Research Ideas and Outcomes 3: e12431
doi: 10.3897/rio.3.e12431

Guidelines

Strategies and guidelines for scholarly publishing of biodiversity data

Lyubomir Penev[‡], Daniel Mietchen[§], Vishwas Shravan Chavan[|], Gregor Hagedorn[¶], Vincent Stuart Smith[#], David Shotton[¤], Éamonn Ó Tuama[«], Viktor Senderov[»,^], Teodor Georgiev[»], Pavel Stoev[˅], Quentin John Groom[¦], David Remsen[ˀ], Scott C. Edmunds[˂]

‡ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
§ EvoMRI Communications, Jena, Germany
| Global Biodiversity Information Facility, Copenhagen, Denmark
¶ Museum für Naturkunde Berlin, Berlin, Germany
# The Natural History Museum, London, United Kingdom
¤ University of Oxford, Oxford, United Kingdom
« Independent Researcher, Cork, Ireland
» Pensoft Publishers, Sofia, Bulgaria
^ Institute of Biodiversity & Ecosystem Reearch, Sofia, Bulgaria
˅ National Museum of Natural History and Pensoft Publishers, Sofia, Bulgaria
¦ Botanic Garden Meise, Meise, Belgium
ˀ Marine Biological Laboratory, Woods Hole, United States of America
˂ GigaScience, BGI HK Ltd., Tai Po Industrial Estate, Hong Kong, Hong Kong

Corresponding author: Lyubomir Penev (penev@pensoft.net)

Reviewed v1

Received: 26 Feb 2017 | Published: 28 Feb 2017

Citation: 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

Abstract

The present paper describes policies and guidelines for scholarly publishing of biodiversity and biodiversity-related data, elaborated and updated during the Framework Program 7 EU BON project, on the basis of an earlier version published on Pensoft's website in 2011. The document discusses some general concepts, including a definition of datasets, incentives to publish data and licenses for data publishing. Further, it defines and compares several routes for data publishing, namely as (1) supplementary files to research articles, which may be made available directly by the publisher, or (2) published in a specialized open data repository with a link to it from the research article, or (3) as a data paper, i.e., a specific, stand-alone publication describing a particular dataset or a collection of datasets, or (4)

© Penev L et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

integrated narrative and data publishing through online import/download of data into/from manuscripts, as provided by the Biodiversity Data Journal.

The paper also contains detailed instructions on how to prepare and peer review data intended for publication, listed under the Guidelines for Authors and Reviewers, respectively. Special attention is given to existing standards, protocols and tools to facilitate data publishing, such as the Integrated Publishing Toolkit of the Global Biodiversity Information Facility (GBIF IPT) and the DarwinCore Archive (DwC-A).

A separate section describes most leading data hosting/indexing infrastructures and repositories for biodiversity and ecological data.

Keywords

biodiversity data publishing, data publishing licenses, Darwin Core, Darwin Core Archive, data re-use, data repository

Data Publishing in a Nutshell

Introduction

Data publishing in this digital age is the act of making data available on the Internet, so that they can be downloaded, analysed, re-used and cited by people and organisations other than the creators of the data (Altman and King 2007, Green 2009). This can be achieved in various ways. In the broadest sense, any upload of a dataset onto a freely accessible website could be regarded as “data publishing”. There are, however, several issues to be considered during the process of data publication, including:

- Data hosting, long-term preservation and archiving
- Documentation and metadata
- Citation and credit to the data authors
- Licenses for publishing and re-use
- Data interoperability standards
- Format of published data
- Software used for creation and retrieval
- Dissemination of published data

The present guidelines are based on an earlier version published in PDF on Pensoft's website in 2011 (Penev et al. 2011). However, the process of implementation of data publishing practices in Pensoft's journals started earlier (Penev et al. 2009a, Penev et al. 2009b). Since that time, several novel approaches in both biodiversity and general research data publishing have been developed, mostly due to large-scale international efforts through networks such as FORCE11 (Future of Research Communication and e-

Scholarship), CODATA (The Committee on Data for Science and Technology), RDA (Research Data Aliance) and others.

The FORCE11 group dedicated to facilitating change in knowledge creation and sharing, recognising that data should be valued as publisheable and citable products of research, has developed a set of principles for publishing and citing such data. The FAIR Data Publishing Group formulated the following four FAIR principles of fata publishing (Wilkinson et al. 2016):

- Data should be **Findable**
- Data should be **Accessible**
- Data should be **Interoperable**
- Data should be **Re-usable.**

A key outcome of FORCE11 is the Joint Declaration of Data Citation Principles (see also Martone, M (Ed.) 2014 and Altman et al. 2015). These principles, organised under eight groupings, are abstracted here:

