Data management aspects of public engagement with biodiversity documentation

Alvaro Ortiz-Troncoso
Museum für Naturkunde, Leibniz Institute for Research on Evolution and Biodiversity at the Humboldt University Berlin, Invalidenstraße 43, 10115 Berlin, Germany

E-mail: alvaro.ortiztroncoso@mfn-berlin.de

Abstract

Technological developments open up new opportunities for collaboration between biodiversity researchers and the general public. Three exemplary use cases were examined: digitizing museum specimens, text-mining archived expedition journals and handling environmental monitoring data. Data management principles were applied to refine and map the ensuing requirements to specific deliverables: data policy, standards and procedures; workflows, integration architectures and data products; data quality awareness and improvement methods. Implications for data governance and quality control are discussed.

Keywords: citizen science, crowdsourcing, data governance, data integration, data policy, data quality, digitization, environmental monitoring, primary biodiversity data, text-mining.

Introduction

Primary biodiversity data

Primary biodiversity data records the presence or absence of a certain taxon (of plant or animal etc.) in a particular place and time; this data has many applications: evolutionary research questions, ecological management issues (climate change, invasive species), epidemiology or natural disaster management (Soberón & Peterson, 2004; Lukyanenko, Parsons & Wiersma, 2011).

Primary biodiversity data is obtained from:

- Natural history collections (i.e. vouchered with a specimen; Ellwood et al., 2015)
- Historical observation records (i.e. archived expedition journals; Thomer et al., 2012)
- On-site environmental monitoring (Sullivan et al., 2014).

Citizen science

Public engagement in science has a long tradition. The relatively recent term 'citizen science'

(Irwin, 1995; cited in Catlin-Groves, 2012) reflects technological developments enabling new modes of public engagement on a larger scale than was possible previously (Rubio Iglesias, 2014). A pragmatic approach to the concept is to consider the common principles any citizen science project should adhere to (Robinson, 2014):

- Scientific goals should be pursued.
- While pursuing these goals, volunteers are actors, not research subjects.
- Volunteers should potentially participate in setting hypotheses, designing processes, collecting data, analysis and publication.
- Data should be shared; results published in open access journals.
- The volunteers' contributions should be acknowledged in research publications.
- Scientists should strive to increase the volunteers' scientific literacy.
- Projects should be steered by volunteers and scientists at eye level.
- Participation should be accessible to different groups of volunteers.
- Participants should strive to bridge the gap between science and society.
- Results should be evaluated for their scientific significance, the quality of the data they produce and their social impact.

Data management

All of the principles above could be expected to entail a controlled use of data assets at some level, yet three aspects of citizen science explicitly call for a managed data environment:

Collecting data

How can science institutions leverage the effort of volunteers, which data policy should be adhered to?

Sharing data

Which integration architecture, standards and information products are necessary for distributing this data to scientists, decision makers and the general public?

Evaluating data quality

Which quality control measures and training should be implemented to fully realize the benefits of citizen science and increase its relevance for research?

The Data Management Association (DAMA) compiled the DAMA Data Management Body of Knowledge (DMBOK) to serve as a comprehensive guide to data management activities (Mosley et al., 2009). Using this framework, citizen science requirements can be further refined and mapped to specific deliverables and responsibilities (Table 1).

***Table 1**: Data management activities required by citizen science refined and mapped to specific deliverables and responsibilities using the DAMA DMBOK framework.*

Requirements	Corresponding DAMA DMBOK activities	Deliverables	Responsible roles
Collecting data			
Develop a data policy	Develop, review and approve data policies, standards, and procedures	• Data policies • Data standards • Data management procedures	Data governance council
Sharing data			
Build an appropriate data integration architecture	• Analyse and align with other business models • Define and maintain the metadata architecture • Define and maintain the data integration architecture	• Information value chain analysis • Data integration architecture • Metadata integration architecture	Data architect
Make information accessible to different audiences	Design, build and test information products	Models, reports	Software developer
Improving data quality			
Train volunteers	Develop and promote data quality awareness	Data quality training	Data steward
Implement quality control measures	Define data quality business rules	Data quality business rules	Data quality analyst

Materials & methods

Three exemplary use cases were selected: digitization of museum specimens, text-mining archived expedition journals and handling environmental monitoring data. For each use case, a recent (as of 2015) peer-reviewed paper describing data management aspects was analysed, using the DAMA DMBOK activities as a guide (Table 1).

