OPEN ACCESS

CrossMark

Workshop Report

Georeferencing for Research Use (GRU): An integrated geospatial training paradigm for biocollections researchers and data providers

Katja C. Seltmann[‡], Sara Lafia[§], Deborah L. Paul[|,¶], Shelley A. James[#], David Bloom[¤], Nelson Rios["], Shari Ellis[¤], Una Farrell["], Jessica Utrup[^], Michael Yost[ˇ], Edward Davis[¦], Rob Emery[?], Gary Motz[ς,¢], Julien Kimmig[ℓ], Vaughn Shirey[ϑ], Emily Sandall[P], Daniel Park[A,ℶ], Christopher Tyrrell[F], R. Sean Thackurdeen[Ŧ], Matthew Collins[N,K], Vincent O'Leary[₵], Heather Prestridge[?,₩], Christopher Evelyn[Ŧ], Ben Nyberg[‡‡]

‡ Cheadle Center for Biodiversity and Ecological Restoration, University of California - Santa Barbara, Santa Barbara, CA, United States of America
§ Department of Geography, Center for Spatial Studies, University of California - Santa Barbara, Santa Barbara, CA, United States of America
| Florida State University, Tallahassee, United States of America
¶ iDigBio, Gainesville, United States of America
# National Herbarium of NSW, Royal Botanic Gardens & Domain Trust, Sydney, NSW, Australia
¤ Florida Museum of Natural History, Gainesville, FL, United States of America
« Yale Peabody Museum, New Haven, Connecticut, United States of America
» Stanford University, Stanford, CA, United States of America
^ Yale Peabody Museum of Natural History, New Haven, CT, United States of America
ˇ New York Botanic Garden, New York, NY, United States of America
¦ University of Oregon, Eugene, OR, United States of America
? Department of Food and Agriculture, Perth, WA, Australia
ς Indiana Geological and Water Survey, Bloomington, IN, United States of America
¢ Indiana University, Bloomington, IN, United States of America
ℓ University of Kansas, Lawrence, KS, United States of America
ϑ The Academy of Natural Sciences of Drexel University, Philadelphia, PA, United States of America
P Frost Entomological Museum at Penn State, University Park, PA, United States of America
A Harvard University, Cambridge, MA, United States of America
ℶ University of Arizona, Tucson, AZ, United States of America
F Milwaukee Public Museum, Milwaukee, WI, United States of America
Ŧ New York Botanical Garden, Bronx, NY, United States of America
N University of Florida, Gainesville, FL, United States of America
K iDigBio, Gainesville, FL, United States of America
₵ Drexel University, Philadelphia, PA, United States of America
? Texas A&M University, College Station, TX, United States of America
₩ Biodiversity Research and Teaching, College Station, TX, United States of America
Ŧ Department of Ecology, Evolution and Marine Biology, University of California - Santa Barbara, Santa Barbara, CA, United States of America
‡‡ National Tropical Botanical Garden, Kalaheo, HI, United States of America

Corresponding author: Katja C. Seltmann (enicospilus@gmail.com),
Deborah L. Paul (dpaul@fsu.edu)

Reviewable   v1

Abstract

Georeferencing is the process of aligning a text description of a geographic location with a spatial location based on a geographic coordinate system. Training aids are commonly created around the georeferencing process to disseminate community standards and ideas, guide accurate georeferencing, inform users about new tools, and help users evaluate existing geospatial data. The *Georeferencing for Research Use* (GRU) workshop was implemented as a training aid that focused on the creation and research use of geospatial coordinates, and included both data researchers and data providers, to facilitate communication between the groups. The workshop included 23 participants with a wide background of expertise ranging from students (undergraduate and graduate), professors, researchers and educators, scientific data managers, natural history collections personnel, and spatial analyst specialists. The conversations and survey results from this workshop demonstrate that it is important to provide opportunities for biocollections data providers to interact directly with the researchers using the data they produce and vice versa.

Keywords

GIS, Geospatial data, Natural history collections, Biocollections, Workshop report, Data quality, iDigBio, Georeferencing, GeoLocate, QGIS

Introduction

Scientific knowledge relating to our environment, human health, climate change and global ecosystems increasingly requires the creation and evaluation of diverse and varied datasets, warranting training to raise the data competencies of researchers and data providers (Hampton et al. 2017, Hampton et al. 2013, McCulloch 2013). One of the important data sources for ecological data is from specimens found in natural history collections (Hampton et al. 2013, iDigBio 2016, Lister 2011, Otero-Ferrer et al. 2017, Shaffer et al. 1998, Arnaud et al. 2016, McGeoch et al. 2016, Krishtalka et al. 2016). Natural history collections, or biocollections as defined in this paper, include any specimen-based biological, zooarchaeological, or paleontological collections (Lieberman and Kimmig

2018). The Georeferencing for Research Use (GRU) workshop was developed jointly by Integrated Digitized Biocollections (iDigBio), UC Santa Barbara Cheadle Center for Biodiversity and Ecological Restoration (CCBER), VertNet, Denver Botanic Gardens, Yale Peabody Museum, Stanford Earth and the GEOLocate project to provide an innovative, integrated forum for discussion between data researchers and data providers about the use and critical creation of geospatial coordinates from biocollection specimens, one aspect of biocollection data. This paper reports our findings and observations, and captures the participants' discussion around changing museum practices, use of error ranges in geospatial data, and needs for continued training and tool development.

Background: Importance of biocollections georeferenced data

Information about the distribution of biological organisms is used to investigate many current global health and human services such as clean water, land preservation and restoration, disease prevention, food safety and security, agricultural pests, drought, land use management, urban planning, and the effects of climate change (Kearney et al. 2014, Sinervo et al. 2010, Arnaud et al. 2016, McGeoch et al. 2016, Nature Publishing Group 2008, Department of Primary Industries and Regional Development: Agriculture and Food 2018). Researchers use locality information associated with vouchered biocollections specimens to evaluate the biological requirements for organisms by linking its presence with many other kinds of data (i.e., climate layers, satellite imagery, and other data resources) (Kearney et al. 2014, Sinervo et al. 2010). In general, researchers create, retrieve, combine, assess quality, clean, and visualize geospatial data before they apply their research methods, such as gap analysis (e.g., taxonomic or geographic, Ariño et al. 2016), species distribution models (e.g., MaxEnt, Anderson et al. 2016) or International Union for Conservation of Nature (IUCN) conservation assessments (Brummitt et al. 2015). Locality information is found on the labels, field books, logs, photographs, etc. that accompany and describe specimens in biocollections. Today, researchers apply geocoordinates immediately when an organism is collected, but before we had readily accessible and reliable Global Positioning System (GPS) units, only a text description of the locality was provided. To make these older specimens useful in modern research practices, geocoordinates (i.e., latitude, longitude) are retroactively applied to a text description of an organism's locality in a process referred to as "georeferencing."

