Project Number: CA17122

Project Acronym: Alien-CSI

Project title: Increasing Understanding of Alien Species Through Citizen Science

Contributors

Quentin Groom, Tim Adriaens, Ana Cristina Cardoso, Franz Essl, Kelly Martinou, Toril Loennechen Moen, Jan Pergl, Michael Pocock, Lien Reyserhove, Sven Schade, Elena Tricarico & Helen Roy

DATA MANAGEMENT PLAN

Based upon the Template for the Data Management Plan provided by the European Commission in the Participants Portal of H2020

1. Data Summary

Alien-CSI is a research network (COST Action https://www.cost.eu/actions/CA17122/) with the overarching aim of increasing public awareness and levels of participation on issues related to invasive alien species and citizen science (Roy et al. 2018). The Action links together researchers, institutions and projects to share best practice, knowledge and policies, but it also aims to discover new ways to coordinate our work across Europe and globally. Alien-CSI is not primarily a data generating project, however, throughout the Action we will be gathering data on many aspects of citizen science and biology to inform our discussions.

With regards to data on invasive alien species a number of recommendations were already put forward (e.g. Groom et al. 2015, 2017); a number of key recommendations are recapped in Table 1.

---

1. Create and implement data management plans to define the alien species data life cycle, good data quality and metadata, standardisation, data sharing options, and long-term data preservation.
2. Increase interoperability and sustainability of existing and new alien species information sources by exposing the data they contain through standard exchange formats.
3. Describe alien species data through metadata, so users can understand its scope and limitations, and use metadata standards (EML, INSPIRE) to facilitate metadata exchange.
4. Format data using existing standards (Darwin Core, GISIN) and engage in their development through TDWG.
5. Adopt controlled vocabularies to further increase interoperability of data and engage with TDWG to make these compatible with existing standards.
6. Increase data availability by making alien species data openly accessible as soon as possible after collection.
7. Ensure long-term preservation of alien species data by archiving these in existing data repositories (GBIF, Zenodo).

Table 1. Recommendations for improving the usefulness of alien species data (Groom et al. 2017).

This data management plan is intended to give further guidance to participants of the Action, and potentially all other users, on how to gather, curate, store and publish data related to our work. It will also give some guidance on how to share data openly so that it can be used beyond the project and persist for use in the future. For wider recommendations regarding data platforms and mobile apps, we refer the interested reader to Sturm et al. (2018) and Luna et al. (2018).

Key questions everyone should think about when handling data resulting from or used within the Action are:

➔ How are my data described i.e. what metadata are required and what metadata standards will be used?
➔ How can I ensure my data are preserved and formatted for future use.
➔ How can I maximize (re)use of the data, what licences will data be shared under and what restrictions (if any) will be put on data reuse?

Types of data

Biodiversity data through BioBlitz

BioBlitzes have become a commonly used method for engaging citizen scientists. They are collaborative events to discover and record as many of the living species within a designated area, over a defined period of time (Robinson et al. 2013). They can have multiple goals, but public engagement, teaching and data gathering are all important outcomes (Roger & Klistorner 2016). Alien-CSI may conduct BioBlitzes to generate data on the BioBlitz methods, for engagement with the Action and for members of the Action to gain experience in science communication and engagement through participation. In the process, data will be gathered on biodiversity and it is intended that these data will be disseminated to make them useful in other research. These data are particularly occurrence records, but can include other forms of observation and measurement.

Data from questionnaires

Alien-CSI will generate data through online and e-mail questionnaires. These will include information from participants of the Action and may also be directed externally to a wider group of stakeholders such as project managers, data managers or participants in citizen science projects. They may be used to gather facts or perceptions about citizen science projects, the management of facts on alien species and facts (metadata) regarding characteristics and outputs of citizen science projects.

Data gathered from online research

Data may also be gathered from public domain websites or printed publications, such as other citizen science initiatives on alien species in Europe. These data will be collated and perhaps annotated with additional data. Other sorts of data in this category might include methodologies used in citizen science or local policies on invasive species.

We might also collect bibliographies related to our aims. These might be used for meta-analysis or more specifically in the process of writing papers.

Data Reuse

Where appropriate we will reuse data collected by other projects assuming they have been published openly or we have obtained explicit permission to use the data. Examples of such data might be observation data from the Global Biodiversity Information Facility (GBIF), but also data on monitoring schemes and projects, such as those in Chandler et al. (2017).

