording uncertainty data is a highly valuable finding which needs to be taken into account. Even in the scientific literature there are very few examples which explicitly mention the impact on results that come after ignoring geographic inaccuracy when using biodiversity data from public repositories (*e.g.* Maldonado *et al.*, 2015). With respect to raising awareness on this issue, several possibilities were mentioned. Some appealed to inform about the importance of the issue at hand and the necessity of not deprioritising georeferencing. Others called for encouraging good practices by unhiding and making palpable the lack of uncertainty data within the GBIF web application interface. Still others claimed to recommend or even enforce good practices through new policy documents. Moreover, the importance of uncertainty data could be made more obvious if this was displayed somehow in the GBIF data portal; *e.g.* by showing buffers of uncertainty in maps of queries and highlighting records with complete georeference information in the results tables. Making the quality of data visible in such a hugely visited portal could potentially transform the overall vision on this issue and make users more aware of the importance of accounting for it in their final uses of the data. Furthermore, some sort of visible icon tag which ranks the quality of the georeference in each dataset or record could be used. These tags could have further implications in that they could be described as granting fitness-for-use for different purposes. Information campaigns through email-list tools, newsletters or white documents could advertise the need for quality georeferences along with the implications of not following the best available practices. A portfolio of georeferencing user stories which highlight the benefits of correctly georeferencing collections could be created. Some available user stories do exist within the DISSCO projects, although they are not written at this level of detail. As an example, a nice case was presented on the role of georeferencing in IUCN Red List Assessments; the work of Nic Lughadha *et al.* (2019) providing further illustration around the importance of accurate georeferencing to the assessment process. Another set of non-detailed user stories from results of georeferencing workshops has been compiled by Vertnet (see Appendix 2). Finally, institutional policies could be enforced for newly collected specimens to ensure that new data meets certain standards. Some sort of certificates could be devised and granted to publishers and datasets as a warranty of quality for final users (although this could also fire back if one thinks the certificate will not be obtained). Policy documents would need to enforce the use of best practices and the attachment of metadata to georeferenced records, such as the protocol followed.

With respect to software tools, there was a consensus on the lack of a friendly and efficient tool for georeferencing. BioGeomancer (Guralnick et *al.*, 2006) was mentioned as an example of a tool being developed in the right direction but which, unfortunately, ran out of funding (John R. Wieczorek informed the group). It was pointed out that it would be possible to further develop the tool from the stalled code base and, without a daunting effort, end having tools which can classify locality types automatically from text strings. Text strings could be broken down into smaller units and georeferenced independently. Similar to BioGeoMancer, Naturalis also developed a georeferencing tool, which too, has stalled. With reference to the absence of adequate tools, it was brought into attention that there can be a lot of solutions to parts of the whole puzzle of semi-automated georeferencing and data quality assessment and improvement. These could be integrated and developed into a comprehensive suite and some sort of *Locality Registry* could be supported and hosted as a service within GBIF. Moreover, additional benefits of publishing data within GBIF could be brought about by a kind of global georeferencing collaboration tool scoped within GBIF-mediated data. To build such a tool, it would first be necessary to pull together a requirements specification document which would set the standards to be met. On the other hand, it was also mentioned that it would make sense to explore the use of the millions of georeferenced data records in GBIF as a georeferencing training dataset for the new emergent AI-tools. Moreover, data mining techniques could be applied to texts within the Biodiversity Heritage Library to discover itineraries and gazetteers. In this respect, one of the presentations showed how itineraries during collection expeditions could be used to improve georeferencing quality by constraining in time neighbouring localities. A potentially important caveat of expliciting itineraries within public datasets is that sensitive locality data may become more apparent and visible.

There is a whole variety of limitations within the existing set of georeferencing tools which makes them less than fit for use in georeferencing tasks: stalled development; difficulties in using them due to installation problems; lack of expertise; institutions not allowing the installation or use in their computer systems; lack of georeferencing fields within collection management systems; lack of a unified starting site where to learn about them and access them; non-existence of locality databases in many institutions; etc. It was also debated how to build a single site with all information regarding georeferencing tools, a sort of *one-stop* *shop for georeferencing* – as named by one NHC georeferencing survey respondent. Currently, there exists the *georeferencing.org* site which attempts to pull together links to all tools and resources in a web page and make them accessible to users. However, this site is not open and, thus, can not be edited and curated by the community, which limits its sustainability and use. A more advanced option would be the creation of a wiki-type site. However, it was convened that there have been several attempts to create such sites (*e.g.* the iDigBio georeferencing wiki) and that it would be better to pull efforts into improving the already existing ones. Also, these kinds of sites are very costly to be properly maintained and tend to degrade over time. A promising initiative currently in active development is the *Darwin Core Questions & Answers Site* github site, which has a wiki page and a webinar on georeferencing. With respect to collection management systems (CMS) and their lack of support for georeferencing fields, these issues could be brought into the attention of CMS vendors by contacting them through the MOBILISE Cost Action or by the organisation of symposia within events such as the Society for the Preservation of Natural History Collections annual meetings. Georeferencing staff of NHC institutions could communicate to CMS vendors the need for including georeferencing fields, mapped to Darwin Core, within their systems, paving the way to improving those systems.