Georeferenced locality data are widely made available through biodiversity data aggregators such as Atlas of Living Australia (ALA), the Global Biodiversity Information Facility (GBIF), Integrated Digitized Biocollections (iDigBio), and VertNet in a digital format. Digital data records for more than 115 million physical specimens have been shared (iDigBio 2018), leading to large amounts of digitally available locality data about organisms. Of those 115 million records, 59 million have coordinates. For example, the points on this map in Fig. 1 indicate geocoordinated specimen location data available through iDigBio for beetles in the genus *Cicindela* (Coleoptera) from California, USA, with each point representing one or more specimen records found in a collection (Fig. 2).
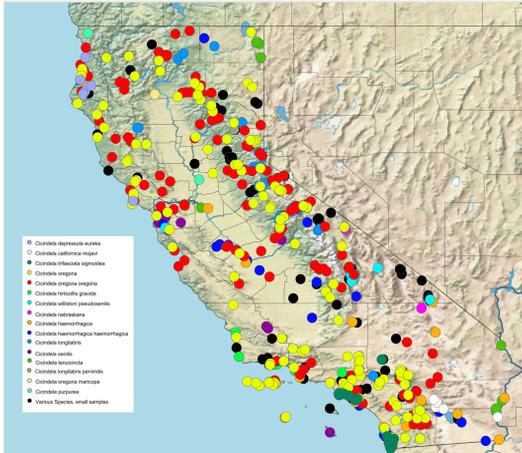
Figure 1. doi

Map created using SimpleMappr (Shorthouse 2010) that illustrates geolocated specimens for Genus=Cicindela in California as found on iDigBio.

Figure 2. doi

This specimen record is an example from the University of California Collection Network Symbiota Portal. The large image is an edit of the record to include a medium size version of the image for easier viewing in this article. The portal software is open source and it is freely available for reuse through the Symbiota GitHub repository. The image is an example of a specimen record that includes an image of the specimen with label data. The image is contributed by the UCSB Invertebrate Zoology Collection at the Cheadle Center for Biodiversity and Ecological Restoration. The usage rights for the image is Creative Commons 0 (public domain).

Rationale: Raising data competencies for researchers and data providers

The data captured from biocollection specimens are typically managed by collections staff who care for specimens and digitize the data associated with those specimens, ultimately making the data available for research. This creates a scenario where the people creating the datasets are often not the same people who use the data in research, which creates a need to provide venues of communication and foster understanding between these stakeholders about the implications of each other's' methods on research products (Tenopir et al. 2015,Zimmermann 2008). Geospatial and biocollections communities benefit from mutual conversations that inform the work of both biocollections staff and the research community, and through exposure to computational technologies, data, and software for working with spatial data. Training aids are specifically needed for biocollections staff to clean and visualize these data to assess its reliability and precision, and they also need to evaluate the data quality feedback that is provided to them after sharing their data with aggregators, which may include issues in the geocoded locality data. Researchers who plan on using the data that is produced need to understand the process of producing the datasets in order to evaluate them for their particular use.

History: Background and provenance of recent efforts in biocollection digitization

Recognizing the value of biocollections for research, education, and society, a diverse group of scientists outlined a coordinated effort they envisioned as a *Networked Integrated Biocollections Alliance* (NIBA), which resulted in strategic and implementation plans for the digitization of US collections information (Hanken 2013, Chapman and Wieczorek 2006). In response, in 2011 the US National Science Foundation (NSF) instituted the program *Advancing the Digitization of Biological Collections* (ADBC). As part of the ADBC, NSF funded the Integrated Digitized Biocollections (iDigBio) 10 year project as the hub for US non-federal collection digitization efforts. iDigBio's missions are to support efforts by biocollections to digitize, mobilize, aggregate, and provide access to biological specimen data for biodiversity research and education globally. Museums holding biocollections collaborate to form Thematic Collection Networks (TCNs) and seek NSF ADBC funding to support digitization of select collections to meet targeted, specific research needs and share this mobilized data via iDigBio.

As part of this initiative, the iDigBio Georeferencing Working Group (GWG) was formed to support the biocollections community to implement best practices in the improvement and maintenance of critical location data. The GWG benefits from the contributions from many, including GEOLocate, VertNet, TCNs, and earlier projects (e.g., Georeferencing.org, MaNIS, FishNet2, ORNIS and HerpNET).

In response to biocollection georeferencing needs, the GWG offered two Train-the-Tainers workshops in 2012 and 2013. These workshops were aimed at training biodiversity and collections professionals to use best practices for georeferencing (Bloom et al. 2017), to mobilize biocollection occurrence data, and to encourage those participants to share what

they learn with others for the benefit of the collections community as a whole. The primary audience for these workshops included the collaborative museum and herbaria ADBC TCNs staff, collection managers, and curators engaged in active transcription and georeferencing of collections data. An additional workshop, Field to Database (F2DB) (2015), targeted researchers and the collection of new, future biocollections data. F2DB incorporated current best practices for creation and sharing of georeferences and locality information to be born digital, that is mapped to data standards, if possible, in electronic format and georeferenced at the time of collection (Online Computer Library Center (OCLC) 2018). This process results in higher quality data, mobilized more efficiently, and avoids adding to the legacy pile of text locality strings to be data-entered and georeferenced. The Biocode Field Information Management System (Deck and Ewing 2016) is an example project supporting the generation of born digital biodiversity data. Other workshops, such as the joint Synthesys-iDigBio: Digitization Software Training Workshop, have similarly shared tools for biocollection data generation and standardization.

Training provided by the iDigBio project initially focused on community-derived digitization best practice discovery and documentation but has increasingly incorporated innovative use of scientific collections data for research into workshops and symposia design. The iDigBio Georeferencing Working Group (GWG) saw the need to offer a georeferencing workshop that would combine best practices for historical and new data with new lessons on evaluating the resulting biocollections data for their fitness for research and other downstream use.