Digitizing museum specimens

In: "Accelerating the digitization of biodiversity research specimens through online public participation", Ellwood et al. (2015) point out that, as digitization is prohibitively expensive, only a small fraction of the specimens available in collections have been digitized. Several digitization tasks are described and their implications for the management of volunteer-contributed data are examined.

Text-mining archived notebooks

In "From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks", Thomer et al. (2012) examine the workflows necessary for converting unstructured text into structured data through a collaboration with the public on an open platform. Data access policies, interoperability issues and quality control are discussed.

Handling environmental monitoring data

In "The eBird enterprise: An integrated approach to development and application of citizen science", Sullivan et al. (2014) describe the workings of eBird, Cornell University's citizen science platform. With 150 000 volunteers contributing species occurrence observations, this platform is setting the standards among citizen science environmental monitoring programs in terms of data access policies, data products and quality assurance.

Results

Data policy

Using "complete open access" practices, such as those championed by Wikipedia, using open source software and promotion through social media have proven workable methods for increasing the outreach of the projects (Thomer et al, 2012). However, beyond the standardization of tools and methods, a need exists to provide a framework for streamlining negotiations between data custodians (e.g. collection curators) and project managers (Ellwood et al., 2015).

Value-chain analysis

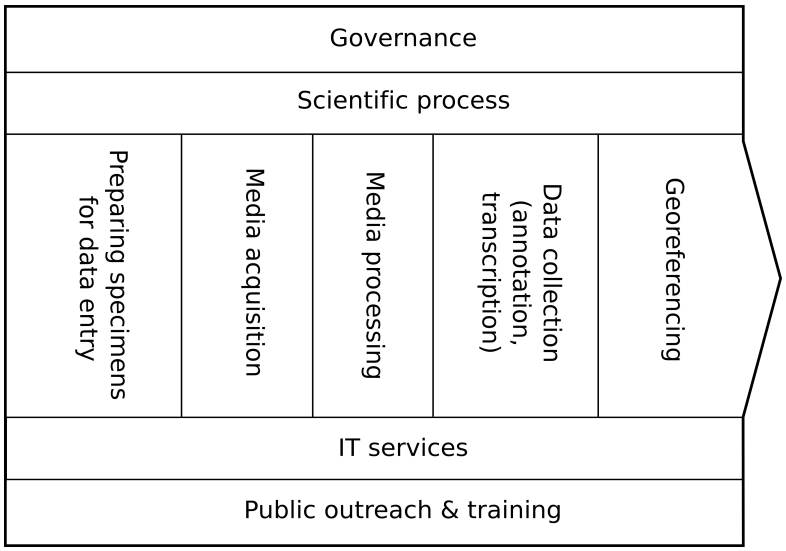

***Figure 1**: Capturing primary biodiversity data is embedded in a scientific process, supported by governance structures, IT services and in a citizen science context, public outreach & training.*

A workflow for capturing primary biodiversity data follows a basic blueprint (Fig. 1): Preparing specimens for data entry, media acquisition, media processing, data collection and geo-referencing (Ellwood et al., 2015). For field monitoring, the first step is skipped, as this data is not vouchered with a specimen.

Metadata integration

The 'data collection' and 'geo-referencing' steps in the digitalization workflow (Ellwood et al., 2015) necessitate the integration of specific metadata standards and taxonomies:

Transcription

Transcription refers to the conversion of unstructured text into structured data. Transcription can be supported by generic resource descriptor standards such as the Dublin Core (http://dublincore.org), while the Text Encoding Initiative standard (http://www.tei-c.org) can be applied to the mark-up of scholarly texts (Ellwood et al., 2015; Thomer et al., 2012).

Annotation

Specimen annotation can be backed by the Darwin Core metadata schema (http://rs.tdwg.org/dwc) for describing biodiversity data (Ellwood et al., 2015) or the Access to Biological Collections Data schema (http://www.tdwg.org/activities/abcd/). On the other hand, projects may choose to develop their own standard as one of their deliverables (e.g. the 'user-friendly' taxonomy maintained by the eBird platform; Sullivan et al., 2014). Alternatively, records can be annotated using templates for machine-readable metadata, as maintained by Wikimedia (Thomer et al., 2012).

