PERSPECTIVE
https://doi.org/10.1038/s41559-017-0458-2
nature ecology & evolution

Navigating the unfolding open data landscape in ecology and evolution

Antica Culina[1*], Miriam Baglioni[2], Tom W. Crowther[1,3], Marcel E. Visser[1], Saskia Woutersen-Windhouwer[1] and Paolo Manghi[2]

Open access to data is revolutionizing the sciences. To allow ecologists and evolutionary biologists to confidently find and use the existing data, we provide an overview of the landscape of online data infrastructures, and highlight the key points to consider when using open data. We introduce an online collaborative platform to keep a community-driven, updated list of the best sources that enable search for data in one interface. In doing so, our aim is to lower the barrier to accessing open data, and encourage its use by researchers hoping to increase the scope, reliability and value of their findings.

Open data (see Box 1) have the potential to transform the sciences[1], providing a new depth of information that can facilitate advances across disciplines ranging from engineering to artificial intelligence, to economics, to medicine and to social sciences. Facilitated by recent advances in internet technologies and tools, and statistical approaches[2], these openly available data are beginning to provide unparalleled insights into complex systems. As scientific fields that are motivated by the search for unifying mechanisms, ecology and evolution inherently lend themselves to the value of open data[3–6]. Yet, until now, the application of open data has not pervaded these natural sciences[6–8].

Within ecology and evolution, the value of open data has been recognized in a few fields characterized by 'big data' (such as genomics, systematics and biogeography, see refs [3,9,10]), many of which also benefit from data originating from other scientific disciplines (medicine, geology or climate sciences, for example). However, the 'long tail' of ecological research (many individual projects producing small-scale data, see Box 1) has failed to fully embrace the open data movement[7,11], probably because of the heterogeneous nature of ecological research (for example, specific taxa, systems, regions or methodologies).

The increasing demand for the use of open data in ecology and evolutionary biology is exemplified best by the need to identify broader ecological and evolutionary patterns and processes across species, space and time[10]. Further benefits include the re-analysis of data using new statistical approaches, error checking or use of existing data to address new questions[11,14]. The relevance of ecological data to addressing many challenges of the Anthropocene largely depends on the power of combined ecological data, supplemented with the data from other disciplines such as geosciences, or economics[7,13,14]. For example, data reuse has been indispensable to our understanding of the climate system, and it has been pivotal in allowing us to constrain projections about future changes including warming[15] and biodiversity loss[16].

Thus, the aim of this Perspective is to provide ecologists and evolutionary biologists with the tools to overcome the daunting task of navigating the unfolding open data landscape, and to increase the use of this valuable resource for more robust and comprehensive analysis and conclusions.

A scattered landscape of open data in ecology

The number of scientific data repositories and data journals (and consequently the amount of open data) has dramatically increased in recent years largely as a result of recent efforts (for example, journal and funder policies on data archiving[4,6]) to enable a transparent, reproducible and efficient science where the previous work is preserved, and can easily be reused, validated and built upon[4,11,18,19]. Archiving data in repositories, or publishing them in an article form in the data journals, are two of the best venues to achieve long-term, findable, accessible, interoperable, and reusable data (FAIR data[6,11,20,21]). The Registry of Research Data Repositories (http://re3Data.org) currently lists more than 3,500 data repositories, out of which around 2,000 are classified under natural and life sciences. Other methods of data archiving, such as publishing data in the paper supplements or on personal websites, prevent data from being easily found, or attributed when used[6,7].

Ecological and evolutionary data are scattered across a large number of community specific and general repositories at present[4], because the culture of data sharing in these fields has started relatively recently, and because the data types and methods used to obtain these data are extremely diverse[3,4,6,17]. Locating the relevant data in this fragmented landscape is today partly mitigated by the places that harvest these primary data sources (that is, collect information) and provide one interface to search for data sets of interest (Fig. 1). Given the vast array of data sources, this valuable resource can be daunting to approach, particularly for researchers in the long tail of ecological research.

Enabling easy discovery of research data

To facilitate access to, and reuse of, data in ecology and evolution, we provide scientists with an up-to-date and evolving list of relevant online data discovery sources that allow searches for (as well as open access to) the data of interest. These data discovery sources harvest across many different primary data sources (for example, data repositories) in the same search interface, facilitating speed and breadth of data acquisition. To use a familiar analogy, this is equivalent to the search for articles via journal databases (such as the Web of Science and Scopus) that search through a wide range of the individual journals.