Georeferencing for Research Use (GRU) Workshop Design and Implementation

To design the Georeferencing for Research Use workshop, the GWG first reviewed materials used in prior workshops (e.g., Train-the-Trainers, Field to Database) to determine what lessons about the evaluation of geospatial data should be integrated into the course curriculum. Additional content was then included by the course instructors from Biodiversity Informatics Training Curriculum (BITC), Cheadle Center for Biodiversity and Ecological Restoration (CCBER), and University of California, Santa Barbara National Center for Ecological Analysis and Synthesis (NCEAS) and iDigBio. A specimen dataset downloaded directly from iDigBio was used as the example dataset (Suppl. material 1). The dataset contained 25,429 records of ground beetles (Family: Carabidae) found in California. Accessed on 29 August 2016, this dataset was used in the data access demonstration, visualizations, data cleaning and QGIS tutorials during the workshop. Geospatial expertise outside the existing GWG group was recruited to develop a new Quantum Geographic Information System (QGIS) tutorial for biocollections data. The resulting tutorial content is now available on GitHub as a series of QGIS Natural History Collection Lessons. The final course outline can be found on the iDigBio Georeferencing for Research Use (GRU) Wiki.

Topic choice and time spent on specialized topics was in part guided by the applicants' expectations of the course. Using an online form that called for participation and evaluated expectations of participants (Suppl. material 2), applicants were asked to describe: 1) the reason(s) for their interest in the course, 2) any current and/or future projects they were involved with that would benefit from their receiving training, 3) knowledge of, and experience with, georeferencing, and 4) their ideal syllabus for a 4-day georeferencing workshop. Desired learning outcomes of the applicants were summarized and used to develop the course outline (Suppl. material 6, also see the GRU Wiki).

Participants were asked to prepare in advance for the workshop. Pre-workshop assignments included a review of materials for a sufficient level of understanding of the fundamental principles of GIS and best practice for biocollection georeferenced data, such as projection information and coordinate reference systems.

The GRU workshop was held 4-7 October 2016 in Santa Barbara, California, hosted by the NCEAS and CCBER. The first two days of the workshop training provided a summary of biocollection georeferencing and data standards, legacy collection data issues, and best practices for the creation of new locality (geospatial) data in an effort to avoid increasing the legacy-data backlog. The final two days encompassed strategies for data standardization before research use, such as the use of spreadsheets and OpenRefine (http://openrefine.org/) software to evaluate data and to select, adjust, and remove data as appropriate, and the visualization of common geolocation issues using QGIS and other tools.

Workshop Objectives

The major objectives of the workshop were to gain feedback from the participants on research data needs for the future and to enhance and improve the quality of collections geospatial data. Objectives of the 4-day workshop included:

- Demonstration of tools (hardware and software) and geospatial data standards, especially as relating to the Darwin Core standard.
- Discussion of best practices for data repositories (e.g., obstacles and minimization of data loss).
- How to evaluate already georeferenced data for quality, or the fitness of the data for a specific use (fitness-for-use).
- Current tools for visualization and evaluation.
- Introduction of open-source QGIS software and selected plug-ins to participants to demonstrate data visualization methods.
- Sharing best practices for researchers for in-the-field creation of new locality data.
- Gain insight into the challenges faced by researchers for georeferencing through shared research experiences using biocollection geospatial data.
- Gain insight into the challenges data and collection managers experience in generating and managing georeferenced data.
- Gain input from participants about their needs and learning experience during the workshop.

- Gain an understanding of how participants use their learning for research and other present and future research and curation of collections and data needs.

Participant demographics

The results presented here represent views from our participants (23 persons) who were selected for their interest in georeferencing and biocollections at the time that they registered for the workshop. All but one of the participants were affiliated with U.S institutions. The participants represented a cross-section of users of biocollections data that include students (graduate and undergraduate), professors, collections managers, curators, an agriculture specialist, and data managers.

Results and Outcomes

In total, 42 applications were received in response to the open call for participants. Twenty-three participants were accepted, with priority given to persons who had not participated in prior iDigBio workshops, and whose expectations best matched the proposed workshop goals and scope. Participants primarily self-identified as researchers or research students (14 participants) who use biocollections spatial data in their research projects, or as data/collection managers (9 participants). Several participants also self-identified as having multiple career needs for participating in the training, for example, a researcher who is also managing a biocollection, or a data manager who is also a student. The combined audience fostered collaboration and understanding across these different domains.

Conversations around GIS, QGIS, and other tools

Workshop participants self-identified as a researcher, data manager, GIS expert, or some combination of identities. The needs of each of these groups for data management and georeferencing skills were similar, yet each group had particular goals. For example, researchers were interested in the functionality of OpenRefine software for standardizing biodiversity data sets, new tools for efficient field data collection, such as the Biocode Field Information Management System, and efficient georeferencing of large datasets with high-throughput analyses of specimen coordinate accuracy. There was a strong desire for efficient and accurate georeferencing of biocollections to create "research ready" data that is available for complex, data-driven analyses (Seltmann et al. 2017). Developing the skills available through OpenRefine for data cleaning was expressed to be of great benefit for basic cleaning of collections datasets prior to publication and ingestion by data aggregators such as iDigBio and GBIF. Data visualization tools, such as QGIS, for determining the accuracy of geopoints were perceived to be of great benefit to researchers and data/collections personnel alike. Researchers were more interested in more advanced applications, such as the use of R and Python scripting due to the desired downstream applications and analyses.

Workshop participants with a strong GIS background were primarily interested in learning more about the current practices of researchers and data managers in data cleaning, georeferencing, and spatial analysis to identify areas of opportunity for improving the ability of domain experts to assess data quality and gain new insights to visualize their data spatially. The QGIS tutorial was modified in response to the interest in new skills expressed by participants in the first two days of the workshop. For example, a desire to know how to leverage coupled spatial data layers to perform a spatial selection and drill down through several layers of information to choose records that met a specific criterion, was of great interest to the participants. Also, learning how to subset data, edit record entries (e.g., coordinates) using QGIS, and save the cleaned and appropriately georeferenced records as comma-delimited (or comma-separated) text files (CSV) was of great interest and added to the tutorial.

