

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

## Abstract

Natural history collections constitute an enormous wealth of information of Life on Earth. It is estimated that over 2 billion specimens are preserved at institutions worldwide, of which less than 10% are accessible via biodiversity data aggregators such as GBIF. Moreover, they are a very important resource for eco-evolutionary research, which greatly depends on knowing the precise location where the specimens were collected in order to characterize the environment in which they lived. Yet, only about 55% of the accessible records are georeferenced and only 31% have coordinate uncertainty information, which is critical for conducting rigorous studies. The awareness of this gap of knowledge which hinders the enormous potential of such data in research led to the organization of a workshop which brought together key players in georeferencing of natural history collections. The discussion and outcomes of this workshop are here presented.

Natural history collections are a superb record of life on Earth (Holmes et al., 2016). In contrast to simple observations of occurrence, physical samples held in museums, herbaria and other institutions, allow support for reproducible and repeatable research and for new data extraction from the collected individual or sample (e.g., molecular or genetic markers) on a much richer scale than other kinds of representation; *that is* photographs (but see Lunghi et al., 2020). Furthermore, and as with other observations of occurrence, their recorded date and place of collection makes it possible to link them to the abiotic and biotic conditions in which they lived. For this, one can infer spatio-temporal ecological and evolutionary patterns in their occurrence. The global set of preserved specimens collected over centuries represents a large potential resource for future research (National Academy of Sciences, Engineering, and Medicine, 2020). Some estimations on the total number of preserved specimens held in institutions worldwide (e.g. natural history museums and herbaria) are in the order of 2 billion (Ariño, 2010).

In the last decade, a myriad of digitization initiatives, combined with the growth of computer and information technologies, have yielded a growing stream of data flowing from natural history collections institutions into aggregators, such as the Global Biodiversity Information Facility (GBIF) and the Ocean Biodiversity Information System (OBIS). GBIF is an international, publicly funded research infrastructure that plays a key role in channelling these data to end users, mainly researchers. As of November 2020, around 11.6% of GBIF records come from natural history collections. The task of digitization is gargantuan and, despite all this work, accessible digital specimen records still represent, at most, only about 10% of the collection holdings worldwide. Most digitization has been funded and taken place in data-rich regions such as Europe, the Americas and Australia.

Moreover, it is crucial for research, notably in species distribution and ecological niche modelling for biogeographical, evolutionary and conservation studies, that these records have been georeferenced, a process by which geographical coordinates are assigned to physical specimens that only have a textual description of their geographic origin. Special consideration needs to be given to sensitive data in order to prevent potential threats to biodiversity (Chapman, 2020; Lunghi et al., 2019; Tulloch et al., 2018). The rigorous resolution of the coordinates where the specimen was collected, together with their uncertainty, is paramount to correctly characterize the environmental conditions and the habitat where an organism lived. It determines the spatial resolution at which research can be safely conducted. Yet, only about 55% of published records purporting to be specimens in GBIF have coordinates and only 31% of these have uncertainty information. In OBIS, all records are georeferenced but also only 31% have coordinate uncertainty.

When coordinates are present, but their spatial uncertainty is not, it is not always possible to rigorously extract useful information from environmental datasets. Regrettably, it is still not unusual to find research studies using such data which have overlooked the need for coordinate uncertainty values. Both the lack of information on spatial uncertainty in georeferenced specimens and the disregard of it on the part of the researchers represent an obstacle to the proper and full exploitation of collections data.

Georeferencing is a skilled, labour-intensive process which is hard to automate. It generally starts with the interpretation of the

documented location information, which in most cases is hand-written on labels. Locations can be described in multiple and idiosyncratic ways; from clearly detailed and precise places to vaguely defined and sometimes large regions. However, despite its complexity, georeferencing is a well-researched process for which clear and detailed guidelines (e.g., Chapman & Wieczorek, 2006, 2020; Wieczorek et al., 2004) and information standards (Darwin Core Task Group, 2009) have long existed and are known by the collections community. Yet the speed of georeferencing is still slow, and there is a need for training, particularly among smaller collections without digitization experience.