[1]NIOO-KNAW, Wageningen, The Netherlands. [2]Istituto di Scienza e Tecnologie dell'Informazione, CNR, Pisa, Italy. [3]Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland. *e-mail: A.Culina@nioo.knaw.nl

Box 1 | Glossary

Here we provide the list of terms that researchers might be unfamiliar with. These include the description of the principles of FAIR data.

Open data. A piece of data that anyone is free to use, reuse and redistribute — subject only, at most, to the requirement to attribute and/or share-alike. Equivalent to FAIR data.

FAIR data. To enable data to be found and used, data should ideally adhere to the FAIR principles. FAIR data are equivalent to open data.

Metadata. Data about data. Standardized structured information explaining the purpose and the origin of data, describing the structure of the data, time references, geographic location, creator, access conditions and terms of use of a data collection. Metadata answer the questions: why and how data were collected, what data have been collected, by whom, when and where.

Licensing. Policies and rules for data use (data released with a clear and accessible data usage licence).

Non-proprietary format. A format that allows the general use of data (the decoding and interpretation of this data is easily accomplished without a particular piece of software or hardware that was developed by a company or organization. For example, .csv belongs to this type of format (while Excel represents the opposite, proprietary format).

DOI. A unique and stable identifier that ensures that a digital object can be permanently found on the World Wide Web, regardless of changes in the web address where the object is found. A central registry ensures that the user of a DOI will be referred to its current location (for example, see http://www.datacite.org).

Long tail of science. Dispersed scientific research that is conducted by many individual researchers/teams, and is often of a limited spatial and temporal scale. Data produced in the long tail tend to be small in volume, and less standardized within the same field of study. The majority of scientific funding is spent on this type of research.

Reproducible research. Research that is documented in such a way that it allows for methods reproducibility (the ability to implement, as exactly as possible, the experimental and computational procedures) and results reproducibility ('replication', obtaining the same, or supporting results in a new study if the same procedures are followed).

Pertinence. A measure of relevance of the subset of the data source associated with the domain.

Open science. Science conducted in a way that makes all of the components of the research life-cycle available to anyone (preferably online). It includes open access to data sets, code (software), publications and peer review.

(Research) data life-cycle. A cycle composed of the stages through which research data go, from being collected (created), recorded, processed, analysed and finally published. It also includes preserving the data, giving access to data and reusing the data.

In this section we provide a classification of data sources in the domain of ecology and evolution, hereafter the EcoEvo domain, deliver a list of currently available sources to ease the identification of relevant data and establish a community-driven online platform that provides an evolving list (and description) of these sources and can be amended by the community members. We provide the ontology that we used to describe data sources in the Methods and Supplementary Information.

Data sources in ecology and evolution

Data sources on the Internet that a, EcoEvo researcher can refer to in order to find data of interest can be classified into five categories.

Data repositories host metadata (data describing and/or supporting the primary data set, Box 1) and files with research data. They can be thematic (for example, PANGAEA, which stores data for Earth and environmental science) or general purpose (Zenodo, FigShare, Dryad). To describe research data, data repositories typically adopt standard scientific metadata formats (such as DataCite) or domain-specific metadata formats (for example, Ecological Metadata Language (EML), Biological Metadata Language (BML) or EURING-code).

Aggregators of data repositories harvest or host metadata from a set of data repositories. One example is DataCite, which is an organization that issues digital object identifiers (DOIs, see Box 1) for several data repositories and keeps a searchable aggregation of their metadata records. Another is World Data System, which aggregates geo-located metadata records collected from around 70 data sources.

Virtual research environments (VREs) provide web user-interface (Web-UI) tools for scientists to collaborate or process/manipulate data. Examples include the VREs provided by D4Science.org to the biodiversity, fishery and aquaculture research communities.

Registries of data sources are directories of data sources that are intended to provide an organized, up-to-date and searchable collection of data sources. One example is re3data.org, which can be used to find international repositories for research data.

Data sources with links to data sets are data sources that may not be intended to preserve data set objects but offer the possibility to reach data sets via links from other scholarly objects, such as literature (scientific articles, theses, reports). Examples include advanced aggregators such as OpenAIRE, the Data-Literature Interlinking Service or CrossRef.