Specifically with regards to citizen science project metadata and in light of activities in several Alien-CSI working groups an important dataset is the Citizen Science Project Inventory (https://ec.europa.eu/eusurvey/runner/CSProjectInventory). This project aims to collect information about citizen science projects to be included in the Joint Research Centre (JRC) project inventory of citizen science activities for environment policies . It is available online in the JRC data catalogue (https://data.jrc.ec.europa.eu/) (Bio Innovation Service 2018). This inventory already contains

metadata for citizen science projects with controlled vocabularies, including typologies of level of engagement and engagement methods in citizen science.

2. FAIR data

We will follow the FAIR data principles. FAIR is described as a set of guiding principles to make data findable, accessible, interoperable and reusable (Wilkinson et al. 2016), see e.g. https://www.force11.org/group/fairgroup/fairprinciples and https://www.go-fair.org/fair-principles/ . The FAIR data principles are not the only guide to good management and sharing of scientific data and users of this DMP are encouraged to consult other guidelines such as those from the Organisation for Economic Co-operation and Development (Pilat & Fukasaku 2007) and the Group on Earth Observations (2015).

Detailed authoritative guidelines for publishing biodiversity data following FAIR Data Principles have been published by Penev et al. (2017).

2.1. Making data findable, including provisions for metadata

Biodiversity Observation Data

We aim to make biodiversity observation data available to GBIF. The flow of data to GBIF will depend on the nature of the data and the way that they were collected. For example, users of the iNaturalist app will be encouraged to set their application setting so that verified observations are shared with GBIF automatically. This can be done in their preferences by setting their data sharing to CC0, CC-By or CC-By-NC. We recommend the use of the CC0 public domain dedication, because this makes the data most widely useable to the whole community.

Observations not automatically fed to GBIF will be formatted into Darwin Core (Wieczorek et al. 2012) and published using the Integrated Publishing Toolkit (Robertson et al. 2014).

Metadata associated with the data might include the following:

- ➔ A title and summary of the content
- ➔ Names and affiliations of all the participants, preferably with their ORCID IDs
- ➔ Specifications with respect to the language, rightsholder, license and type of data
- ➔ Methodology, including the effort associated with the surveys
- ➔ A statement on data quality, and conformity with standards or legislation.
- ➔ Details of any specimens that were collected; which institution they are deposited in and what their accession numbers are
- ➔ Geographic information on the sites surveyed, which might include shapefiles of sites and of routes taken
- ➔ Temporal information referring to the date and time of the survey
- ➔ A reference list of literature and websites used for the identification of organisms. Including their Digital Object Identifiers.

Dataset version numbers are created automatically on data repositories such as GBIF and Zenodo (https://zenodo.org/). Data will be provided with suitable keywords to identify the project, the

geographic area, country and the nature of the project. Data published on GBIF will be documented using the Ecological Metadata Language (EML) (Fegraus et al. 2005).

2.2. Making data openly accessible

The default for the project will be that all data will be open. However, participants should reflect on whether there are reasons of privacy or conservation why data should not be made open. We do not anticipate collecting personal data that cannot be shared under this project. Project participants will also have to follow the data sharing policies of their own organizations.

We intend that data will be shared in certified scientific data repositories. We recommend GBIF and Zenodo, but we do not exclude other options such as Dryad Digital Repository (https://datadryad.org/) for data associated with scientific papers. However, we caution participants against making data available from institutional servers and as supplementary data that are not certified data repositories. These generally lack the standards and infrastructure to ensure data are secure for long-term storage and retrieval.

We anticipate that the data collected under this project will not require specialised software tools to access. Much of the data will be held in simple data tables of columns and rows. These can be stored in basic text formats. If software is required to use the data we recommend that these tools are deposited together with the data in the repository. A full description of how to access the data should be provided together with examples.

For software there are many options for Open Source licence, however we suggest use of the MIT Licence (https://opensource.org/licenses/MIT). This licence is highly permissive to users.

Research articles

Article publication in the framework of COST should follow the COST guidelines on open access (COST 2015): *Action publications benefitting from COST funding shall - whenever possible - be made available as Open Access by means of self-archiving (also referred to as "green" Open Access) in an online repository before, during or after being published.* Explicit reference is made in the guidelines to the Directory of Open Access Journals (DOAJ, https://doaj.org/) and the Open Access Infrastructure for Research in Europe (OpenAIRE) in relation to choosing an appropriate open access target journal or the type of repository for archiving COST action publications respectively.