Summary of the workshop presentations

Please find the actual presentation slides in the presentations folder bundled with this workshop report. Presentations are named as 'Pr - <presenter> - <presentation title>'.

Location data in ecological research

Presented by: Arturo Ariño

Arturo's talk dealt with the importance of locational data in ecological research, *i.e.*, the ecologists' need for georeferencing of NHC to establish a link between organisms and the environment they live in. He illustrated this with some ecological studies from the literature in different areas of ecology, *e.g.* pollination, mosquito richness, climate-change driven migrations, fisheries, etc. He then analyzed word usage in over 1500 journal articles and reported that the words *diversity* and *biodiversity* were the most linked to the use of georeferenced data and that the number of papers that explicitly mention georeferencing started to steadily grow in 2005 and have continued to do so till 2020. He then went on showing the growing use of GBIF data in ecological studies and the exponential growth in data at GBIF, with also a big increment in the number of georeferenced records, *e.g. Lilliopsid* species underwent a rise from 40% georeferenced records in 1995 to up to 90% in 2020. Further on, he analyzed the relation between general keywords mentioned in the literature and countries of origin, revealing differences among them. A retrospective scrutiny of uncertainty in georeferencing was the next work presented. Among other findings, he pointed out that, despite current quality issues in the georeference of data, these are being currently added with less uncertainty than before. Also, georeferencing quality increases proportionally to the closeness of georeferencers to the georeferenced sites. Finally, he reasoned about the risks associated with species protection when disclosing occurrences with a high degree of precision and the trade-offs that need to be taken into account between being open in the release of data and protecting biodiversity.

Survey on georeferencing of Natural History Collections

Presented by: Arnald Marcer

Arnald presented the results of a survey on current georeferencing practices at institutions holding natural history collections around the world. The aim of this anonymous survey was to reveal the actual tools, guidelines and overall practices carried out by georeferencers at those institutions. Eventually, such an overview could shed light on how current practices affect georeference data quality. The survey consisted of a total of 39 questions concentrated on the georeferencing process. It also asked for the kinds and sizes of both institution's holdings and the collections being reported. The survey used the SurveyMonkey platform to invite over 4000 contacts to participate, of which 552 responded (13%). 200 respondents opted for asking for a copy of the final report. Geographically, North America (41%) and Europe (36%) accounted for the majority of respondents, while taxonomically the plant (58%) and animal (23%) kingdoms prevailed. Role-wise, curators (37%) and collection managers (22%) dominated. Overall, respondents represented about 3.4 billion specimens across their institutions and the overall number of specimens for which they were responding was on the order of 1.3 billion. 28% of collections were reported to be completely digitized, 16% to be only in analog form and the rest partially digitized. The average proportion of georeferenced records in collections was 42%, with only 5% of the collections totally georeferenced. 34% reported they were not using any georeferencing protocol, 33% developed their in-house protocol and 17% were using Chapman and Wieczorek's Best Practices for Georeferencing (2006). Microsoft Access (13%) and Specify (9%) dominated the software tool used for managing georeferenced locations, while 34% reported not using any software tool for this purpose. With respect to coordinate uncertainty, only 48% provide this kind of information. 31% do not apply any methodology to detect georeferencing errors after the fact. Finally, 76% agree that many of the georeferencing work may be duplicated between institutions since their site names are most probably repeated.

Some preliminary results on NHC georeferencing quality at GBIF

Presented by: Arnald Marcer

In this short presentation, Arnald showed some preliminary results on the state of georeferencing quality in NHC data at GBIF, *i.e.* those records tagged as *basisOfRecord=PRESERVED_SPECIMEN*. Results refer to data at GBIF as of February 2, 2019. Of a total of 159 million records, 55% had coordinates and only 28% had information on coordinate uncertainty (coordinateUncertaintyInMeters field in Darwin Core). Norway and Finland were the countries with a higher percentage of records with coordinate uncertainty, followed by Spain, Turkey, Afghanistan, Pakistan and Australia. Of the countries with over 1M (M for million) specimens in GBIF, U.S.A. had the highest number of records with coordinates (~15M) followed by Australia (~10M). Norway and Finland are the countries with a higher percentage of records with reported uncertainty. Animals and plants dominate the number of records, with ~75M and ~73M, respectively. 63% of animals, dominated by *Arthropoda*, and 49% of plants, dominated by *Tracheophyta*, are georeferenced.

Darwin Core & Georeferencing

Presented by: John Wieczoreck

John first introduced a new update on the 2006 version of the classic *Guide for Best Practices for Georeferencing* by Chapman and Wieczorek on which he and Arthur are working on. The new document was kindly shared for comments among the participants before the workshop. This new document, out of a contract with GBIF, will be accompanied by an update of the 2012 version of the *Georeferencing Quick Reference Guide* which will be authored by Zermoglio *et al.*, and by both the newly created Georeferencing Calculator (by Wieczorek and Wiezoreck) and Manual (by Wieczorek, Bloom and Zermoglio) which incorporate best practices on how to calculate uncertainties. John went on explaining the differences between prospective and retrospective georeferencing, the different concepts regarding geography, locality and georeference within Darwin Core. He clarified the fact that Darwin Core does not allow for one-to-many relationships, thus not allowing to keep multiple georeferences per locality. He outlined the different location term categories in Darwin Core (higher geographies, localities and georeferences), the types of georeferenced input terms, constraining terms and output terms, stressing the differences in georeferences based on point-radius and those based on shapes. The essential fields for point-radius are longitude, latitude, uncertainty and datum while for shapes are footprintWKT, footprintSRS and footprintSpatialFit.