Although data repositories generally host data (primary data sources), the aggregators of data repositories, registries of data sources and data sources with links to data sets collect information (usually metadata) from data repositories, or other data sources, and provide links to these data sets. Thus, they facilitate the discovery of data sets across many different data repositories (or other data sources) in one interface (data discovery sources).

Where to find EcoEvo data

To facilitate the discovery and use of EcoEvo data (which are probably scattered across many data repositories) in Table 1 we provide a list of the main data discovery sources that contain and/or refer to data sets that are relevant to the EcoEvo domain. These sources allow us to search for data sets (hosted at different places across the data landscape) through one interface, thus increasing search speed, efficiency and coverage (for example, OpenDOAR harvests more than 300 relevant repositories). They also partly overlap in the content (the primary data sources, commonly repositories) they harvest information from. To secure the most comprehensive list of relevant search results, our recommendation is to use all of the listed sources. For example, we were interested in all of the open access data on the pedigrees of non-domesticated species: the degree of overlap in the results we obtained by the search using nine different aggregators was substantial (Fig. 2), however, almost all of them did provide at least one unique record.

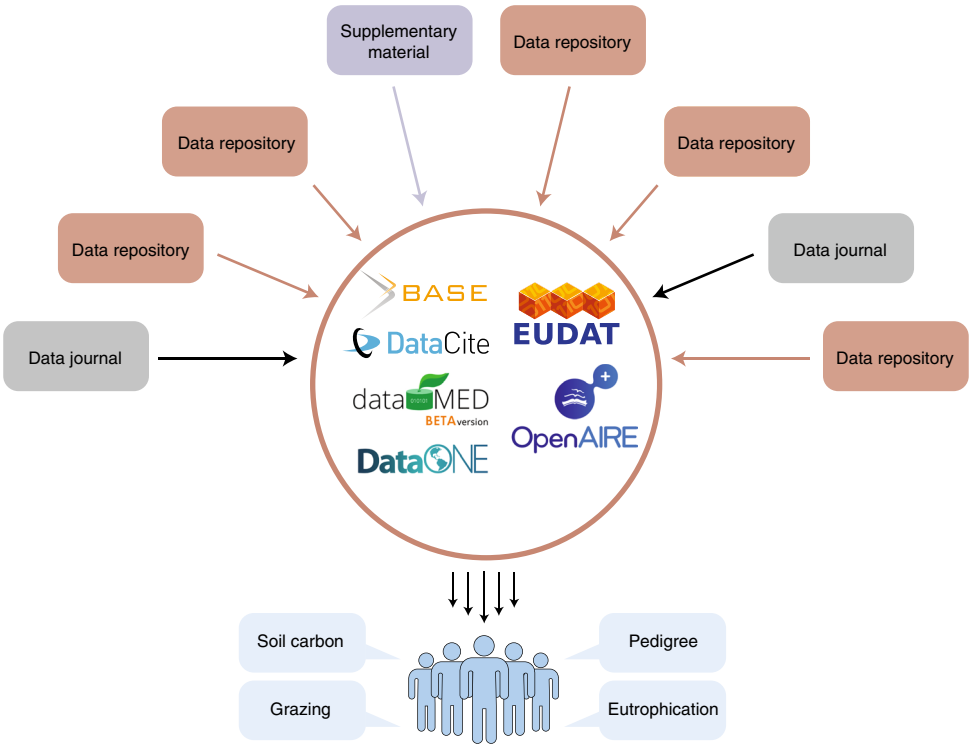

Fig. 1 | A schematic representation of the layered structure of the open data landscape. Data are stored in data repositories, data journals and supplements of scientific papers. The information provided by these primary data sources is then harvested by the data discovery sources (a few examples of these are presented). The researcher can use these aggregators to simultaneously find relevant data across many primary data sources.

For researchers interested in more domain- or community-specific data, searching directly in one or more domain-specific repositories (for example, Flybase) would be a more functional approach. To locate specific relevant repositories (according to keywords or subject areas) we recommend using the Registry of Open Access Repositories (ROAR; http://roar.eprints.org/cgi/search/advanced) and re3data.org.

Community-driven EcoEvo data source catalogue

As a part of our vision to increase the reuse of existing EcoEvo data, we have created an online, interactive VRE on the D4Science platform[22]. This VRE provides an up-to-date, searchable list of the best places to search for the data (data discovery sources) within the EcoEvo domain and can be accessed at https://ckan-ecoevo.d4science.org.