Georeferencing among data managers and researchers in biocollections generally use the point-radius method (Fig. 3) to account for coordinate uncertainty (Chapman and Wieczorek 2006), although polygons are also commonly used for narrow geographical features (Fig. 3). In the geospatial community, bounding boxes are often used to delineate the extent of resources rather than points with a given radius. Additionally, many field record entries reference named places which can be reconciled against a gazetteer. Participants learned how this presents an opportunity for exploiting existing geometries to use as a record boundary from existing sources, such as OpenStreetMap's Application Program Interface (API), or shape files for publicly available datasets.
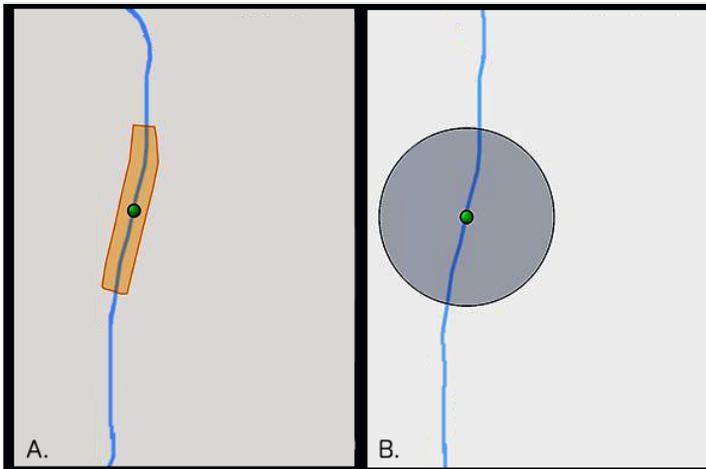

Figure 3. doi

An illustrative example of the two methods of uncertainty capture when georeferencing specimens. Method A, or polygon, creates a shape around the river (in blue). Method B, or point-radius, creates a circle of uncertainty around the origin. The illustration is based on output from GeoLocate software (Rios 2018) for both polygon and point-radius.

Researchers and data managers were taught how to import existing georeferenced collections into software such as QGIS. Once spatially referenced, it is possible to couple the georeferenced data with existing layers of spatial information that can supplement assessment and analysis. Some layers of interest included administrative boundaries (for checking centroids, or center points of polygons, against existing records), ecoregions, cultivated gardens and zoos, and parks and protected areas. Other researchers expressed interest in acquiring and incorporating data about elevation, federally managed sites, and climate data. A conversation about where to find data resources, like local municipality GIS office websites and ArcGIS Online's Living Atlas of the World, proved helpful.

On the first day of the QGIS tutorial, the topics covered included importing tutorial data obtained from iDigBio, adding additional layers, and saving a map project. The following day included advanced topics, such as performing a spatial join across layers to query attributes, sub-setting the dataset based on spatial selections and intersection, changing the symbology of the data, and performing summary statistics on the attributes. Additionally, visualizing the data over time and producing heat maps of observation locations provided new views on trends within the dataset that would be impossible to detect by simply viewing the records in a table.

Reactions from the data managers and researchers about the QGIS tutorial was positive. Many were excited about the prospect of incorporating QGIS and other spatial analysis techniques into their georeferencing workflows. While fewer participants indicated interest in using QGIS for data cleaning, many expressed interest in using the techniques for data exploration and public communication. Participants who brought their own datasets to the workshop spent time on the last day running similar analyses against their collections in QGIS and were interested in learning how to perform specific GIS operations to spatialize their research questions. For example, several participants were interested in assessing the spatial distribution of specimens in their study area to find a region with a greater significant number occurences. Others were interested in testing predictions about regions where the specimens were not likely to occur (e.g., studying absence) and using other spatial information to account for the trends in distribution.

Desired skills to develop included coupling QGIS with Python scripting to enable researchers and data managers to work with larger datasets. Researchers who also wanted to incorporate QGIS into georeferencing workflows were interested in plugins such as QGIS Gazetteer Plugin that would allow them to cross-reference observations and flag records for removal from further analysis. Best practices for editing CSV files in QGIS were also desired, and a lack of current consensus surrounding when to remove a point from a dataset resulted in inconsistent practices surrounding editing of data using QGIS. For example, once a point has been flagged because it corresponds too closely to a county centroid, should it be removed from the dataset or can its position be rectified? Should the record be noted but not included in analysis? Ideally, the modified and annotated research dataset would be published with original identifiers to enable linking of the data to the original data record and the physical specimen.

Many researchers and collections managers expressed interest in a follow-up tutorial on more advanced techniques in QGIS and were highly motivated to incorporate QGIS into their workflows upon completion of the tutorial. Interest in using QGIS for georeferencing tasks, data exploration, and public communication indicated that planned future tutorials could help users apply more advanced techniques to their existing research or collections datasets.

Participant surveys and evaluations

We gathered data from the participants using pre, post and follow-up surveys. Results from these were used to gauge enthusiasm from participants for the workshop and the direction participants thought future training should progress. The pre-workshop data was collected online through the workshop application using a google forms (Suppl. material 2). The post workshop (Suppl. material 3) and follow-up surveys (Suppl. material 4) were distributed electronically via Qualtrics Survey software licensed to the University of Florida. The application form, post-workshop surveys, and follow-up survey questions and protocols were reviewed and approved by the University of Florida Human Subjects Committee Institutional Review Board (IRB) (University of Florida IRB201601849).

Pre- and post-workshop surveys

Post workshop surveys were given to participants at the end of each day. All participants recorded their knowledge of georeferencing best practices and resources, and their skills for creating and using spatial data were higher after the workshop, with 63-68% of participants indicating a much higher knowledge level after the workshop.

The post-workshop surveys show that more than 50% of participants rated all but two topics covered in the workshop as "most valued." Comments on the survey suggest that the two, less-highly rated topics *Good-Bad Localities* and *Getting Datasets*, were seen as review rather than introduction of new information. Participants also responded that the most useful topics provided during this workshop were *OpenRefine* and *QGIS*, the advanced features of *GEOLocate*, and how to incorperate available APIs into workflows for both basic georeferencing and research.

Table 1 summarizes the participants priorities for future workshops in order to train the next generation in sharing of research-ready biocollection data. Detailed participant responses available for review and study can be found in Suppl. material 4.

Table 1.

Perceived community needs for tools, standards, and skills and training needs.