Data Location Quality at GBIF

Presented by: John Waller (part of it on behalf of Tim Robertson)

John first gave a general overview of some issues currently meriting attention at GBIF such as the new data derived from metagenomics and all the associated errors that need to be dealt with, the automatic checking of errors on data quality (*e.g.* a very high number of suspicious occurrences located at exactly 0 degrees latitude and 0 degrees longitude). He mentioned that the whole backend at GBIF has recently been updated in order to improve performance and that this has been done transparently for the end-user. With respect to the automatic check for errors, it is now part of the GBIF processing of data to flag occurrences with suspicious information (*e.g.* country derived from coordinates, geodetic datum invalid, invalid coordinate uncertainty in meters, etc.). GBIF is currently developing new flags to enrich the information on errors in occurrence records (*e.g.* country, state and province centroids). Brazil, Mexico and India are the countries with most occurrences within a 1km buffer around their centroids and plants are by far the kingdom with most reported country centroids. John presented a new methodology based on nearest-neighbour clustering for detecting occurrences derived from gridded datasets. The number of gridded datasets within GBIF is around 400, most of them (228) from France. Finally, the specific location of botanic gardens, zoos and herbaria can be used to detect those occurrences which refer to these locations instead of those found in the wild.

Georeferencing & Data Quality

Presented by: Arthur D. Chapman

Arthur started by giving an introduction on how collection characteristics may affect georeference quality, both from a perspective on the information associated to them and the georeferencing processes used to put their specimens onto a map. In order to assess this, a workable definition of quality is needed. Different definitions were presented and finally these were summarized under the workable and generalizable concept term *fitness for use.* Arthur further explained the new georeferencing best practices update, previously introduced by John Wieczorek. He stressed the fact that this will be a complete revision which, among other improvements, will come with new and updated references, redefined terms (*e.g.* extent and radial), new concepts (e.g. corrected center), expanded information (*e.g.* elevation, GPS, smartphones, marine data, subterranean locations). He then dissected the different phases in the georeferencing process and the planning of georeferencing projects. Lastly, a focus was given on the issue of uncertainty with the presentation of the spatial fit concept, the incidence of how localities are described on the determined uncertainty, the necessity of testing for georeferencing errors after processing and the effects onto uncertainty of not knowing the coordinate reference system or datum.

The Georeferencing Process. An evaluation of available georeferencing tools and protocols, advantages and shortcomings

Presented by: Sarah Phillips and Jack Plummer

Sarah and Jack divided their presentation into two parts. The first one centered on a revision of software done under the ICEDIG project (Section 6 of Report on New Methods for Data Quality Assurance, Verification and Enrichment) and the second one on the application of georeferenced data to IUCN assessments. Sarah talked about various software tools used in the georeferencing process (*e.g.* R packages, GEOLocate, BioGeomancer, Georeferencing calculator, etc.) and the fact that some of them were actually unavailable or could not be installed (R biogeo and GeoNames packages, BiogeoMancer and the Edinburgh Geoparser). Problems in the use of these tools range from sustainability issues regarding their maintenance, lack of sufficient knowledge on using supporting platforms such as Github, R or APIs, institutionally custom-build pipelines, and the impossibility to fully automate the process. Sarah finally pointed out different methods to speed up georeferencing such as using collector information, collaboration among users and institutions, and enhancing collection management systems with fields for georeferencing data. Jack's presentation was about the utility of georeferenced data in IUCN Red List Assessments, emphasizing the considerable relative effort that georeferencing plays. He showed how manual georeferencing in assessments of species with restricted distributions is needed to avoid potential misclassifications, while for very common species with widespread distributions, automated georeferencing has greater potential for application. In all scenarios, clear documentation of geographic unit delineation, uncertainty and source of information is key.

(Innovative) methodologies to approach locational data quality issues

Presented by: Nicky Nicolson

Nicky presented how data mining techniques can help in the georeferencing process through the use of contextual information from collectors and collecting events. These techniques are enabled by the use of large aggregated datasets which can provide contextual information. The methodology takes advantage of collector practices and habits and uses the *recordedBy*, *eventDate* and *recordNumber* fields of Darwin Core as inputs to a clustering algorithm. Through an iterative process it is possible to cluster occurrences of preserved specimens and detect collecting trips. Further down, these collection events can be even separated according to the collecting intensity, *i.e.* intense days of collecting versus more slow-paced leisure collecting events. This methodology can help detect specimens from the same collecting event held at different institutions and georeferenced separately. This could pave the way for different institutions to collaborate in their georeferencing efforts while improving the overall quality of georeferenced data in repositories.