The items registered within the data catalogue are described with (and searchable by) features that characterize the data discovery source itself (such as the name and organization) and with features of EcoEvo data sets that the source hosts or collects information on (for example, available metadata formats, content reuse policies). The full list of the descriptor fields is provided in the Supplementary Information. The catalogue cannot be used to search for data themselves, but only to locate the best sources for data search. Given that each data discovery source contains (harvests from) a number of unique primary data sources (or other data sources), we advise community members to utilize all of listed data discovery sources when searching for a certain type of data. This approach ensures the retrieval of the most comprehensive list of relevant data sets.

Because the data landscape is rapidly evolving, we encourage community participation to keep this list current by accessing the VRE at https://services.d4science.org/group/ecoevo and requesting the rights to publish items in the catalogue. Once registered, each community member can add a new data source to the list by using the 'Datasources' option and then the 'Publish item' button. This will lead to 'Item information list' where some fields (for example, Name, and the link to the website containing the data source) are mandatory, and some are not (such as First Appearance). The explanation of each descriptor field can be found under the 'i' button on the right-hand side. Members of the VRE can share messages among themselves via the 'Share Update' functionality provided by the environment. The current version of the catalogue is a beta version; we welcome any suggestions to improve the functionality of the catalogue.

Recommendations when (re)using data

We outline four main sets of recommendations for researcher when reusing EcoEvo data sets: check any legal considerations, credit/acknowledge the authors (owners) of data sets, consider potential analysis issues when using others data and consider technical aspects while searching for data.

Legal considerations. Researchers might refrain from searching for and using open access data because they are unsure about the legal implications. Rights that may apply to research data are: intellectual property rights (copyrights, database rights, patent rights), privacy law, national security laws and contractual agreements (including trade secrets). However, in many cases research data can be reused without a fear of liability of infringement or breach of contract[23]. Here we briefly outline three main scenarios that researchers might encounter, and what to do in each case. First, data sets that come with copyright and database rights attached can be reused when permission is granted by a public unequivocal and non-exclusive licence. The most common licences that EcoEvo researchers might encounter are creative commons (CC) licences. For research data these are mainly CC-BY and CC0 licences[24]. A CC-BY licence means that the user is allowed to share and adapt the data, and only

Table 1 | A list of the main data discovery sources for searching for EcoEvo data in one interface, and for later accessing the data of interest

Name	Type of content	Web link
Aggregators of data repositories		
DataCite	Data sets	https://search.datacite.org/
WorldWideScience	Data sets, literature, multimedia	http://worldwidescience.org/index.html
BASE	Data sets, literature, multimedia, software, other	https://www.base-search.net/
Share	Data sets, literature, multimedia, projects, other	https://share.osf.io/discover
Dataone and One Mercury[a]	Data sets	https://search.dataone.org/#data/page/0https://cn.dataone.org/onemercury/
Science research	Data sets, literature, software, multimedia, other	http://scienceresearch.com/scienceresearch/advancedsearch.html
Research Data Australia	Data sets	https://researchdata.ands.org.au/
B2Find	Data sets, literature, other	http://b2find.eudat.eu/
DataHub	Data sets	https://datahub.io/dataset
Dliservice portal	Data sets, linked with publications	https://dliservice.research-infrastructures.eu/index.html#/
DataMed	Data sets	https://datamed.org/index.php
UK Research data discovery service	Data sets	http://ckan.data.alpha.jisc.ac.uk/dataset
ZanRan	Data sets	http://www.zanran.com/q/
DataSearch	Data sets	https://datasearch.elsevier.com/#/
Mendeley Data	Data sets	https://data.mendeley.com/
Data sources with links to data sets		
Europe PMC	Literature (links to data sets)	http://europepmc.org/
OpenAIRE	Data sets, literature, software, services	https://www.openaire.eu/search/
BioStudies	Descriptions of studies, links to their data	http://www.ebi.ac.uk/biostudies/
GoOA	Open access journals and additional files that include tables and supplementary materials, so one can search for data	http://gooa.las.ac.cn/external/about-us.jsp
Registries of data sources		
ROAR	Repositories and data sets	http://roar.eprints.org/content.html
OpenDOAR	Data sets, literature, software, multimedia	http://www.opendoar.org/search.php
Virtual research environments		
D4Science Integrated Data Catalogue	Databases, data sets, repositories	https://www.d4science.org/integrated-data-catalogue
Marine LifeWatch	Databases, repositories, methods	http://marine.lifewatch.eu