	High	Somewhat High	Neither High nor Low	Somewhat low
Improving georeferencing efficiency (tools, training)	11	7	1	
Georeferenced data sharing and reintegration	10	7	2	

Quality or fitness-for-use indicators for georeferenced data (standards)	9	7	2	1
Visualizing georeferenced data (training)	9	8	1	1
QGIS for spatial analysis (training)	8	7	4	
R scripting (training)	8	5	5	1
GEOLocate (training)	7	7	3	2
OpenRefine (training)	7	9	3	
Developing georeferencing expertise (training)	7	7	5	
Darwin core georeferencing fields (standards)	4	10	5	
Gazetteer development/ availability (tools. standards)	3	12	2	2

Follow-up survey

Fifty-nine percent of the participants responded to the follow-up survey distributed three months following the workshop (Suppl. material 4,Suppl. material 5). Some of the results are featured in Fig. 4. All participants indicated an increased use of OpenRefine, 77% indicated an increased use of GEOLocate, including one participant who indicated the use of R for GEOLocate, and 54% indicated an increased use of QGIS. About half of the participants had increased their use of documentation relating to georeferencing best practices and the Georeferencing Calculator relative to before the GRU workshop. The majority of participants indicated that into the future, their use of GEOLocate, OpenRefine, QGIS, and georeferencing best practices will continue to increase. QGIS was primarily being used for data visualization, but also for spatial analysis; OpenRefine was being used primarily for data cleaning, error detection, and data reconciliation. Almost all participants that responded provided one-on-one training to institutional colleagues, gave a group presentation, or posted information via social media or blog post. Several participants also reported sharing the knowledge they gained with colleagues outside their institution, and with students, and a publication was inspired by the workshop (Park and Davis 2017). Such sharing of information is critical to helping train the biodiversity workforce (e.g., Hampton et al. 2017, Biodiversity Literacy for Undergraduate Training (BLUE), TDWG Biodiversity Informatics Curriculum Interest Group). All participants that responded desired advanced training on the software and tools (primarily GEOLocate, QGIS and OpenRefine) introduced during the GRU workshop. Participants all indicated an increase in confidence in the use of GEOLocate, QGIS, and OpenRefine, regardless of their initial knowledge of the software prior to training.

Topics for future workshops

A complete list of topics and new tools that participants identified during the workshop as something they wanted to learn more about is available in the supplementary files (Suppl. material 8). In general, the topics of interest trended toward automation, analysis, and data complexity. 32 items were added to the list during the workshop discussion, and out of

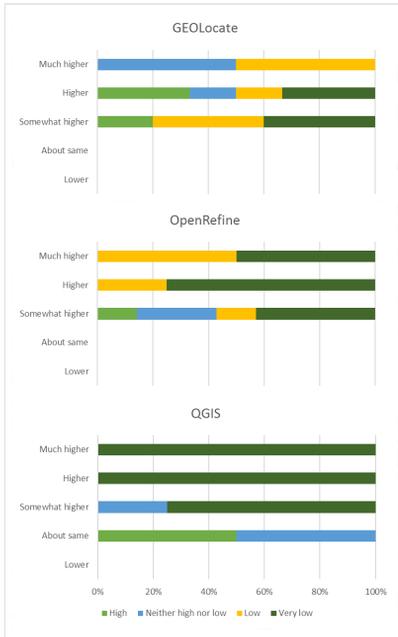those, 11 have been addressed at previous iDigBio workshops, and 19 directly involve software training.

Figure 4. doi

Initial expertise (color of the bar) vs final confidence (y-axis) after the GRU workshop for participants responding to final survey. Example for how to interpret this graphic: the blue color bar at the top indicates that *before the workshop* roughly 50% of respondents said their knowledge of GEOLocate was "neither high nor low" but *after the workshop* these same respondents selected "much higher" for their knowledge of GEOLocate.

Discussion

Digitization of biocollections is somewhat unique in the biological and paleontological data community because the major funding sources that create the data often do not include financial support for research using the resulting digitized data products. This produces large-scale data capture projects that do not participate in research efforts using those data, and these data providers (both collections and aggregators) need input from data users to improve their data products.

One of the expectations between data providers and data users involves the quality of the georeferenced coordinates. Data providers have a mandate to deliver data efficiently, and often data users require data with a quality than may not be cost-efficient or feasible to produce. Other times, locations are deliberately obfuscated (see dwc:informationWithheld) in public databases due to biosecurity, privacy, sensitive species, and other legal reasons.

Original locations are sometimes stored, and the less-precise georeferences are displayed. No currently existing methods that we know of offer easy ways to ensure these withheld data get published when finally appropriate to do so, if ever.

The strongest finding was that 62% of workshop participants increased their use of QGIS since the workshop and 77% expect to use the software increasingly over time, highlighting the importance of training opportunities for career professionals in all sectors of the biodiversity data realm. Expressions of interest by participants for future training on other topics of interest related to georeferencing and the use of biocollections data have been outlined in the iDigBio GRU Wiki requests for the future, and in the post-workshop survey.

Biocollection workflows are changing

Key conversations at the workshop centered around capturing historical versus future biocollections data. All participants, data researchers and data providers noted changes occurring in biocollections data workflows, yet outstanding challenges continue. For all of these reasons and more, there exists a need for ongoing initiatives to address these topics.

- Significant quantities of specimen data collected remain on paper specimen labels, in notebooks, catalogs, and field cards, limiting online open access and resulting in patchy datasets.
- At the same time, new specimens are still being accessioned into biocollections in solely paper-based data formats, thus further contributing to a largely inaccessible backlog of biodiversity specimen data.
- Backlogged legacy data is labor-intensive, expensive, and sometimes difficult to georeference, and often has significant uncertainty.
- Not all georeferences (new or legacy) come with the metadata needed for researchers (or algorithms) to effectively evaluate geospatial data fitness-for-use.
- Many collections undertake the time-consuming process to provide uncertainty data using point-radius or polygons, however, these data seem to be underused in scientific analysis.

Collections and researchers need workflows that reduce unnecessary future collections data management and speed access to data. On this topic, our discussions focused on best practices for creating new georeferences when collecting specimens, georeferencing existing collection locality data, and methods for use and evaluation of geospatial coordinates from historic biocollections records. Optimistically, technology is making it possible for new specimen data to be born-digital (August et al. 2015) and biocollections are beginning to develop policies and subsequent workflows (Nicole Fisher, Digital Collections and Informatics, National Research Collections Australia (NRCA), Personal Communication, August 2018) to prevent the creation of an even bigger collection legacy data backlog (La Salle et al. 2016, Young 2015). There is a push to capture coordinates and other relevant data upfront using mobile devices and other methods, saving time, money and providing valuable, much-needed, more accurate data, faster (Morrison et al. 2017). Born-digital data has the potential to speed access to georeferenced specimen

data, although those using these data will still need to evaluate usefulness and accuracy as born-digital does not necessarily equate to high geospatial data quality.