NHM Georeferencing & Mass Digitization

Presented by: Laurence Livermore

Laurence gave an overview of digitisation and georeferencing at Natural History Museum London. He mentioned the main issues hindering the process of digitisation and georeferencing such as the inefficiency of specimen-by-specimen georeferencing driven by insufficient organisation of collections on a geographical basis, poor implementation of locational data holding fields in collection management software and not having consensus in transcription and georeferencing software and standards. For many thousands of records transcription and georeferencing is limited to higher geographies. He presented three project examples with which he showed the high effort that georeferencing represents, the high skills needed by georeferencers for records which may be recorded in a mix of several languages, and the absence of a well-defined protocol for the georeferencing process. Crowdsourcing as a means to help solve the problem of the vast personnel resources needed was commented upon, though it was considered not efficient given that NHM public engagement programs are meant for deeper scientific endeavours. Moreover, crowdsourcing policy at NHM is meant to target a set of people with a very broad background and the experiences so far have not met this objective. On the other hand, relying on a volunteer program is considered a better option. Other options which might help to improve the georeferencing project are the enforcement of georeferencing practices by collection managers, awarding a sort of georeferencing certificate of good practices on georeferencing to institutions, and the open calls for a-la-carte georeferencing projects within the SYNTHESYS+ project.

Georeferencing natural history museum specimens

Presented by: Torkild Bakken

Torkild from the NTNU (Norwegian University of Science and Technology) University Museum in Trondheim, Norway, presented the digitisation of collections as a process which started back in 1997 as a series of distinct initiatives and which had a major push thank to both the 2006-2015 national programme Revita (Revitalise) and the building of a common national database (MUSIT). This process has resulted in the digitisation of about 90% of the estimated 1.5 million natural history specimens held at the NTNU University Museum. These collections are publicly shared only if there is locational information available. A key point in having achieved this degree of success lies in the availability of personnel resources such as the institution's staff, students and, to a high degree, an agreement with the Norwegian Labour Authority for the recruiting of unemployed people and people on work training. The infrastructure consists of a Darwin Core backbone which is exported every night to GBIF. Torkild made emphasis on the balance between the need to both follow standard guidelines and procedures and being pragmatic at the same time. He mentioned several key points which affect the quality of georeferencing: the handling of old site names and changes in the administrative organisation, the use of supplementary information such as field notes to improve accuracy, the use of centroids for vaguely defined sites, the lack of technical knowledge on georeferencing, the need for having a user handbook, keeping the verbatim locality and, being aware of the right number of decimal places which are necessary for any given degree of accuracy. Lastly, he gave the link to the map and to records of official place names of the Norwegian Mapping Authority which they use. More information on experiences from the NTNU University museum can be found here, here and here.

Identification of geographical free text information via OpenRefine

Presented by: David Fichtmüller

David gave a presentation on behalf of himself, Dominik Röpert and Marcus Ernst centred on the use of the OpenRefine software tool (https://openrefine.org) for converting messy, unstructured and multi-language free text locational data into valid georeferenced site names. He introduced the use of this tool at the *Botanischer Garten und Botanisches Museum Berlin* (BGBM) for a dataset of about 400 000 free text locations with the goal of georeferencing as many site names as possible with the minimum effort possible. After acknowledging that automatic georeferencing may be feasible for some specific datasets, he exposed the unavoidable fact that, to a varying degree, oversight and manual curation of data may always be necessary. BGBM transforms free text site names into georeferenced localities with an OpenRefine workflow which starts by first pruning the easy records and gradually dealing with the more complicated ones. It involves treating data facetted by country and locality and applying filtering and clustering actions. The process is driven forward by a curator detecting common patterns in the data that are then matched and processed using regular expressions. This is repeated until the selected subset of the data is ready for automated matching against the GeoNames API. The user-authenticated API is limited to 1000 requests per hour or 20000 requests per day. Records where coordinates are available in the collection data are validated by calculating distances to coordinates provided by GeoNames and setting a warning if a certain threshold is exceeded. From the original dataset of 414500 records, 189824 have matching names in Geonames, 35801 could not be identified and 188875 have not been processed yet. So far distinct 2363 locations names have been identified.

Georeferencing @NHM-Vienna

Presented by: Heimo Rainer

Heimo explained the case of the Natural History Museum in Vienna. He started by summarizing the legacy data which needs to be dealt with. Data is divided by organisation departments. In order of size, zoology (15.2M), geology and paleontology (5.7M) and botany (5.5M) account for 90% of a total number of specimens which exceeds 29M. Only a very small fraction of these, less than 1%, have been georeferenced. Data are of global scope in all departments and highly relevant from local to global scales. With respect to ongoing collecting activities, georeferencing is mandatory only at the NUTS Level 3. Georeferencing is done manually with the support of the online Austrian map (http://www.austrianmap.at) on a pay-per-use basis. Some highly specific referencing systems need to be dealt with such as the BMN Mercator referenced to the Ferro Meridianian, and a 3′-by-5′ cartographic grid system. Heimo presented a 1.7M 3-year project for georeferencing plant collections which had to deal with the particular working habits of each collection team and which succeeded in georeferencing 17 000 unique site names; though far short of the overall total. Historical location data is still to be processed. NHM herbarium data transferred to GBIF contains georeferencing only to some extent since georeferencing efforts were halted because of the needed effort and the lack of sufficient resources. As for future needs, Heimo pointed at repositories based on expeditions, Biogeomancer-style tools, text mining and feedback to local systems and the pooling of efforts across institutions worldwide.