Sources are listed by type: aggregators of data repositories, data sources with links to data sets, registries of data sources and VREs. Each source is provided with the name and type of scholarly content it covers, and a link. A list with the full description of each data source is provided in Supplementary Table 1. [a]DataONE Search Tool for Scientific Data.

attribution (citing the data set, for example) is required. A CC0 licence waives all copyrights and database rights (in most countries) and dedicates the data to the public domain (completely free use of data, without any obligations). For example, all data sets hosted in Dryad (http://datadryad.org/pages/faq) and BioMed Central (https://www.biomedcentral.com/about/policies/open-data) have a CC0 licence attached. However, in accordance with academic norms reusers should still consider citing data[11]. A second common situation is when research data are protected by terms of use, in which case the user has to comply with these terms. Third, when no permission has been granted (the data set is protected by copyrights and database rights and no public licence for reuse has been granted, or a licensor has protected research data by a term of use), permission for reuse or republication should be requested from the data owner (usually the creator of the research data, or her/his employer). Finally, even when the full data set (including its metadata) is protected by any of the above-mentioned rights, using only parts of the data set ('entities' within the data set), and without using metadata or the organizational structure of the data set, there are no copyrights to comply with. This is because numeric values at the item level are 'uncopyrightable' data elements in a data set in most of the world, and can be copied and reused without copyright restrictions[25]. Metadata documentation can also give information about the possibilities for reusing data by providing intellectual property rights (including ownership) and regulations for reuse of the research data[26,27]. All of the above also applies to data that are not

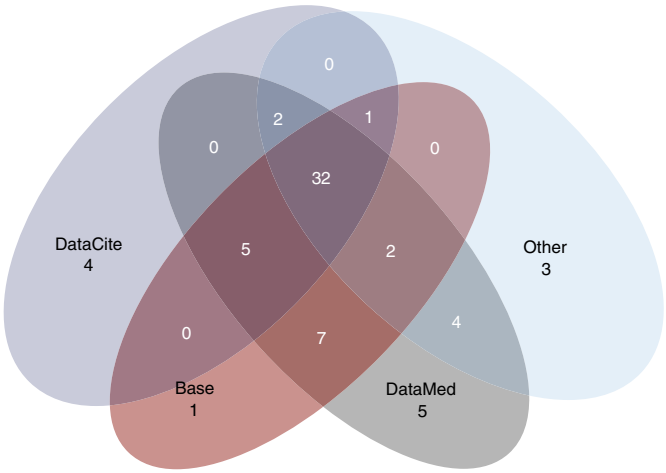

Fig. 2 | The distribution of search results on pedigree data (relatedness matrix) in natural and experimental animal populations. We used nine different data discovery sources. The diagram represents overlap between the three sources that returned the most results (DataCite, Base search engine and DataMed) and the combined results from other sources (Europe PMC, OpenAIRE, ScienceSerach, DataOne, Data Citation Index and the DLI Service). Overall, our search resulted in 66 relevant pedigree data sets. DataCite identified 4, Base 1, DataMed 5, and other sources 3 unique data sets.

public and will only be shared between a limited number of parties. In this case, the agreement should also consider privacy, national security, trade secret and patent rights.

Crediting authors of the data sets. The shift towards making data the first class research objects (that is, equivalent to the current status of a journal paper[28]), which can be cited and attributed[18,19] is an important component of the transition to open science (Box 1). Increasing numbers of data sets now have a DOI that can be used to cite data sets. Furthermore, many data sets (for example, those in Dryad) are accompanied with information on how to cite them, and whether to cite the related publication when citing the data.