- **Importance**: Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
- **Credit and Attribution**: Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
- **Evidence**: In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
- **Unique Identification**: A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
- **Access**: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
- **Persistence**: Unique identifiers — and metadata describing the data and its disposition — should persist, even beyond the lifespan of the data they describe.
- **Specificity and Verifiability**: Data citations should facilitate identification of, access to, and verification of the specific data or datum that support a claim. Citations or citation metadata should include information about provenance and permanence sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.
- **Interoperability and Flexibility**: Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

The Research Data Alliance (RDA) promotes the open sharing of data by building upon the underlying social and technical infrastructure. Established in 2013 by the European Union, the National Science Foundation and the National Institute of Standards and Technology (USA) as well as the Department of Innovation (Australia), it has grown to include some 4,200 members from 110 countries who collaborate through Work and Interest Groups "to develop and adopt infrastructure that promotes data-sharing and data-driven research, and accelerate the growth of a cohesive data community that integrates contributors across domain, research, national, geographical and generational boundaries" (Research Data Alliance (RDA) 2017). These groups develop recommendations and outputs which, to date, have tended to address the common foundations for a data sharing infrastructure. For example, among those recommendations endorsed or in process of endorsement are:

- Data Description Registry Interoperability Model
- Persistent Identifier Type Registry
- Workflows for Research Data Publishing: Models and Key Components
- Bibliometric Indicators for Data Publishing
- Dynamic Data Citation Methodology
- Repository Audit and Certification Catalogues

One RDA output, the Scholix Inititive, under the RDA/WDS (ICSU World Data System) Publishing Data Services Work Group is of particular relevance, as it seeks to develop an interoperability framework for exchanging information about the links between scholarly literature and data, i.e., what data underpins literature and what literature references data.

Within RDA, a Biodiversity Data Integration Interest Group has been established, which aims to "increase the effectiveness of biodiversity e-Infrastructures by promoting the adoption of common tools and services establishing data interoperability within the biodiversity domain, enabling the convergence on shared terminology and routines for assembling and integrating biodiversity data."

With regard to biodiversity, some recently published papers emphasise the importance of publishing of biodiversity data (Smith 2009, Costello 2009, Costello et al. 2013, Smith et al. 2013, Hardisty et al. 2013). The urgent need for open, comprehensive, discoverable, interoperable, and reliable biodiversity data was further reinforced by the Aichi Biodiversity Targets of the United Nations' Strategic Plan for Biodiversity which have set an ambitious plan to stop biodiversity loss by 2020 (Convention on Biological Diversity 2011). The key prerequisite for progressing, monitoring and achieving the Aichi targets is the implementation of policies, strategies and actions. These should be based on new approaches, methods and infrastructure for the collection, aggregation, curation, publication and dissemination of data. On the way to it, scientists and policy makers have to overcome several barriers and fill in many gaps in both our knowledge of biodiversity and associated ecosystem services and in the means we obtain, handle, process, and publish data (Wetzel et al. 2015).

The EU BON project funded by the European Union's Framework Program Seven (FP7) (Building the European Biodiversity Observation Network, grant agreement ENV30845) was launched to contribute towards the achievement of these challenging tasks within a much wider global initiative, the Group on Earth Observations Biodiversity Observation Network (GEO BON), which itself is a part of the Group of Earth Observation System of Systems (GEOSS). A key feature of EU BON is the delivery of near-real-time data, both from on-ground observation and remote sensing, to the various stakeholders to enable greater interoperability of different data layers and systems, and provide access to improved analytical tools and services; furthermore, EU BON is supporting biodiversity science-policy interfaces, facilitate political decisions for sound environmental management (Hoffmann et al. 2014, Wetzel et al. 2015). A sound basis for pursuing these goals is the GEOSS 10-year Implementation Plan adopted in 2005, which has outlined a set of Data Sharing Principles (DSPs) (see also Uhlir et al. 2009).

The present paper outlines the strategies and guidelines needed to support the scholarly publishing and dissemination of biodiversity data, that is publishing through the academic journal networks.

What Is a Dataset

A dataset is understood here as a digital collection of logically connected facts (observations, descriptions or measurements), typically structured in tabular form as a set of records, with each record comprising a set of fields, and recorded in one or more computer data files that together comprise a data package. Certain types of research datasets, e.g., a video recording of animal behaviour, will not be in tabular form, although analyses of such recordings may be. Within the domain of biodiversity, a dataset can be any discrete collection of data underlying a paper – e.g., a list of all species occurrences published in the paper, data tables from which a graph or map is produced, digital images or videos that are the basis for conclusions, an appendix with morphological measurements, or ecological observations.

More generally, with the development of XML-based publishing technologies, the research and publishing communities are coming to a much wider definition of data, proposed in the BioMed Central (BMC) position statement on open data: "the raw, non-copyrightable facts provided in an article or its associated additional files, which are potentially available for harvesting and re-use" (BioMed Central 2010).