Georeferencing at the MnhnL

Presented by: Paul Braun

Paul gave an overview of georeferencing tasks at the Luxembourg *Musée national d'histoire naturelle (MnhnL).* MnhnL holds a collection of about 3M zoological specimens, 110k (k for thousand) plant and fungi specimens, 40k fossils and 35k minerals and rocks, plus DNA and tissue samples. Collection data are managed using Recorder 6 and a dedicated collection module (http://www.recorder6.info) and transferred currently to GBIF and soon to the Atlas of Living Luxembourg using the IPT toolkit (https://www.gbif.org/ipt). Georeferencing from research projects is done in lat/long in WGS84 with very limited additional metadata and no information on coordinates uncertainty. Georeferencing from site names is done using topographic maps while maintaining site names' georeferences separated from specimen data. Luxembourg's national official geoportal (https://geoportail.lu/en) is frequently used to find toponyms. Uncertainty is seldom reported. Many data are georeferenced using the coordinate grid system of Luxembourg (LUGRID), which can be clearly seen in the regularly spaced occurrences in the GBIF repository map. MnhnL holds a hierarchical gazetteer of multilingual site names with no coordinate information associated. Future georeferencing tasks at MnhnL will need to make use of and cite georeferencing protocols, raise awareness for data providers on the importance of georeferencing to data and, actually use the whole location fields set in the tools already at their disposal and publicate data that is complete for Darwin Core location fields.

References

- Chapman, Arthur, and J. Wieczorek. 2006 (eds). Guide to Best Practices for Georeferencing. Copenhagen: Global Biodiversity Information Facility.

- Guralnick, Robert P, John Wieczorek, Reed Beaman, Robert J Hijmans, and the BioGeomancer Working Group. 2006. "BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data." *PLoS Biology* 4 (11): e381.

- Nic Lughadha E.M., Graziele Staggemeier V., Nogales da Costa Vasconcelos T., Walker B.E., Canteiro C., Lucas E.J. 2019. Harnessing the potential of integrated systematics for conservation of taxonomically complex, megadiverse plant groups. *Conservation Biology* **33**: 510– 521

- Maldonado, Carla, Carlos I. Molina, Alexander Zizka, Claes Persson, Charlotte M. Taylor, Joaquina Albán, Eder Chilquillo, Nina Rønsted, and Alexandre Antonelli. 2015. "Estimating Species Diversity and Distribution in the Era of Big Data: To What Extent Can We Trust Public Databases?: Species Diversity and Distribution in the Era of Big Data." *Global Ecology and Biogeography* 24 (8): 973–84.

- Wieczorek, John, David Bloom, Heather Constable, Janet Fang, Michelle Koo, Carol Spencer and Kristina Yamamoto (2012). Georeferencing Quick Reference Guide. Version 2012-10-08.

Appendix 1 - Planned workshop schedule

Note: This is the initial planning document for the workshop. The workshop followed it almost exactly except for: presentations number 4 and 5 which were collapsed into a single presentation; an extra presentation on uncertainty data at GBIF was given; and presentation number 10 for which the presenter could not finally come.

MOBILISE Workshop: Quality issues in georeferencing: From physical collections to digital data repositories for ecological research

February 10-12, 2020, Warsaw, Poland

Natural history collections represent a largely vast untapped resource for ecological research. Together, they add up to a huge repository of life on Earth collected over space and time, with the potential of helping to uncover spatio-temporal patterns in the distribution and evolution of species on Earth. Their value surpasses that of species observations in that they can offer access to analyses such as taxonomic identification, microscopic examination, genetic profiling, etc.

A crucial aspect which determines the final usefulness of this wealth of data in ecological research is the appropriate georeferencing of preserved specimens. It is of utmost importance to have the most complete data possible on the location where they have been collected, especially information on the accuracy of the coordinates. When information on location uncertainty is not available, the data record loses almost all its potential for ecological research due to the difficulty in assigning environmental data to the location.

Currently, museums and botanical gardens across the world are investing in the digitisation of their biological collections to make them available through public digital biodiversity data repositories such as GBIF. Despite the existence of protocols, guidelines and recommendations on how to translate a myriad of cases of tagged textual information into georeferenced locations, the fact is that a large amount of digital specimen data lack complete georeferencing information. Without it, datasets become useless for most ecological research such as species distribution modelling and niche estimation, rendering worthless the georeferencing efforts. Additionally, all these separate digitising efforts represent a large collective effort which could eventually be made more efficient if it were possible to pull them together into a sort of federated locally-curated gazetteers. This could make the overall collective investment of resources dedicated to georeferencing more efficient.

This workshop brings together experts from the institutions involved in the different steps of the data pipeline that guides the digitisation process, from the raw specimen to the data repository institution and the ecological researcher finally using the data. The aim of the workshop is to evaluate current georeferencing practices and data workflows in order to pinpoint which factors may be preventing the fulfillment of having complete digital georeference data and propose actions which could be taken to solve this issue.

Objectives

- Determine NHC georeferencing data shortcomings.
- Evaluate the current state of implementation of georeferencing efforts across Europe, the existing panorama of available tools, protocols, standards and cartographic resources and their possible role in the problem at hand.
- Present significant examples of digitization efforts at major European museums and botanical gardens to illustrate the georeferencing process workflow and data pipeline.
- Know first-hand the GBIF harvesting process and quality check in relation to georeference information.
- Outline possible actions and initiatives which can lead to better georeferencing data at the user end-point, *i.e.* GBIF.