Last February, we held a workshop to discuss the state of georeferencing quality of natural history collections as a critical issue for ecological research (for a detailed account of its outcomes, here summarized, see Marcer et al., 2020). The workshop brought together key players in the study and application of georeferencing to biodiversity collections in order to explore the reasons behind the insufficient quality of georeferenced records in data aggregators such as GBIF. To focus the discussion, the participants were given the following two questions, which were analysed and debated in four sessions in two days:

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

From the workshop, it became clear that no single cause can be attributed to this situation. In response to the first question above, the participants converged on a list of different types of causes leading to the current situation:

- a. Awareness-related—the need for the collections community to better appraise the importance of quality georeferencing through the use of current existing guidelines and standards;
- b. Collection management systems and databases—most of them are still not fit for the purpose probably due to a lack of sufficient dialog between software vendors and the user community;
- c. Staff workload—digitization is a time consuming process and georeferencing is often of low priority;
- d. Tool friendliness—georeferencing tools still require improvement in terms of user friendliness and interoperability;
- e. Geographic features—there is a lack of publicly shared, global, hierarchical, time-aware, community-vetted geographical directories, gazetteers.

After much debate and discussion and in response to question two above, a list of needed actions were identified and prioritized in the following categories:

- a. Resource availability—it is essential to create shared gazetteers, formulate crowdsourcing and volunteer programs, and make better use of funding while searching for additional funds;

- b. Centralized support—provide institutional support programs and centralized information resources to georeferencers;
- c. Automated tools—there is a need to review and enhance existing software tools and develop new ones to enable bulk text processing and interpretation; a cost-effective option would be to start from existing codebases (e.g., the Biogeomancer project (Guralnick et al., 2006));
- d. Better databases—Collection management software and databases need to be enhanced with georeferencing capability by means of a two-way dialog between software vendors and the user community; and,
- e. User stories—there is a need to compile, document and disseminate concrete working experiences from the georeference community which can influence improved georeference practices.

Natural history collections have already had a massive impact by documenting life on Earth. With this letter, we make a call to the global collections and research communities to pull together and refine current procedures towards improving georeferencing and research practise. A joint effort will allow us to move forward and capitalize on the enormous wealth of information that natural history collections represent. The development of accurate and thorough georeferencing tools and protocols, and the rigorous use of the generated data in research can be a means to integrate communities with benefits for all. Natural history collections represent a unique science infrastructure which can enable novel and larger scale uses of the global collections resource, delivering vital research and public interpretation.

## KEYWORDS

eco-evolutionary research, global biodiversity information facility, georeferencing, natural history collections, uncertainty, workshop

## ACKNOWLEDGEMENTS

This work has been possible thanks to the EU Cost Action CA17106: “MOBILISE. Mobilizing Data, Experts and Policies in Scientific Collections.” We would also like to acknowledge the facilities and hospitality provided by the personnel hosting the event (February 2020) at the Biological and Chemical Research Centre of the University of Warsaw in Poland.

- Arnald Marcer<sup>1,2</sup>   
 Elspeth Haston<sup>3</sup>   
 Quentin Groom<sup>4</sup>   
 Arturo H. Ariño<sup>5</sup>   
 Arthur D. Chapman<sup>6</sup>   
 Torkild Bakken<sup>7</sup>   
 Paul Braun<sup>8</sup>   
 Mathias Dillen<sup>9</sup>   
 Marcus Ernst<sup>9</sup>   
 Agustí Escobar<sup>1</sup>   
 David Fichtmüller<sup>9</sup>   
 Laurence Livermore<sup>10</sup> 