EU General Data Protection Regulation (GDPR)

One particular aspect of research on citizen science and of performing surveys is the storage and treatment of personal data e.g. personal information (names, email addresses) of people filling questionnaires or participants of citizen science projects and events. Personal information is subject to the General Data Protection Regulation (GDPR) rules in the European Union (https://gdpr.eu/). Also, there might be specific privacy policies from the COST programme office and Action participants are encouraged to consult the COST Guidelines for the communication, dissemination and exploitation of COST Action results and outcomes (COST 2015).

A general application of the GDPR legislation applies to information collected about people, for example surveys of members of the COST Action or surveys of citizen science project organisers. In this case we will only collect information that is relevant and necessary and we will be open and

honest about the use of the personal data, as required by GDPR. In general, data should be reported and stored in a way that anonymises the survey participant by removing personally identifiable information.

One specific application of GDPR which is of concern to us is personal information associated with biological observations (i.e. the identity of the observer). Specifically for the case of biodiversity observations Articles 5 and 89 of the GPPR regulations allows for legitimate use of personal data if it is in "*the public interest, scientific or historical research purposes or statistical purposes*" (European Parliament 2016). Most biodiversity observation and specimen data have been annotated with the name of the observer, usually by the observer themselves. Use of these data for biodiversity research and alien species management is compatible with the initial purpose of data collection. This is the position taken by other aggregators of biodiversity observations (Copas 2019, PLAZI 2019). Information on data protection regulations can be found here https://gdpr-info.eu/.

2.3. Making data interoperable

Where appropriate we will follow the standards of the Biodiversity Information Standards organization (TDWG). For biodiversity observations this will be achieved by using Darwin Core. For vocabularies we will follow community standards or frameworks where they exist. For example, the invasion pathway scheme of the Convention on Biological Diversity (2014) is widely used, as is the invasion impact classification of Blackburn et al. (2014).

We will follow ISO standards where applicable (i.e. ISO 8601 & ISO 3166 for dates and country codes respectively). For measurements we will use the metric SI units as appropriate.

For terminology related to citizen science projects Eitzel et al. (2017) provide advice.

Throughout the COST action, we will also align with the data models and standards, as currently under development and promoted by the CSA International Working Group on Citizen Science Data and metadata. Notably the Public Participation in Scientific Research (PPSR) Core metadata standards. More information can be found in the deliverable of Deliverable 1 of working group 5 of the COST Action CA15212 (Citizen Science to promote creativity, scientific literacy, and innovation throughout Europe) https://www.cs-eu.net/sites/default/files/media/2018/10/Deliverable%201%20-%20Citizen-science%20ontology%202018_09_13%20%28report%29.pdf.

2.4. Increase data re-use (through clarifying licences)

Participants will have to follow the data management policies of their own institutions, but we encourage participants to use the most open licensing possible. In order to allow the best possible data re-use, data policies should be clearly described and well recognised licenses should be applied (see, for example, Williams et al. 2018). This is because the use of standard and well known licences helps to avoid in-depth analysis of the access and use conditions on a case-by-case basis. Furthermore, we promote the use of machine readable licences, so that the conditions for data re-use can be automatically assessed.

The preference is to put data into the public domain by using the Creative Commons Zero Public Domain Dedication (https://creativecommons.org/publicdomain/zero/1.0/). The Creative Commons

Attribution licence (https://creativecommons.org/licenses/by/4.0/) is a more restrictive alternative, and if this is used then a clear statement of the expected attribution should be given with the licencing information.

3. Allocation of resources

Depositing and accessing data on Zenodo and GBIF are free to users. The time to prepare data will be at the cost of the organizer's institutions. However, some resources for data management might be made available through short-term scientific missions within the project. Any further publication costs may be applied for on the Alien CSI budget (CA17122) if appropriate. The COST Vademecum gives details of the costs eligable for reimbursement https://www.cost.eu/wp-content/uploads/2019/02/Vademecum-20190218-feb-update.pdf.

Long-term preservation costs for the data are guaranteed by Zenodo for at least 20 years, but they anticipate keeping the data indefinitely.

The individual researchers are responsible for the data they manage. However, guidance can be given by the Working Group leaders and the Alien-CSI Chair and Vice Chair. If there are data that may be lost as a result of a lack of resources, participants are encouraged to escalate the issue so that a solution can be found.

4. Data security

To ensure data are not lost or corrupted they should be backed-up on a regular basis and ultimately stored in a long-term repository, such as Zenodo. For working documents of non-sensitive information we encourage Alien-CSI participants to use cloud-based tools, such as GitHub, Google Docs and the Open Science Framework. Use of such platforms avoids creating a local infrastructure for data preservation.