Expected outcomes

- Workshop report with findings and recommendations for action.
- An action work plan for MOBILISE in relation to georeferencing.
- (Potentially) A scientific paper on the workshop results and conclusions.

Schedule

Presentations length: ~ **20** minutes.

Workshop conveners: *Arnald Marcer, Elspeth Haston*

Session 1: 10/02/2020, Afternoon 15:30 - 17:00 (1.5h)

Lead: *Arturo Ariño, Universidad de Navarra, Spain*

Workshop introduction and presentations

Very brief explanation of the workshop objectives and presentation of participants.

Setting the scene

Use of Natural History Collections data for ecological research and the problem of incomplete, inconsistent or missing locational uncertainty information. What are the consequences ? Current state of the Darwin Core standard and best practices for georeferencing. The future ahead.

1. The importance of location data in ecological research

Presenter: Arturo Ariño, Universidad de Navarra, Spain

An introduction on how important location data is in ecological research and why documented uncertainty is paramount to the usefulness of the data.

2. Current georeferencing practices at Natural History Museums and Botanical Gardens around the world. A report on a recent survey.

Presenters: Arnald Marcer, CREAF and Francesc Uribe, Museu de Ciències Naturals de Barcelona, Spain

3. Darwin Core standard for location information and Georeferencing best practices

Presenter: John Wieczorek, University of California, Berkeley, USA

Overview of the current state of Darwin Core in relation to georeferencing and an update of Chapman&Wieczorek's Best Practices Protocol document. Is there room for improvement, especially with regard to reporting uncertainty ? Georeferencing and the future of the Darwin Core.

Discussion

Aim: Set the scope for the workshop, put the georeferencing data issues in context, especially information on geospatial uncertainty.

Session 2: 11/02/2020, Morning 08:30 - 12:30 (4h with 30m break)

Lead: *Arnald Marcer, CREAF*

The Georeferencing process (raw materials and tools)

How are the diversity of sources of information converted into digital records and transferred to GBIF. Data harvesting and post-processing at GBIF and their relation to georeferencing quality. What tools, protocols, standards are used. Availability and quality of reference cartography. Gazetteers, site names lists and the need for a common curated federated list of locally versioned georeferenced site names.

4. GBIF, a view from the endpoint of the data pipeline

Presenter: Tim Robertson and John Waller, GBIF, Denmark

Overview of data harvesting and post-processing at GBIF in relation to georeferencing. What is GBIF's view on the issue?

5. GBIF's locational data quality

Presenter: John Waller, GBIF, Denmark

An overview of the actual state of georeferencing quality in GBIF's preserved and fossil specimens data. What is their degree of completeness and the major problems detected.

6. A general overview on how the nature of collection information affects quality, or should it ?

Presenter: Arthur D. Chapman, Australian Biodiversity Information Services, Australia

An overview of the plethora of different cases of original sources of location information and how they facilitate or difficult the complete georeferencing of specimens, especially including uncertainty issues.

7. An evaluation of available georeferencing tools and protocols, advantages and shortcomings

Presenter: Sarah Phillips and Jack Plummer, Royal Botanic Gardens, Kew, UK

Overview of available software tools used in the georeferencing process. The georeferencing process workflow. To what degree is automation possible.

8. Innovative methodologies to approach locational data quality issues

Presenter: Nicky Nicolson, Royal Botanic Gardens, Kew, UK

How data mining techniques can help in the detection of locational data quality issues and their improvement throughout the digitization process.

Discussion

Aim: Critically evaluate the existing tools, protocols, standards and cartographic resources. What are their shortcomings when generating complete quality georeference information. Suggest possible initiatives to improve these shortcomings. What can be improved in the Darwin Core that can positively impact georeferencing data quality.

Session 3: 11/02/2020, Afternoon 13:30 - 17:00 (3.5h with 30m break)

Lead: *Elspeth Haston, Royal Botanic Garden Edinburgh, UK*

The Georeferencing process (digitization workflows and data pipelines through case studies)

An overview of georeferencing practices at some major European institutions. What can their shared experiences contribute to improving the overall quality of georeferencing.

9. Case 1: Natural History Museum London, UK

Presenter: Laurence Livermore, Natural History Museum London

10. Case 2: Moscow University Herbarium, Moscow, Russia

Presenter: Alexey Seregin, Lomonosov Moscow State University, Moscow, Russia

11. Case 3: Botanischer Garten und Botanisches Museum Berlin, Germany

Presenter: David Fichtmüller and Dominik Röpert, BGBM Berlin, Germany

12. Case 4: NTNU University Museum, Trondheim, Norway

Presenter: Torkild Bakken, Department of Natural History, NTNU University Museum, Norway

13. Case 5: Naturhistorisches Museum Wien, Wien, Austria

Presenter: *Heimo Rainer, Staff Scientist*

14. Case 6: Musée National d'Histoire Naturelle, Luxembourg

Presenter: *Paul Braun, Digital Curator*

Session 4: 12/02/2020, Morning 08:30 - 10:30 (2h)

Lead: *Arthur D. Chapman, Australian Biodiversity Information Services, Australia*

Final wrap-up discussion and conclusions

Aim: Draw up the major points raised during the previous sessions, how they interrelate and come up with a realistic plan of initiatives to help tackle the problem. What about incomplete georeference data already at GBIF ?