Geo-referencing

Ellwood et al. (2015) identify the Open Geospatial Consortium (http://opengeospatial.org/) and the implementation supplied by the Environmental Systems Research Institute (http://esri.com) as major sources of geo-referencing standards.

Integration of data sources

In order to realize its outreach potential, a citizen science platform should accommodate data flows originating in portals serving different user groups or language communities; furthermore, the data collection protocol should be modifiable to serve different research objectives (Fig. 2; Sullivan et al., 2014).

Information products

Ellwood et al. (2015) point out that many current digitization projects make their data accessible to the data custodians but not to the volunteers, and that this situation hinders a truly collaborative creation and management of information. Nevertheless, three types of information products can be identified:

Primary data

Observational data can be aggregated through data clearinghouses (e.g. the Global Biodiversity Information Facility; http://gbif.org); additionally, users may download their own data (Sullivan et al., 2014). Scans can be made available for download in PDF, OCR-augmented PDF or DjVu multipage image file (Thomer et al., 2012).

Annotated data

A dataset containing primary data (taxon, place, time) and metadata describing the observation event (protocol used, observer, equipment) can be made available for download (Sullivan et al., 2014; Thomer et al., 2012).

Predictive models

Spatiotemporal exploratory models can be provided to organizations seeking to estimate the environmental impact of conservation policy (Sullivan et al., 2014).

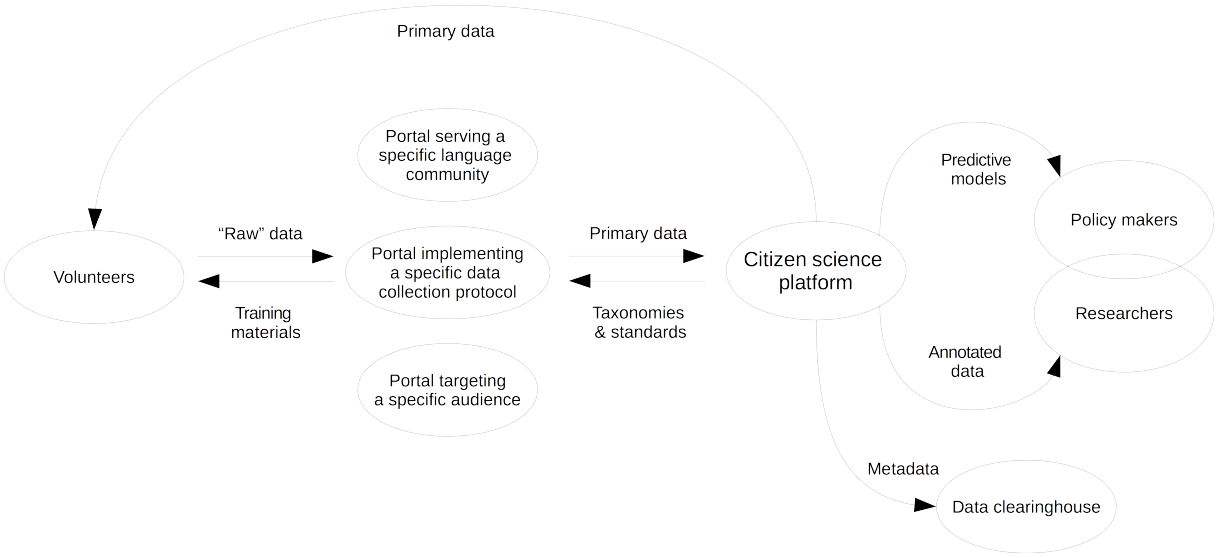

***Figure 2**: Data flows in and out of a citizen science platform.*

Training

Training should develop and promote data quality awareness by combining the scientific and the public outreach processes. Training can be structured along the tasks required by the digitalization workflow outlined by Ellwood et al. (2015):

Transcription

Scientific jargon, label and date formats can be clarified, as well as the identification and resolution of inconsistencies (Ellwood et al., 2015).