If personal or sensitive data were to be collected then bespoke storage solutions will have to be considered. This will ensure that the data cannot be accidentally lost, and that they are secure from unauthorized access. For the most part we do not foresee sensitive data , in the sense used in GDBR, being gathered under Alien-CSI.

5. Ethical aspects

If project participants suspect there are ethical issues related to data they intend to collect they are encouraged to consult their institutional ethics guidelines. They are also welcome to consult the chairs and working group leaders of the COST action.

If personal data are collected in questionnaires then it is for the principal investigator to ensure that informed consent is given to store and share these data.

References

Bio Innovation Service (2018). Citizen science for environmental policy: development of an EU-wide inventory and analysis of selected practices. Final report for the European Commission, DG Environment under the contract 070203/2017/768879/ETU/ENV.A.3, in collaboration with Fundacion Ibercivis and The Natural History Museum, November 2018. https://publications.europa.eu/en/publication-detail/-/publication/842b73e3-fc30-11e8-a96d-01aa75ed71a1/language-en

Blackburn T, Essl F, Evans T, Hulme P, Jeschke J, Kühn I, Kumschick S, Marková Z, Mrugała A, Nentwig W, Pergl J, Pyšek P, Rabitsch W, Ricciardi A, Richardson D, Sendek A, Vilà M, Wilson JU, Winter M, Genovesi P, Bacher S (2014). A Unified Classification of Alien Species Based on the Magnitude of their Environmental Impacts. *PLoS Biology* 12 (5): e1001850. https://doi.org/10.1371/journal.pbio.1001850

Chandler M, See L, Copas K, Bonde AM, López BC, Danielsen F, Legind JK, Masinde S, Miller-Rushing AJ, Newman G, Rosemartin A (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation* 213: 280-294. https://doi.org/10.1016/j.biocon.2016.09.004

Convention on Biological Diversity (2014). Pathways of Introduction of Invasive Species, Their Prioritization, and Management. In CBD. UNEP/CBD/SBSTTA/18/9/Add.1, Montreal, Canada, June 2014, 18 pp

Copas K (2019, June 7). The GDPR-compliance waltz: balancing necessity, legitimate interests and data-subject rights in the GBIF network. https://doi.org/10.17605/OSF.IO/9R26B

COST (2015). Guidelines for the communication, dissemination and exploitation of COST Action results and outcomes. The COST Association. https://www.cost.eu/funding/how-to-get-funding/documents-and-guidelines/

Eitzel MV, Cappadonna JL, Santos-Lang C, Duerr RE, Virapongse A, West SE, Kyba CCM, Bowser A, Cooper CB, Sforzi A, Metcalfe AN, Harris ES, Thiel M, Haklay M, Ponciano L, Roche J, Ceccaroni L, Shilling FM, Dörler D, Heigl F, Kiessling T, Davis BY, Jiang Q (2017). Citizen science terminology matters: Exploring key terms. *Citizen Science: Theory and Practice* 2(1) 1-20. https://doi.org/10.5334/cstp.96

European Parliament (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. Official Journal of the European Union (OJ), 59(1-88), 294.

Group on Earth Observations (2015). Data Management Principles Implementation Guidelines. GEO-XII https://www.earthobservations.org/documents/geo_xii/GEO-XII_10_Data%20Management%20Principles%20Implementation%20Guidelines.pdf

Groom QJ, Adriaens T, Desmet P, Simpson A, De Wever A, Bazos I, Cardoso AC, Charles L, Christopoulou A, Gazda A, Helmisaari H, Hobern D, Josefsson M, Lucy F, Marisavljevic D, Oszako T, Pergl J, Petrovic-Obradovic O, Prévot C, Ravn HP, Richards G, Roques A, Roy HE, Rozenberg MAA, Scalera R, Tricarico E, Trichkova T, Vercayie D, Zenetos A, Vanderhoeven S (2017). Seven recommendations to make your invasive alien species data more useful. *Frontiers in Applied Mathematics and Statistics* 3:13. https://doi.org/10.3389/fams.2017.00013

Groom Q, Desmet P, Vanderhoeven S, Adriaens T (2015). The importance of open data for invasive alien species research, policy and management. *Management of Biological Invasions* 6(2): 119‑125. https://doi.org/10.3391/mbi.2015.6.2.02