Alphabetical list of participants/collaborators

1. **Ariño, Arturo** [artarip@unav.es] *Universidad de Navarra, Pamplona, Spain*
2. **Bakken, Torkild** [torkild.bakken@ntnu.no] *NTNU University Museum, Trondheim, Norway*
3. **Braun, Paul** [paul.braun@mnhn.lu] Musée National d'Histoire Naturelle, Luxembourg
4. **Chapman, Arthur D.** [biodiv_2@achapman.org] *Australian Biodiversity Information Services, Melbourne, Australia*
5. **Dillen, Mathias.** [mathias.dillen@plantentuinmeise.be] *Botanic Garden, Meise, Belgium*
6. **Escobar, Agustí**. [a.escobar@creaf.uab.cat] *CREAF, Bellaterra, Spain*
7. **Fichtmüller, David** [d.fichtmueller@bgbm.org] Botanischer Garten und Botanisches Museum, Berlin, Germany
8. **Groom, Quentin** [quentin.groom@plantentuinmeise.be] *Botanic Garden, Meise, Belgium*
9. **Haston, Elspeth** [e.haston@rbge.org.uk] Royal Botanic Garden, Edinburgh, UK
10. **Livermore, Laurence** [l.livermore@nhm.ac.uk] *Natural History Museum, London, UK*
11. **Marcer, Arnald** [arnald.marcer@uab.cat] *CREAF, Bellaterra, Spain*
12. **Nicolson, Nicky** [n.nicolson@kew.org] *Royal Botanic Gardens, Kew, UK*
13. **Paragamian, Kaloust** [k.paragamian@gmail.com] *Hellenic Institute of Speleological Research, Irakleio, Crete, Greece*
14. **Petterson, Lars B.** [lars.pettersson@biol.lu.se] *Lund University, Lund, Sweden*
15. **Phillips, Sarah** [Sarah.Phillips@Kew.org] *Royal Botanic Gardens, Kew, UK*
16. **Plummer, Jack** [j.plummer@kew.org] *Royal Botanic Gardens, Kew, UK*
17. **Rainer, Heimo** [heimo.rainer@nhm-wien.ac.at] *Naturhistorisches Museum, Wien, Austria*
18. **Rey, Isabel** [isabel.rey@csic.es] *Museu Nacional de Ciencias Naturales, Madrid, Spain*
19. **Robertson, Tim** [trobertson@gbif.org] *GBIF, Copenhaghen, Denmark*[1]
20. **Röpert, Dominik** [d.roepert@bgbm.org] Botanischer Garten und Botanisches Museum, Berlin, Germany
21. **Santos, Joaquim** [joaquimsantos@gmail.com] *University of Coimbra, Coimbra, Portugal*
22. **Seregin, Alexey** [botanik.seregin@gmail.com] *Lomonosov Moscow State University, Moscow, Russia*
23. **Uribe, Francesc** [furibe@bcn.cat] *Museu de Ciències Naturals, Barcelona, Spain*[2]
24. **Waagmeester, Andra** [andra@micel.io] *Micelio, Antwerp Ekeren, Belgium*
25. **Waller, John** [jwaller@gbif.org] *GBIF, Copenhaghen, Denmark*
26. **Wieczorek, John** [tuco@berkeley.edu] *University of California, Berkeley, USA*

[1] Presentation given by John Waller (GBIF)

[2] Francesc will not be attending the workshop meeting in Warsaw although he actively participates in the georeferencing survey analyses.

Appendix 2 - User Stories and Tools

Vertnet User Stories

This is a summary of 11 user stories from VertNet reported by John R. Wieczorek. Stories are given as short bullet points which give the gist of the wanted capacity, tool, functionality, etc. Stories are written in a casual, informal way.

User 1: As a collection manager, I want …

1. a tool to look for locations, so that I can retrieve georeferences with minimal effort and improve the quality of my database.
2. a tool to test the quality of my locations, so that I can prioritize students' work in the collection
3. a tool to georeference my locations easily, so that I can improve the quality of my collection and get more money for the collection
4. a tool to georeference historical collections, so that I can improve the quality of my collection and get more funding
5. a tool to visualize my locations, so that I can prioritize, clean and show to museum director and ask for more funding
6. to have and manage my own list of locations
7. a tool to assess the effort to georeference that part of my collection that still needs it, so that I can write a grant with a solid estimate of the cost of that part of my project

User 2: As an aggregator, I want…

1. to show the total number of distinct locations in the aggregation and how many of those
   a) have coordinates
   b) have putatively best-practice georeferences
2. to do the same as above by country of origin of the record
3. to show the number of distinct locations in the aggregation and how many occurrence records correspond with those
4. to be able to interpret incoming verbatim geographic name data to standardized geographic name data

User 3: As User X, I want…

1. a reliable tool to assess locations DQ for niche modelling and build models

2. a tool to georeference non-georeferenced locations, to use them for niche modelling.