Data misinterpretation and potential biases. Although an ideally described data set (how and where the data were collected, processed and analysed) should minimize any room for data misinterpretation, many ecological data sets still lack the complete information to enable a full understanding[12,29,30]. Furthermore, ecological data are often context specific, and their interpretation and informed use can only take place if this context is properly described (which is sometimes difficult to achieve, and it fully relies on whether the data owners had reuse in mind). For example, data misinterpretation due to many subtle biological and study-system specific details has been outlined as one of the main concerns about public data archiving for long-term studies[31]. Although this situation has been rapidly changing due to many initiatives to promote FAIR data (for example, Making Data Count, the Research Data Alliance), at the moment, contacting the authors of the data set before use is a good way to avoid data misuse. Second, working with a large amount of data requires careful consideration of the possible biases, statistical issues and inferences that can be drawn when using these data. For example, one recent study[32] identified multidimensional biases, gaps and uncertainties in global plant occurrence information data in the GBIF database, while another work[33] examined spatial biases in collected data sets used in two different meta-analysis that (wrongly) concluded that there was no net loss of biodiversity due to anthropogenic disturbances. This does not warn against data reuse, but rather calls for a rigorous scientific approach that identifies and accordingly addresses potential issues.

Technical considerations. Similar to the search platforms that researchers use to locate studies, different platforms for data search vary in their search functionality, and in the ways they harvest information from the primary data providers. To fully understand how the results of a search have been obtained, we suggest consulting the documentation on the updated search functionally of each platform that can be found on their website. We provide links to the relevant content in the Supplementary Information.

The future of open data in ecology and evolution

The benefits of ecological and evolutionary research pervade all aspects of society. However, the major historical drawbacks of ecological research (the challenge to standardize, validate and generalize findings) often limit the relevance of ecological findings for most urgent societal and scientific needs. Following advances in other scientific disciplines, a move towards increased utilization of open data across ecological and evolutionary disciplines can allow us to overcome some of these limitations. Comprehending systems that are as complex and extensive as natural ecosystems necessitates that we embrace the possibilities that are offered in the new open access era. By providing a structured overview of best data discovery sources for navigating the open data landscape and highlighting the necessary considerations when reusing others' data, we hope that this Perspective will encourage ecologists to embrace this valuable and ever improving scientific resource.

The open data landscape is not perfect, and so navigating it still requires a number of different considerations. This is particularly true for researchers in the long tail of ecological research, where data sources may be specialized, disjointed and difficult to interpret. This resource will, however, improve as it is increasingly adopted by the community. For example, at present there are varying levels of overlap between the major data aggregators, so a comprehensive data search must involve a combination of different search engines (Fig. 2). As the demand for these resources increases, and certain data aggregators emerge to guide the market, the efficiency and simplicity of data acquisition will probably improve.

EcoEvo biologists face a considerable (and relatively fast) leap to this data-intensive landscape. A crucial next step to increase the use and reuse of existing data sets is to raise awareness within the EcoEvo research communities, to inform them on the best places to easily search and access these data sets and to publish a number of 'benchmark studies' that will showcase the great potential of open data. By synthesizing the data landscape, we hope that our Perspective will promote the utilization of existing open data, driving a positive feedback loop that will ultimately encourage people to contribute and make use of more truly FAIR data sets, which will hopefully initiate a new era of open science.

Methods

An ontology is a formal vocabulary that describes the properties that characterize the domain of interest, and relationships between the components of this domain. The main purpose of an ontology is to enable the description, comparison and selection of entities (data sources in our case) according to a common conceptual schema. The ontology is typically agreed on and shared in the (scientific) domain. For the purpose of our work, we have defined an ontology for describing data sources that contain or refer to data sets relevant to the EcoEvo. The ontology makes a distinction between the data source and the collection of EcoEvo data sets that the data source contains. It enables the description and identification of data sources based on: (1) the identity of an EcoEvo data source (that is, features that characterize the data source itself, such as name and organization) and (2) the FAIRness of the EcoEvo data sets within the EcoEvo data source (that is, relevant data sets hosted/referred to within the data source; examples include the available metadata formats and content reuse policies). Indeed, most of the identified data sources host (or refer to) data sets from multiple disciplines of which only a subset

is relevant to the EcoEvo community, here called EcoEvo data sets. Three main use scenarios are: a data repository, an aggregator of data repositories and a data source with links to data sets. In the first case the EcoEvo data sets are hosted within the data source itself, whereas in the second and third cases the EcoEvo data sets are hosted elsewhere, and the data source contains the information about these data sets (and thus enable the researcher to find the data sets). We provide detailed description on how we developed this ontology in the Supplementary Information. We used the ontology to describe the complete list of sources that can be used to search for the EcoEvo data sets in one search interface. The list of sources can be used by scientists to search for the location of the data of interest, while new sources can be added to the list (using the ontology).