Importance of access to centralized resources

Finding materials and expertise can be challenging and time-consuming. Development of new tools and workflows continues. Participants appreciated consolidated resources like those found at iDigBio and the Biodiversity Catalogue. Aggregation of this information saves time and effort. Some topics and tools covered stood out from the rest for all participants. These include: template generation for data collection, such as the Biocode Field Information Management System, application development with Open Data Kit (mobile-first data collection platforms), and discussion and development of a data validation process (Suppl. material 7).

Cross-discipline interactions

Hands-on interaction with new software tools and APIs for data refinement, along with in-person interactions enabled by the workshop improved the efficiency of the learning process. Sharing standards of practice and needs across disciplines highlighted the need for changes and transparency in future data collection, evaluation, and data sharing processes.

Exploring and encouraging use of uncertainty

Our workshop experience suggests the need to improve creation, use, and publication of uncertainty data. Whether expressed as a polygon or point-radius, this information appears to be underused and quite time-intensive to create. The biocollections and research communities are encouraged to utilize best practices for creating and sharing coordinate uncertainty information, and the collections community need research examples using these data in order to justify such effort. This may require future training in the research use and value of uncertainty data.

Improved communication about data quality

Data aggregators like GBIF, Atlas of Living Australia, and iDigBio are working through the Biodiversity Science and Standards TDWG/GBIF Biodiversity Data Quality Interest Group to harmonize biocollections data quality (DQ) feedback. Aggregators provide these standardized DQ assertions: 1) to those browsing the data online, 2) to the data providers, and 3) in any downloaded datasets. However, this DQ process, along with the tools, methods, and data standards used during data capture and publishing are not always understood in the research community. For instance, data aggregators currently use taxonomic name assemblages to facilitate searching and indexing of aggregated data. But as a result, the taxonomy may not be what some expect or need (Mesibov 2018,Franz and Sterner 2018). These issues and others, can affect data use and the impressions of the data value (Maldonado et al. 2015,Mesibov 2018). Lacking the understanding of the DQ

process can make it difficult for researchers to evaluate and use the data, and make it tricky for researchers to offer fruitful feedback. This issue is now a priority of the TDWG Biodiversity Data Quality Interest Group which recognizes this barrier to the use of collections data and is actively working to resolve some of these issues.

Proposed actions to speed up biocollection data in research

Some possible actions to advance and speed up the use of biocollection data in research are listed here, and the worldwide biodiversity collections network, industry, and groups like the iDigBio GWG are encouraged to collaborate to support:

- increased outreach and training of necessary skills, standards, and literacy in the biodiversity data community,
- georeferencing of specimens at the time of collection (including uncertainty and source of the coordinates),
- further development and integration of tools such as GEOLocate into collection management systems,
- streamlining batch processing (Guo et al. 2008),
- the development of shared locality services and more gazetteer resources to reduce repeated georeferencing efforts and improve usefulness, and
- the development of techniques to publish and link georeferenced research data sets to the original occurrence records and physical specimens.

Conclusions

The Georeferencing for Research Use workshop was a successful workshop based on the discussion, survey results, and issues reflected in the captured conversations. It created an important platform for biocollections data providers to learn directly from the researchers who hope to use the data they provide and vice versa. As we continue to provide data about specimens, and learn to use the data in research, workshops that provide this kind of cross talk will continue to be important learning platforms that will improve the quality of research and data products. Workshops of this type also offer strategic opportunities to discover future leaders and innovators in our community as the role of collection and data managers evolve to support faster data mobilization and more robustly standardized and complete datasets. We anticipate data collated and summarized in this survey report will contribute valuable information for planning future activities.

Funding program

Hosting institution

University of California, Santa Barbara National Center for Ecological Analysis and Synthesis and University of California, Santa Barbara Cheadle Center for Biodiversity and Ecological Restoration.

Ethics and security

The application form, post-workshop surveys, and follow-up survey questions and protocols were reviewed and approved by the University of Florida Human Subjects Committee Institutional Review Board (IRB) (University of Florida IRB201601849).

Author contributions

Deborah L. Paul and Katja C. Seltmann conceptualized the workshop; Deborah L. Paul, David Bloom, Nelson Rios, Shelley A. James, Sara Lafia, Shari Ellis, Katja C. Seltmann, Una Farrell, Jessica Utrup and Michael Yost developed the workshop teaching materials; Edward Davis, Rob Emery, Gary Motz, Julien Kimmig, Vaughn Shirey, Emily Sandall, Daniel Park, Christopher Tyrrell, R. Sean Thackurdeen, Matthew Collins, Vincent O'Leary, Heather Prestridge, Christopher Evelyn, and Ben Nyberg participated in the workshop and contributed to the manuscript.

References

- Anderson R, Araújo M, Guisan A, Lobo J, Martínez-Meyer E, Peterson AT, Soberón J (2016) Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling. GBIF Secretariat, Copenhagen. URL: http://www.gbif.org/system/files_force/gbif_resource/resource-82612/Final_report_of_the_task_group_on_GBIF_data_fitness_for_use_in_distribution_modelling_ver1.1.pdf?download=1
- Ariño A, Chavan V, Otegui J (2016) Best Practice Guide for Data Gap Analysis for Biodiversity Stakeholders. https://www.gbif.org/document/82566/best-practice-guide-for-data-gap-analysis-for-biodiversity-stakeholders. Accessed on: 2018-5-18.
- Arnaud E, Castañeda-Álvarez NP, Cossi JG, Endresen D, Jahanshiri E, Vigouroux Y (2016) Final report of the Task Group on GBIF data fitness for use in agrobiodiversity. GBIF Secretariat, Copenhagen. URL: https://www.gbif.org/document/82283/report-of-the-task-group-on-gbif-data-fitness-for-use-in-agrobiodiversity