User 4: As a User Y, I want…

1. a service to georeference the locations where I collected, to better report in my papers
2. a tool to test GBIF-mediated data geographic quality, so that I can use it (or not) in my papers.
3. a place where to store my location information, because I'm required to share the outcome of my government-funded project

a) A means to provide others with the ability to find and use my localities for research or confirmation that I have fulfilled my obligations for funding.

User 5: As a public data repository, I want to…

1. Provide a service for my users to give them an opportunity to georeference records occurrences that have not been georeferenced.
2. Provide a means to check and validate georeferences for quality and completeness when records are uploaded into my index.
3. Check my existing georeferences against other known and validated georeferences for comparison and quality control.
4. Provide our set of georeferences to the public for use in research and data quality management
   a) Ability to name or tag my dataset so that the GBIF aggregated locality set can be located and used
   b) A means to track usage of the GBIF locality set
5. Have access to the full set of georeferenced localities to identify gaps in research activity, gaps in knowledge for specific areas.
   a) To use the above as an argument to encourage specific regions or counties to participate in GBIF.
   b) To be able to make gaps public so that researchers, students, etc., can easily identify areas of interest for research and exploration.

User 6: As a georeferencing project manager, I want to…

1. Identify accurate localities described in my project or to target locations for field operations.
2. Provide localities with coordinates, etc., to all members of the project for continuity and standardized descriptions/locations.

3. Check localities described/coordinates recorded in the field for accuracy.
4. Manage a list of localities specific to my project that can be accessed and updated only by team personnel.
    a) Create a printable/plottable list or map of project localities as a whole or in subsets for reporting or planning.
    b) Compare localities recorded taken over time (e.g., year to year) or by multiple teams in the field.
5. Upload results to a repository for future projects or institutional use and/or to comply with local laws or institutional/funder requirements to make project "products" available to the public.
    a) Create a project or institution-based list or portal for my project's products. (users can find my dataset and use it)

User 7: As a museum, I want…

1. to show museum geographical representation, so that I can show its importance to potential funders.
2. a means to prioritize where the money goes among collections, to increase value/ money
3. See Collection Manager story above.

User 8: As a ministry of environment I want to…

1. See if areas/locations under my control or within my jurisdiction have been georeferenced and by whom.
2. Know how accurate our ministerial georeferences are.
3. Know if any of the georeferences provided represent specific locations of sensitive or protected areas and how to
    a) Generalize protected localities
    b) Contact people using localities within protected areas
4. Comply with local/national/funder-driven requirements or laws for the public posting and availability of data funded by tax money or other public awards and/ or by foundations that require the public dissemination of data.
5. - Create a project or institution-based list or portal for projects, areas, and products under ministry jurisdiction.
6. Have a resource where I can acquire accurate localities, perhaps from my own projects, for the purpose of presentations and public documents.

User 9: As administrator of the production-level Locality Service I want to…

1. hear problems and pitfalls from real users
2. hear from real users what they actually need that the system can't do
3. be able to monitor the usage of the Service
4. not have to worry about load on the system
5. not have to do system administration on servers
6. be able to run a test suite that tells me that all is functioning as expected

User 10: As a web developer I want to…

1. use an API to the Locality Service to build pretty web pages with summary data about Locations in the Service

User 11: As a georeferencing trainer I want to…

1. use the same tools to teach the concepts as are used in actual georeferencing
2. have tools that are simple and clear to facilitate training as well as consistent results across users

List of desirable tools (compiled by John Wieczorek)

- Textual locality geographic feature extraction (multilingual) (get the named place out of the longer description)
- Unambiguous geographic feature interpretation/standardization (multilingual) (non-naive geography standardization to current geography using the DwC geography terms - see https://github.com/tucotuco/DwCVocabs)
- Spatial representations of geographic features (gazetteers that can give back shapes, such as the OpenStreetMap relations)
- Textual locality translator (multilingual to English) (this would be used as a step in locality interpretation, once in English, the english parsers and natural language engines could be used)
- Textual locality typifier (English) (once clauses are parsed out in English, determine the locality type, and therefore the georeferencing algorithm to use)
- Gazetteers of features from all real-world biodiversity data (capture geographic features that are found in biodiversity data, as we know those are used, get them into shapes, and therefore make automated georeferencing feasible for those)
- Location scrubber (enough to make a location interpretable, clustering tool?) (research needed to know if cleaning localities actually helps or not, and if so, what kind of cleaning helps)
- Location storage (a global registry and Location identifier minter for Locations and their georeferences - a resources that could be checked for existing georeferences following best practice)
- Location discovery (search into the registry described above)
- Location annotator (means to add commentary, or even georeferences, to localities in the above-mentioned registry)
- Location quality assessor (a tool that takes a Darwin Core Location as input, reports on all Tests and Assertions on it, and reports on georeferencing fitness for use against common uses)
- Collector itinerary locality constrainer (given a locality, date, collector, and collector number as input, constrain the geographic scope based on "neighboring" records)
- Georeferencing effort estimator (given locations or occurrences as input, estimate how much effort it would take to produce best practice georeferences for them)

- Automatic georeference calculator (resurrect BioGeomancer)
- Visual Georeference editor (Integrate the strengths of GEOLocate with georeferencing engines that can do better than GEOLocate and that GEOLocate can call upon)