- Nicky Nicolson<sup>11</sup>   
 Kaloust Paragamian<sup>12</sup>   
 Deborah Paul<sup>13</sup>   
 Lars B. Pettersson<sup>14</sup>   
 Sarah Phillips<sup>11</sup>   
 Jack Plummer<sup>11</sup>   
 Heimo Rainer<sup>15</sup>   
 Isabel Rey<sup>16</sup>   
 Tim Robertson<sup>17</sup>   
 Dominik Röpert<sup>9</sup>   
 Joaquim Santos<sup>18</sup>   
 Francesc Uribe<sup>19</sup>   
 John Waller<sup>17</sup>   
 John R. Wiczorek<sup>20</sup> 
- Elsbeth Haston*  <https://orcid.org/0000-0001-9144-2848>  
*Quentin Groom*  <https://orcid.org/0000-0002-0596-5376>  
*Arturo H. Ariño*  <https://orcid.org/0000-0003-4620-6445>  
*Arthur D. Chapman*  <https://orcid.org/0000-0003-1700-6962>  
*Torkild Bakken*  <https://orcid.org/0000-0002-5188-7305>  
*Paul Braun*  <https://orcid.org/0000-0002-3620-6188>  
*Mathias Dillen*  <https://orcid.org/0000-0002-3973-1252>  
*Agustí Escobar*  <https://orcid.org/0000-0002-6856-0480>  
*David Fichtmüller*  <https://orcid.org/0000-0002-0829-5849>  
*Laurence Livermore*  <https://orcid.org/0000-0002-7341-1842>  
*Nicky Nicolson*  <https://orcid.org/0000-0003-3700-4884>  
*Kaloust Paragamian*  <https://orcid.org/0000-0001-7372-733X>  
*Deborah Paul*  <https://orcid.org/0000-0003-2639-7520>  
*Lars B. Pettersson*  <https://orcid.org/0000-0001-5745-508X>  
*Sarah Phillips*  <https://orcid.org/0000-0002-9155-8573>  
*Jack Plummer*  <https://orcid.org/0000-0002-1575-5241>  
*Heimo Rainer*  <https://orcid.org/0000-0002-5963-349X>  
*Isabel Rey*  <https://orcid.org/0000-0002-2122-5124>  
*Tim Robertson*  <https://orcid.org/0000-0001-6215-3617>  
*Dominik Röpert*  <https://orcid.org/0000-0001-6565-8450>  
*Joaquim Santos*  <https://orcid.org/0000-0002-2160-4968>  
*Francesc Uribe*  <https://orcid.org/0000-0002-0832-6561>  
*John Waller*  <https://orcid.org/0000-0002-7302-5976>  
*John R. Wiczorek*  <https://orcid.org/0000-0003-1144-0290>

<sup>1</sup>CREAF, E 08193, Bellaterra (Cerdanyola del Vallès), Catalonia, Spain

<sup>2</sup>Universitat Autònoma de Barcelona, E 08193, Bellaterra (Cerdanyola del Vallès), Catalonia, Spain

<sup>3</sup>Royal Botanic Garden Edinburgh, Edinburgh, UK

<sup>4</sup>Meise Botanic Garden, Meise, Belgium

<sup>5</sup>Universidad de Navarra, Pamplona, Spain

<sup>6</sup>Australian Biodiversity Information Services, Melbourne, Victoria, Australia

<sup>7</sup>Norwegian University of Science and Technology, NTNU University Museum, Trondheim, Norway

<sup>8</sup>Musée National d'Histoire Naturelle, Luxembourg City, Luxembourg

<sup>9</sup>Botanic Garden and Botanical Museum Berlin, Freie Universität Berlin, Berlin, Germany

<sup>10</sup>Natural History Museum, London, UK

<sup>11</sup>Royal Botanic Gardens, Kew, UK

<sup>12</sup>Hellenic Institute of Speleological Research, Crete, Greece

<sup>13</sup>Florida State University, iDigBio, Tallahassee, Florida, USA

<sup>14</sup>Biodiversity Unit, Department of Biology, Lund University, Lund, Sweden

<sup>15</sup>Naturhistorisches Museum Wien, Vienna, Austria

<sup>16</sup>Museu Nacional de Ciencias Naturales (CSIC), Madrid, Spain

<sup>17</sup>GBIF, Copenhagen, Denmark

<sup>18</sup>University of Coimbra, Coimbra, Portugal

<sup>19</sup>Museu de Ciències Naturals, Barcelona, Spain

<sup>20</sup>University of California, Berkeley, California, USA

#### Correspondence

Arnald Marcer, CREA, E 08193, Bellaterra (Cerdanyola del Vallès), Catalonia, Spain.