- August T, Harvey M, Lightfoot P, Kilbey D, Papadopoulos T, Jepson P (2015) Emerging technologies for biological recording. Biological Journal of the Linnean Society 115 (3): 731-749. https://doi.org/10.1111/bij.12534
- Bloom TDS, Flower A, DeChaine EG (2017) Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis. Ecology and Evolution 8: 765-777. https://doi.org/10.1002/ece3.3516
- Brummitt N, Bachman S, Griffiths-Lee J, Lutz M, Moat J, Farjon A, Donaldson J, Hilton-Taylor C, Meagher T, Albuquerque S, Aletrari E, Andrews AK, Atchison G, Baloch E, Barlozzini B, Brunazzi A, Carretero J, Celesti M, Chadburn H, Cianfoni E, Cockel C, Coldwell V, Concetti B, Contu S, Crook V, Dyson P, Gardiner L, Ghanim N, Greene H, Groom A, Harker R, Hopkins D, Khela S, Lakeman-Fraser P, Lindon H, Lockwood H, Loftus C, Lombrici D, Lopez-Poveda L, Lyon J, Malcolm-Tompkins P, McGregor K, Moreno L, Murray L, Nazar K, Power E, Tuijtelaars MQ, Salter R, Segrott R, Thacker H, Thomas L, Tingvoll S, Watkinson G, Wojtaszekova K, Nic Lughadha E (2015) Green plants in the red: A baseline global assessment for the IUCN sampled Red List Index for plants. PLOS One 10 (8): e0135152. https://doi.org/10.1371/journal.pone.0135152
- Chapman A, Wieczorek J (Eds) (2006) Guide to best practices for georeferencing. Global Biodiversity Information Facility, Copenhagen, Denmark, 28 pp. URL: http://www.gbif.org/orc/?doc_id=1288
- Deck J, Ewing R (2016) Biocode Field Information Management System. http://www.biscicol.org/. Accessed on: 2017-3-03.
- Department of Primary Industries and Regional Development: Agriculture and Food (2018) What is Biosecurity? https://www.agric.wa.gov.au/biosecurity-quarantine/biosecurity. Accessed on: 2018-9-23.
- Franz NM, Sterner BW (2018) To increase trust, change the social design behind aggregated biodiversity data. Database: The Journal of Biological Databases and Curation 2018 (1): bax100. https://doi.org/10.1093/database/bax100
- Guo Q, Liu Y, Wieczorek J (2008) Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. International Journal of Geographical Information Science 22 (10): 1067-1090. https://doi.org/10.1080/13658810701851420
- Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH (2013) Big data and the future of ecology. Frontiers in Ecology and the Environment 11 (3): 156-162. https://doi.org/10.1890/120103
- Hampton SE, Jones MB, Wasser LA, Schildhauer MP, Supp SR, Brun J, Hernandez RR, Boettiger C, Collins SL, Gross LJ, Fernández DS, Budden A, White EP, Teal TK, Labou SG, Aukema JE (2017) Skills and knowledge for data-intensive environmental research. BioScience 67 (6): 546-557. https://doi.org/10.1093/biosci/bix025
- Hanken J (2013) Biodiversity online: Toward a network integrated biocollections alliance. BioScience 63 (10): 789-790. https://doi.org/10.1525/bio.2013.63.10.4
- iDigBio (2016) Why Digitize? https://vimeo.com/184552432. Accessed on: 2016-1-01.
- iDigBio (2018) iDigBio Portal Search. www.idigbio.org/portal/. Accessed on: 2018-9-23.
- Kearney MR, Isaac AP, Porter WP (2014) microclim: Global estimates of hourly microclimate based on long-term monthly climate averages. Scientific Data 1 https://doi.org/10.1038/sdata.2014.6

- Krishtalka L, Dalcin E, Ellis S, Ganglo JC, Hosoya T, Nakae M, Owens I, Paul D, Pignal M, Thiers B (2016) Accelerating the discovery of biocollections data. http://www.gbif.org/resource/83022. Accessed on: 2018-4-04.
- La Salle J, Williams KJ, Moritz C (2016) Biodiversity analysis in the digital era. Philosophical Transactions of the Royal Society B: Biological Sciences 371 (1702): 20150337. https://doi.org/10.1098/rstb.2015.0337
- Lieberman BS, Kimmig J (2018) Museums, Paleontology, and a Biodiversity Science Based Approach. In: Rosenberg GD, Clary RM (Eds) Museums at the Forefront of the History and Philosophy of Geology: History Made, History in the Making. Vol. 535. Geological Society of America https://doi.org/10.1130/2018.2535(22)
- Lister A (2011) Natural history collections as sources of long-term datasets. Trends in Ecology & Evolution 26 (4): 153-154. https://doi.org/10.1016/j.tree.2010.12.009
- Maldonado C, Molina C, Zizka A, Persson C, Taylor C, Albán J, Chilquillo E, Rønsted N, Antonelli A (2015) Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Global Ecology and Biogeography 24 (8): 973-984. https://doi.org/10.1111/geb.12326
- McCulloch ES (2013) Harnessing the power of Big Data in biological research. BioScience 63 (9): 715-716. https://doi.org/10.1525/bio.2013.63.9.4
- McGeoch M, Groom Q, Pagad S, Petrosyan VG, Ruiz GM, Wilson J (2016) Task group on data fitness for use in research into invasive alien species. GBIF Secretariat, Copenhagen. URL: https://www.gbif-uat.org/document/82958/data-fitness-for-use-in-research-on-alien-and-invasive-species
- Mesibov R (2018) An audit of some processing effects in aggregated occurrence records. ZooKeys 751: 129-146. https://doi.org/10.3897/zookeys.751.24791
- Morrison S, Sillett TS, Funk WC, Ghalambor C, Rick T (2017) Equipping the 22nd-century historical ecologist. Trends in Ecology & Evolution 32 (8): 578-588. https://doi.org/10.1016/j.tree.2017.05.006
- Nature Publishing Group (2008) A place for everything. Nature 453 (7191): 2-2. https://doi.org/10.1038/453002a
- Online Computer Library Center (OCLC) (2018) OCLC Research: Demystifying Born Digital. https://www.oclc.org/research/themes/research-collections/borndigital.html. Accessed on: 2018-3-19.
- Otero-Ferrer F, González J, Freitas M, Araújo R, Azevedo JN, Holt W, Tuya F, Haroun R (2017) When natural history collections reveal secrets on data deficient threatened species: Atlantic seahorses as a case study. Biodiversity and Conservation 26 (12): 2791-2802. https://doi.org/10.1007/s10531-017-1385-x
- Park D, Davis C (2017) Implications and alternatives of assigning climate data to geographical centroids. Journal of Biogeography 44 (10): 2188-2198. https://doi.org/10.1111/jbi.13029
- Rios N (2018) GEOLocate - Software for Georeferencing Natural History Data. [Web application software]. http://www.geo-locate.org. Accessed on: 2018-4-09.
- Seltmann KC, Cobb NS, Gall LF, Bartlett CR, Basham MA, Betancourt I, Bills C, Brandt B, Brown RL, Bundy C, Caterino MS, Chapman C, Cognato A, Colby J, Cook SP, Daly KM, Dyer LA, Franz NM, Gelhaus JK, Grinter CC, Harp CE, Hawkins RL, Heydon SL, Hill GM, Huber S, Johnson N, Kawahara AY, Kimsey LS, Kondratieff BC, Krell F-T, Leblanc L, Lee S, Marshall CJ, McCabe LM, McHugh JV, Menard KL, Opler PA, Palffy-Muhoray N, Pardikes N, Peterson MA, Pierce NE, Poremski A, Sikes DS, Weintraub JD,