Annotating

Training should focus on the identification of specific taxa and the correct use of taxonomical terms; volunteers should familiarize themselves with possible variations

within a taxon as well as artefacts induced by the imaging process (Ellwood et al., 2015).

Geo-referencing

Training should emphasize skills such as understanding geographic jargon, projections and descriptions, using maps as well as dealing with inconsistencies (Ellwood et al., 2015).

Suitable vehicles for training materials are: online forums, tutorials and videos (Ellwood et al., 2015). Training can also build upon existing resources provided by the platform itself: by delivering its content through Wikisource, the notebook transcription project described by Thomer et al. (2012) piggybacks on the community-driven forums of Wikipedia. Training can also take place within formal school curricula developed in cooperation between citizen science portals and teachers (Sullivan et al., 2014).

Data quality

Ellwood et al. (2015) note that quality issues are the main source for concern when using data contributed by volunteers in research. Dealing with deviations from expected quality standards is a three step process (Mosley et al., 2009):

1. Identifying faulty data values

An automated plausibility check can be performed to identify records which do not meet reasonable expectations (Sullivan et al., 2014). What constitutes a reasonable expectation can be inferred from formal data quality rules, which are available for geo-referencing and transcription tasks (Ellwood et al., 2015). The threshold beyond which a record is considered unreasonable can be fine-tuned by applying statistical methods (Sullivan et al., 2014). Additionally, faulty values can be identified by proof readers (Thomer et al., 2012)

2. Notifying the person in charge

Once a record has been flagged as dubious, a data steward can be notified; Sullivan et al. (2014) recommend assigning to this function a person with expert knowledge of the region where the record originated.

3. Establishing a process to correct the fault

If the number of volunteers allows it, the data steward can provide feedback to improve the volunteer's skills (Sullivan et al., 2014). If the number of volunteers calls for a collective evaluation of the data, known problems (e.g. correcting taxonomical and geographical bias) can be handled by applying statistical techniques (Ellwood et al., 2015). Inconsistencies in the values assigned to attributes (e.g. taxonomical or geographical names) can be reconciled by computing the best fit against reference records (Thomer et al., 2012).

Discussion

Limitations on an open data policy

Many publishers are uncomfortable with the idea of an open data policy, notwithstanding that opening-up data generally fosters the dissemination of knowledge (Hagedorn et al., 2011) and in some cases, researchers have claimed exclusive access to citizen science data prior to publication (Hampton et al., 2014).

Enforcing an open data policy also has some practical drawbacks: Data distribution should comply with privacy and property regulations, and sensitive data (e.g. the location of endangered species) should be protected (Crall et al., 2010).

These obstacles underline the need for developing a framework for standardizing data handling procedures and constraints in the citizen science domain (Ellwood et al., 2015). Such a framework could use the categorization of governance structures for citizen science data proposed by Conrad & Hilchey (2011) as a starting point:

- If protection of sensitive data has the greatest priority, or right-of-first-publication issues exist, implement consultative / functional governance (i.e. initiated by a central authority, which can be a government or a research institution).
- If protection of privacy and private property is the major issue, implement collaborative governance (i.e. share responsibility among representatives of different interest groups).
- If maximizing outreach is the main goal, implement transformative governance (i.e. a community-based form of data governance).

Achieving trust

As noted by Ellwood et al. (2015), quality issues are the main source for concern when using data contributed by volunteers, and citizen science data is known to suffer from geospatial and taxonomical biases (Sullivan et al., 2014). However, Catlin-Groves (2012) points out that, given adequate tasks and guidance, volunteers gather data of comparable quality than professionals.

Lukyanenko, Parsons & Wiersma (2011) have compiled a list of options for increasing data quality in a citizen science context:

Training

Training is the common method for increasing quality, it is however expensive and not always practicable for large projects.

Verification

Verification by professional experts is contrary to the spirit of citizen science, according to Lukyanenko, Parsons & Wiersma (2011).

Social networking

Relying on a web of trust created by a social network can also be a practicable solution for increasing quality (e.g. the notebook transcription project described by Thomer et al. relies on the Wikisource community for support). However this solution is only applicable to projects which are modelled after a social network principle.