Identity of a data source. The identity of data source is characterized by a persistent identifier (if any), an official name, a textual description of the data source, the type of data source (for example, the type of data source of DataCite is aggregator of data repositories), the languages used to describe the data source objects (data sets or publications), the list of organizations maintaining and supporting the data source and a degree of pertinence to the EcoEvo domain. Data source persistent identifiers are not mandatory, but when present are typically issued by a directory/registry of sources, such as re3data.org for data repositories or OpenDOAR for literature repositories. The presence of the organizations behind the data source may be important to discover data sources of interests, but ultimately, based on their level of branding, may indirectly suggest the level of trust and reliability of the data source (that is, the organization supporting and maintaining the data can be in some cases a guarantor of quality). Finally, the degree of domain pertinence represents a novel but key measure of the correlation between a data source and the EcoEvo domain. Such a measure can be quantified by (1) the proportion of the overall content of the data source that contains EcoEvo data sets (that is, if all of the data sets related to the data source are EcoEvo data sets, the data source is 'highly pertinent') and (2) the degree of discipline focus of the EcoEvo data sets (that is, a data source with a small subset of EcoEvo data sets that are strongly related with the domain is 'highly pertinent'). For example EuropePMC is a highly pertinent data source, which allows the user to search articles with links to data sets from the same domain and in a domain pertinent way (that is, by exploiting the Medical Subject Heading (MeSH) terms and category). Cross-domain sources that have subsets of data sets (to be identified by tag/topic-driven queries or similar) relative to the domain of interest will have a lower degree of pertinence.

FAIRness of data source EcoEvo data sets. A researcher searching for EcoEvo data sets might potentially be interested in identifying data sources based on features of the data sets that these sources contain. We opted for an ontology that represents the characteristics of EcoEvo data sets of a data source in terms of the FAIR principles of data stewardship[20] (also see Box 1).

Findability. To support discovery by findability, the ontology includes a description of how to find/identify/discover the EcoEvo data sets within the data source. Such description will be provided as free text, for example, in the case of EuropePMC it could be 'search articles by MeSH terms and categories in order to identify relevant data sets'.

Accessibility. To support discovery by accessibility, the ontology includes the EcoEvo subjects that are covered by the data sets that the data source contains, the presence of links to other objects and to the web page of the data source. The subjects are terms of a predefined list specific to the EcoEvo domain, useful to filter data sources based on the scientific needs of the interested researcher. The presence of links is an important aspect to be considered in the era of open science, where data sources should not be interpreted as independent 'silos' of content, but rather nodes of an interconnected network. Finally, the data source web page is the means to directly access the data source.

Interoperability. To support discovery by interoperability, the ontology contains the metadata formats (for example, EML) used to describe data sets in the data source and the data set formats (such as a database entry, .csv or time series).

Reusability. To support discovery by reusability, the ontology includes the set of metadata reuse licences and the set of object reuse licences supported by the data source (for example, CC-BY, CC-0).

The interested reader can find the full and detailed list of properties in the Supplementary Information.

Received: 9 May 2017; Accepted: 19 December 2017;
Published online: 16 February 2018