Wikle D, Zaspel JM, Zolnerowich G (2017) LepNet: The Lepidoptera of North America Network. Zootaxa 4247 (1): 73-77. https://doi.org/10.11646/zootaxa.4247.1.10

- Shaffer HB, Fisher RN, Davidson C (1998) The role of natural history collections in documenting species declines. Trends in Ecology & Evolution 13 (1): 27-30. https://doi.org/10.1016/s0169-5347(97)01177-4

- Shorthouse D (2010) SimpleMappr, an online tool to produce publication-quality point maps. Accessed December 4: 2018. URL: http://www.simplemappr.net.

- Sinervo B, Mendez-de-la-Cruz F, Miles DB, Heulin B, Bastiaans E, Cruz MV, Lara-Resendiz R, Martinez-Mendez N, Calderon-Espinosa ML, Meza-Lazaro RN, Gadsden H, Avila LJ, Morando M, la Riva IJD, Sepulveda PV, Rocha CFD, Ibarguengoytia N, Puntriano CA, Massot M, Lepetz V, Oksanen TA, Chapple DG, Bauer AM, Branch WR, Clobert J, Sites JW (2010) Erosion of Lizard Diversity by Climate Change and Altered Thermal Niches. Science 328 (5980): 894-899. https://doi.org/10.1126/science.1184695

- Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, Pollock D, Dorsett K (2015) Changes in data Sharing and data reuse practices and perceptions among scientists worldwide. PLOS One 10 (8): e0134826. https://doi.org/10.1371/journal.pone.0134826

- Young A (2015) Data from drawers: Securing, mobilising and interrogating National Research Collections data. In: National Collections Research Australia (NCRA) Digitization of Biological Collections: A Global Focus. CSIRO (Commonwealth Scientific and Industrial Research Organisation), Blackmountain Labs, Canberra, April 2015.

- Zimmermann AS (2008) New knowledge from old data: the role of standards in the sharing and reuse of ecological data. Science, Technology, and Human Values 33 (5): 631-652. https://doi.org/10.1177/0162243907306704

Supplementary materials

Suppl. material 1: GRU Workshop Carabidae Beetle Dataset `doi`

Authors: Seltmann K, Paul D, et al.
Data type: Darwin Core File
Brief description: Darwin Core Archive file downloaded from the iDigBio portal for use in the Georeferencing for Research Use workshop. Total 25,429 records, accessed on 2016-08-29. Collections contributing to the record set are listed in the archive records.citation.txt file. Dataset GUID: a69d1541-4726-465d-84ad-50c7ed556eee
Filename: a69d1541-4726-465d-84ad-50c7ed556eee.zip - Download file (5.67 MB)

Suppl. material 2: Georeferencing for Research Use - Call for Participation Form and Pre-Workshop Survey (Blank form) `doi`

Authors: Deborah L. Paul, David Bloom, Nelson Rios, Shelley A. James, Sara Lafia, Shari Ellis, Katja C. Seltmann, Una Farrell, Jessica Utrup and Michael Yost
Data type: pdf
Brief description: This document shows just the questions we asked the applicants who applied to participate in this Georeferencing for Research Use workshop. We used a Google Form to deliver these questions and collect responses. It is both an application and serves as our pre-workshop survey.

Filename: GeoreferencingForResearchUseWorkshpApplication.pdf - <u>Download file</u> (39.64 kb)

Suppl. material 3: Georeferencing for Research Use Workshop Informed Consent and Post-Workshop Surveys for Days 1 - 4 (Blank form) `doi`

Authors: Ellis S, Paul D, James S, Seltmann K et al.
Data type: PDF
Brief description: The informed consent request and workshop survey questions given to participants after the workshop each day for 4 consecutive days.
Filename: Georeferencing_for_Research_Use_Post-Workshop_Survey__Days1-4_and_Consent.pdf - <u>Download file</u> (178.88 kb)

Suppl. material 4: Georeferencing for Research Use - Follow-Up Survey Summarized Data `doi`

Authors: Ellis S, Paul D, James S, Seltmann K et al. and workshop participants
Data type: data
Brief description: Three months after the workshop, participants were surveyed to assess what workshop-related knowledge and materials were being used and disseminated to others. This document summarized data collected in this particular survey.
Filename: Georef for Research Use Follow Up.docx - <u>Download file</u> (225.30 kb)

Suppl. material 5: Georeferencing for Research Use - Follow-Up Survey (Blank) `doi`

Authors: Ellis S, Paul D, James S, Seltmann K et al.
Data type: PDF
Brief description: Questions we asked in the Georeferencing for Research Follow Up Survey done 3 months after the workshop.
Filename: Georeferencing_for_Research_Use_Follow-Up_Survey_Rev.pdf - <u>Download file</u> (67.08 kb)

Suppl. material 6: Georeferencing for Research Use - Summary of Participants Desired Learning Outcomes for Workshop `doi`

Authors: Workshop participants
Data type: .doc
Brief description: Summary of topics to be covered in an ideal workshop as identified by workshop applicants in the workshop call for participation. We incorporated as many as possible that also fit our scope.
Filename: Georef workshop desired topics.docx - <u>Download file</u> (8.38 kb)

Suppl. material 7: Georeferencing for Research Use - Participant-generated list of data quality checks to evaluate data suitability `doi`

Authors: Workshop participants
Data type: annotated list
Brief description: This document contains an annotated set of data quality checks that participants report they use when evaluating and cleaning datasets. These items outline how participants are judging if the data suits their purpose.
Filename: Conversation III-Frameworks and Data Quality.docx - <u>Download file</u> (30.67 kb)

Suppl. material 8: Future Workshop Topics - Participant-generated Wish List doi

    Authors: Workshop Participants
    Data type: .doc
    Brief description: Summary of desired future workshop topics that were listed by participants on the last day of the workshop.
    Filename: FutureWorkshops.docx - [Download file](#) (13.33 kb)