Email: arnald.marcer@uab.cat

Editor: Zhixin Zhang

#### REFERENCES

- Ariño, A. H. (2010). Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, 7(2), 81–92. <https://doi.org/10.17161/bi.v7i2.3991>
- Chapman, A. D. (2020). *Current best practices for generalizing sensitive species occurrence data [Community review draft]*. Copenhagen, Denmark: GBIF Secretariat. <https://doi.org/10.15468/doc-5jp4-5g10>
- Chapman, A. D., & Wiczorek, J. (Eds.). (2006). *Guide to best practices for georeferencing*. Copenhagen, Denmark: GBIF Secretariat. <https://doi.org/10.15468/doc-2zpf-zf42>
- Chapman, A. D., & Wiczorek, J. (2020). *Georeferencing best practices [Community review draft]*. Copenhagen, Denmark: GBIF Secretariat. <https://doi.org/10.15468/doc-gg7h-s853>
- Darwin Core Task Group. (2009). *Darwin Core (Kampmeier G, review manager)*. Biodiversity Information Standards (TDWG). <http://www.tdwg.org/standards/450>
- Guralnick, R. P., Wiczorek, J., Beaman, R., Hijmans, R. J. & the BioGeomancer Working Group. (2006). Biogeomancer: Automated georeferencing to map the world's biodiversity data. *PLoS Biology*, 4(11), e381. <https://doi.org/10.1371/journal.pbio.0040381>
- Holmes, M. W., Hammond, T. T., Wogan, G. O. U., Walsh, R. E., LaBarbera, K., Wommack, E. A., Martins, F. M., Crawford, J. C., Mack, K. L., Bloch, L. M., & Nachman, M. W. (2016). Natural history collections as windows on evolutionary processes. *Molecular Ecology*, 25, 864–881. <https://doi.org/10.1111/mec.13529>
- Lunghi, E., Corti, C., Manenti, R., & Ficetola, G. (2019). Consider species specialism when publishing datasets. *Nature Ecology & Evolution*, 3, 319–319. <https://doi.org/10.1038/s41559-019-0803-8>
- Lunghi, E., Giachello, S., Zhao, Y., Corti, C., Ficetola, G., & Manenti, R. (2020). Photographic database of the European cave salamanders, genus *Hydromantes*. *Scientific Data*, 7(1), 171.

#### ORCID

Arnald Marcer  <https://orcid.org/0000-0002-6532-7712>

- 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., ... Wieczorek, J. R. (2020). *Quality issues in georeferencing: From physical collections to digital data repositories for ecological research*. Workshop report. MOBILISE EU Cost Action CA1706. <https://doi.org/10.5281/zenodo.3734848>
- National Academy of Sciences, Engineering, and Medicine. (2020). *Biological Collections: Ensuring Critical Research and Education for the 21st Century*. The National Academies Press.
- Tulloch, A. I. T., Auerbach, N., Avery-Gomm, S., Bayraktarov, E., Butt, N., Dickman, C. R., Ehmke, G., Fisher, D. O., Grantham, H., Holden, M. H., Lavery, T. H., Leseberg, N. P., Nicholls, M., O'Connor, J., Roberson, L., Smyth, A. K., Stone, Z., Tulloch, V., Turak, E., Wardle, G. M., & Watson, J. E. M. (2018). A decision tree for assessing the risks and benefits of publishing biodiversity data. *Nature Ecology & Evolution*, 2, 1209–1217. <https://doi.org/10.1038/s41559-018-0608-1>
- 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. <https://doi.org/10.1080/13658810412331280211>

#### BIOSKETCH

The authors of this work are an *ad hoc* team put together with the aim of exploring and discussing the issues affecting the interplay between georeferencing practices in natural history collections and ecological research, mainly with respect to uncertainty questions. They come from natural history collection institutions, research centres and universities and GBIF. They are interested and have extensive experience in the georeferencing process of natural history collections for research.

**How to cite this article:** Marcer A, Haston E, Groom Q, et al. Quality issues in georeferencing: From physical collections to digital data repositories for ecological research. *Divers Distrib.* 2021;27:564–567. <https://doi.org/10.1111/ddi.13208>