References

1. Masuzzo, P. & Martens, L. Do you speak open science? Resources and tips to learn the language. *PeerJ Prepr.* **5**, e2689v1 (2017).
2. Hey, T., Tansley, S. & Tolle, K. (eds) in *The Fourth Paradigm: Data-Intensive Scientific Discovery* Ch. 3 & 4 (Microsoft, Redland, 2009).
3. Reichman, O. J., Jones, M. B. & Schildhauer, M. P. Challenges and opportunities of open data in ecology. *Science* **331**, 703–705 (2011).
4. Hampton, S. E. et al. Big data and the future of ecology in a nutshell. *Front. Ecol. Environ.* **11**, 156–162 (2013).
A brief overview of the importance of embracing the data as an important research product in ecology, issues it in doing this, and current efforts that help the change to happen.
5. Hampton, S. E. et al. The Tao of open science for ecology. *Ecosphere* **6**, 1–13 (2015).
6. Michener, W. K. Ecological data sharing. *Ecol. Inform.* **29**, 33–44 (2015).
History and future of data sharing in ecology, including the role of sociological changes and cyberinfrastructures, and best practices for data sharing.
7. Wallis, J. C., Rolando, E. & Borgman, C. L. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE* **8**, e67332 (2013).
8. Evans, S. R. Gauging the purported costs of public data archiving for long-term population studies. *PLoS Biol.* **14**, 1–9 (2016).
9. Piwowar, H. A. & Fridsma, D. B. Examining the uses of shared data. *Nat. Preced.* https://doi.org/10.1038/npre.2007.425.3 (2007).
10. Kenall, A., Harold, S. & Foote, C. An open future for ecological and evolutionary data? *BMC Evol. Biol.* **14**, 1–6 (2014).
11. Whitlock, M. C. Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* **26**, 61–65 (2011).
12. White, E. P. et al. Nine simple ways to make it easier to (re)use your data. *Ideas Ecol. Evol.* **6**, 1–10 (2013).
13. Losos, J. B. et al. Evolutionary biology for the 21st century. *PLoS Biol.* **11**, e1001466 (2013).
14. Mcnutt, B. M. et al. Liberating field science samples and data. *Science* **351**, 1024–1026 (2016).
15. Crowther, T. W. et al. Quantifying global soil carbon losses in response to warming. *Nature* **54**, 104–108 (2016).
16. Hawkins, S. J. et al. Data rescue and re-use: recycling old information to address new policy concerns. *Mar. Policy* **42**, 91–98 (2013).
17. Lindenmayer, D. B. et al. Value of long-term ecological studies. *Austral Ecol.* **37**, 745–757 (2012).
18. Duke, C. S. & Porter, J. H. The ethics of data sharing and reuse in biology. *Prof. Biol.* **63**, 483–489 (2013).
19. *Guidelines on Open Access to Scientific Publications and Research Data in Horizon* 2020 (European Commission, 2016).
20. Wilkinson, M. D. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2**, 1–9 (2016).
21. Assante, M., Candela, L., Castelli, D. & Tani, A. Are scientific data repositories coping with research data publishing? *Data Sci. J.* **15**, 1–24 (2016).
22. Candela, L., Castelli, D., Manzi, A. & Pagano, P. Realising virtual research environments by hybrid data infrastructures: the D4Science Experience. In *International Symposium on Grids and Clouds 2014* https://pos.sissa.it/210/022 (Proceedings of Science, 2014).
23. Carroll, M. W. Sharing research data and intellectual property law: a primer. *PLoS Biol.* **13**, e1002235 (2015).
24. *Data* (Creative Commons Wiki, accessed 1 May 2017); https://wiki.creativecommons.org/wiki/Data
25. Reichman, J. H. & Uhlir, P. F. A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment. *Law Contemp. Prob.* **66**, 315–462 (2003).
26. Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B. & Stafford, S. G. Nongeospatial metadata for the ecological sciences. *Ecol. Appl.* **7**, 330–342 (1997).
27. Schmidt-Kloiber, A., Moe, S. J., Dudley, B., Strackbein, J. & Vogl, R. The WISER metadatabase: the key to more than 100 ecological datasets from European rivers, lakes and coastal waters. *Hydrobiologia* **704**, 29–38 (2013).
28. Kratz, J. E. & Strasser, C. Comment: Making data count. *Sci. Data* **2**, 10–14 (2015).
29. Magee, A. F., May, M. R. & Moore, B. R. The dawn of open access to phylogenetic data. *PloS ONE* **9**, e110268 (2014).
30. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol.* **13**, e1002295 (2015).
31. Mills, J. A. et al. Archiving primary data: solutions for long-term studies. *Trends Ecol. Evol.* **30**, 581–589 (2015).
32. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
33. Gonzales, A. et al. Estimating local biodiversity change: a critique of papers claiming no net loss of local diversity. *Ecology* **97**, 1949–1960 (2016).

Acknowledgements

We thank L. Candela, M. Assante, F. Mangiacrapa, C. Perciante and A. Dell'Amico for their support in the deployment and customization of the catalogue. We thank D4Science Infrastructure (www.d4science.org) for hosting the catalogue.

Author contributions

A.C. collected the data (list of resources) and wrote the majority of the manuscript. M.E.V., T.W.C., S.W.-W., P.M. and M.B all contributed to the manuscript. P.M. and M.B. established the methodological approach for the data source description and the D4Science data catalogue. S.W.-W. also provided insights into the legal implications of data use.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-017-0458-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.