Attribute-based data collection

Lukyanenko, Parsons & Wiersma (2011) propose that volunteers should not provide a direct classification of the taxa observed, but describe them. This method purports to be more open to non-experts as well as less prone to classification errors.

Conclusions

Value-chains and workflows for acquiring and processing primary biodiversity data are applicable to a citizen science context, particularly annotation, transcription and geo-referencing tasks. Standard data formats and supporting taxonomies are available. Several data and metadata integration architectures are in operation. More work is needed to standardize data policies and data governance structures, with the long-term goal of facilitating negotiations between the principal stakeholders: data custodians, researchers, policy makers and volunteers. Quality control should strive to widen the scope of fault correction and training methods.

Acknowledgements

I wish to thank Claudia Göbel and Florian Wetzel for letting me attend the European Citizen Science Association general assembly and the European Biodiversity Observation Network stakeholder roundtable at Museum für Naturkunde Berlin, 2014.

References

Catlin-Groves CL. 2012. The citizen science landscape: from volunteers to citizen sensors and beyond. *International Journal of Zoology*, Volume 2012.

Conrad CC & Hilchey KG. 2011. A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental monitoring and assessment*, 176:273-291.

Crall AW, Newman GJ, Jarnevich CS, Stohlgren TJ, Waller DM & Graham J. 2010. Improving and integrating data on invasive species collected by citizen scientists. *Biological Invasions*, 12(10):3419-3428.

Ellwood ER, Dunckel BA, Flemons P, Guralnick R, Nelson G, Newman G, Paul D, Riccardi G, Rios N, Seltmann KG & Mast AR. 2015. Accelerating the Digitization of Biodiversity Research Specimens through Online Public Participation. *BioScience*, biv005.

Hagedorn G, Mietchen D, Morris RA, Agosti D, Penev L, Berendsohn WG & Hobern D.

2011. Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. *ZooKeys*, 150:127-149.

Hampton SE, Anderson S, Bagby SC, Gries C, Han X, Hart E, Jones MB, Lenhardt WC, MacDonald A, Michener W, Mudge JF, Pourmokhtarian A, Schildhauer M, Woo KH & Zimmerman N. 2014. The Tao of Open Science for Ecology. *PeerJ PrePrints*, 2:e549v1 https://dx.doi.org/10.7287/peerj.preprints.549v1 (accessed 23 March 2015)

Irwin A. 1995. *Citizen science: A Study of People, Expertise and Sustainable Development*, London: Routledge.

Lukyanenko R, Parsons J & Wiersma Y. 2011. Citizen science 2.0: Data management principles to harness the power of the crowd. *Service-Oriented Perspectives in Design Science Research*. Berlin: Springer Verlag, 465-473.

Mosley M, Brackett MH, Earley S & Henderson D. 2009. *DAMA guide to the data management body of knowledge, 1st edn*. Bradley Beach: Technics Publications.

Robinson L. 2014. Principles and standards in citizen science: sharing best practice and building capacity. Available at http://ecsa.biodiv.naturkundemuseum-berlin.de/sites/ecsa.biodiv.naturkundemuseum-berlin.de/files/ECSA-GA-2014-04-14-Robinson.pdf (accessed 9 March 2015)

Rubio Iglesias JM. 2014. Citizen science. Improving science-society-policy bridge. A perspective. Available at http://adm.eubon.eu/getatt.php?filename=oo_11927.pdf (accessed 9 March 2015).

Soberón J & Peterson T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 359(1444):689-698.

Sullivan BL, Aycrigg JL, Barry JH, Bonney RE, Bruns N, Cooper CB, Damoulas T, Dhondt A, Dietterich T, Farnsworth A, Fink D, Fitzpatrick JW, Fredericks T, Gerbracht J, Gomes C, Hochachka WM, Iliff MJ, Lagoze C, La Sorte FA, Merrifield M, Morris W, Phillips TB, Reynolds M, Rodewald AD, Rosenberg KV, Trautmann NM, Wiggins A, Winkler DW, Weng-Keen Wong, Wood CL, Jun Yu & Kelling S. 2014. The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation*, 169: 31-40.

Thomer A, Vaidya G, Guralnick R, Bloom D & Russell, L. 2012. From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks. *ZooKeys*, 209:235-253.