Hulme PE, Bacher S, Kenis M, Klotz S, Kühn I, Minchin D, Nentwig W, Olenin S, Panov V, Pergl J, Pyšek P, Roques A, Sol D, Solarz W, Vilà M (2008). Grasping at the routes of biological invasions: a framework for integrating pathways into policy. *Journal of Applied Ecology* 45: 403-414. http://doi.org/10.1111/j.1365-2664.2007.01442.x

Luna S, Gold M, Albert A, Ceccaroni L, Claramunt B, Danylo O, Haklay M (2018). Developing Mobile Applications for Environmental and Biodiversity Citizen Science: Considerations and Recommendations. In: Joly A., Vrochidis S., Karatzas K., Karppinen A., Bonnet P. (eds) Multimedia Tools and Applications for Environmental & Biodiversity Informatics. Multimedia Systems and Applications. Springer, Cham https://doi.org/10.1007/978-3-319-76445-0_2

Penev L, Mietchen D, Chavan V, Hagedorn G, Smith V, Shotton D, Ó Tuama É, Senderov V, Georgiev T, Stoev P, Groom Q, Remsen D, Edmunds S (2017) Strategies and guidelines for scholarly publishing of biodiversity data. *Research Ideas and Outcomes* 3: e12431. https://doi.org/10.3897/rio.3.e12431

Pilat D, Fukasaku Y (2007). OECD principles and guidelines for access to research data from public funding. *Data Science Journal* 6: OD4-OD11.

PLAZI (2019). Reuse of person names on specimen labels, in scholarly publications, and Data Protection (GDPR). http://plazi.org/news/beitrag/reuse-of-person-names-on-specimen-labels-in-scholarly-publications-and-data-protection-gdpr/acd7dec0d7bea41f307100b4048295a1/ (06.06.2019)

Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P (2014). The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PloS One* 9(8): e102623. https://doi.org/10.1371/journal.pone.0102623

Robinson LD, Tweddle JC, Postles MC, West SE, Sewell J (2013). Guide to Running a BioBlitz 2.0. Natural History Museum, Bristol Natural History Consortium, Stockholm Environment Institute York and Marine Biological Association. http://www.bnhc.org.uk/communicate/guide-to-running-a-bioblitz-2-0/.

Roger E, Klistorner S (2016). BioBlitzes help science communicators engage local communities in environmental research. *Journal of Science Communication* 15(3): A06. https://doi.org/10.22323/2.15030206

Roy H, Groom Q, Adriaens T, Agnello G, Antic M, Archambeau A, Bacher S, Bonn A, Brown P, Brundu G, López B, Cleary M, Cogălniceanu D, de Groot M, De Sousa T, Deidun A, Essl F, Fišer Pečnikar Ž,

Gazda A, Gervasini E, Glavendekic M, Gigot G, Jelaska S, Jeschke J, Kaminski D, Karachle P, Komives T, Lapin K, Lucy F, Marchante E, Marisavljevic D, Marja R, Martín Torrijos L, Martinou A, Matosevic D, Mifsud C, Motiejūnaitė J, Ojaveer H, Pasalic N, Pekárik L, Per E, Pergl J, Pesic V, Pocock M, Reino L, Ries C, Rozylowicz L, Schade S, Sigurdsson S, Steinitz O, Stern N, Teofilovski A, Thorsson J, Tomov R, Tricarico E, Trichkova T, Tsiamis K, van Valkenburg J, Vella N, Verbrugge L, Vétek G, Villaverde C, Witzell J, Zenetos A, Cardoso A (2018). Increasing understanding of alien species through citizen science (Alien-CSI). *Research Ideas and Outcomes* 4: e31412. https://doi.org/10.3897/rio.4.e31412

Sturm U, Schade S, Ceccaroni L, Gold M, Kyba C, Claramunt B, Haklay M, Kasperowski D, Albert A, Piera J, Brier J, Kullenberg C, Luna S (2018). Defining principles for mobile apps and platforms development in citizen science. *Research Ideas and Outcomes* 4: e23394. https://doi.org/10.3897/rio.4.e23394

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PloS One* 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. https://doi.org/10.1038/sdata.2016.18

Williams J, Chapman C, Leibovici D, Loïs G, Matheus A, Oggioni A, Schade S, See L, van Genuchten P (2018). Maximising the impact and reuse of citizen science data. In: Hecker S, Haklay M, Bowser A, Makuch Z, Vogel J, Bonn A (eds.) Citizen Science - Innovation in Open Science, Society and Policy. (pp. 321-336). UCL Press: London.http://dx.doi.org/10.14324/111.9781787352339