As these examples illustrate, while the term "dataset" is convenient and widely used, its definition is vague. Data repositories such as Dryad, wishing for precision, do not use the term "dataset". Instead, they describe **data packages** to which metadata and unique identifiers are assigned. Each data package comprises one or more related **data files**, these being data-containing digital files in defined formats, to which unique identifiers and metadata are also assigned. Nevertheless, the term "dataset" is used below, except where a more specific distinction is required.

For practical reasons, we propose a clear distinction between static data that represent specific completed compilations of data upon which the analyses and conclusions of a given scientific paper may be based, and curated data that belong to a large data collection (usually called a "database") with ongoing goals and curation, for example the various bioinformatics databases that curate ever growing amounts of nucleotide sequences (Cochrane et al. 2015). Both forms are of strong potential scientific interest and application. Where a static dataset is inextricably linked to a scientific paper, the data publisher must assure consistent and secure access to it on the same time scale as the text content of the digital article. As a consequence, it is not permissible to upload a new version of such data in ways that would replace the original, unless strict versioning is undertaken and the reader of the published article has easy access to the original version of the data resource as well as to updated versions.

Curated data, on the other hand, are usually hosted on external servers or in data hosting centres. A primary goal of the data publishing process in this case is to guarantee that these data are properly described, up to date, available to others under appropriate licensing schemes, peer-reviewed, interoperable, and where appropriate linked from a research article or a data paper at the time of publication. Especially in cases where the long-term viability of the curated project may be insecure (e.g. in the case of grant funded projects) (Chandras et al. 2009), the publisher may in addition support the publication of a dated and versioned copy of such data (with the option to update these with another version later on, keeping access to all versions).

Why Publish Data

Data publishing has become increasingly important and already affects the policies of the world's leading science funding frameworks and organizations — see for example the NSF Data Management Plan Requirements, the data management policies of the National Institutes of Health (NIH), Wellcome Trust, or the Riding the Wave (How Europe Can Gain From the Rising Tide of Scientific Data) report submitted to the European Commission in October 2010. More generally, the concept of "open data" is described in the Protocol for Implementing Open Access Data, the Open Knowledge/Data Definition, the Panton Principles for Open Data in Science, and the Open Data Manual. There are several incentives for authors and institutions to publish data (after Costello 2009, Smith 2009, with additions and changes):

- There is a widespread conviction that data produced using public funds should be regarded as a common good, and should be openly published and made available for inspection, interpretation and re-use by third parties.
- Open data increases transparency and the overall quality of research; published datasets can be re-analyzed and verified by others.
- Published data can be cited and re-used in the future, either alone or in association with other data.
- Open data can be integrated with other datasets across both space and time.
- Data integration increases recognition and opportunities for collaboration.

- Open data increases the potential for interdisciplinary research, and for re-use in new contexts not envisaged by the data creator.
- Needless duplication of data-collecting efforts and associated costs will be reduced.
- Published data can be indexed and made discoverable, browsable and searchable through internet services (e.g. Web search engines) or more specific infrastructures (e.g., GBIF for biodiversity data).
- Collection managers can trace usage and citations of digitized data from their collections.
- Data creators, and their institutions and funding agencies, can be credited for their work of data creation and publication through the conventional channels of scholarly citation; priority and authorship is achieved in the same way as with a publication of a research paper.
- Datasets and their metadata, and any related data papers, may be inter-linked into research objects, to expedite and mutually extend their dissemination, to the benefit of the authors, other scientists in their fields, and society at large.
- Published data may be structured as "Linked Data", by which term is meant data accessible using RDF, the Resource Description Framework, one of the fundamentals of the semantic web. Since RDF descriptions are based on publicly available ontology terms, ideally derived from a limited number of complementary ontologies, this permits automated data integration, since data elements from different sources have built-in syntactic and semantic alignment.

How to Publish Data

There are four main routes for scholarly publication of data, most of which are available with various journals and publishers:

1. Supplementary files underpinning a research paper and available from the journal's website.
2. Data hosted at external repositories but linked back from the research article it underpins.
3. Stand-alone description of the data resource in the form of scholarly publication (e.g., Data Paper, or Data Note - see, for example, Newman and Corke 2009, Chavan and Penev 2011, and Candela et al. 2015).
4. Data published within the article text and downloadable from there in the form of structured data tables or as a result of text mining. This "integrated data publishing" approach has been implemented by the Biodiversity Data Journal (BDJ), which was developed in the course of the EU funded project ViBRANT (Smith et al. 2013). Other examples of a similar approach are executable code published in an article (Veres and Adolfsson 2011), or linking of a standard article to an integrated external platform that hosts all data associated with the article, and provides additional data analysis tools and computing resources (an example for that are GigaDB and the GigaScience journal - see Edmunds et al. 2016), or various kinds of implementing 3D visualisations on the basis of MicroCT files (Stoev et al. 2013).

Within these main data publishing modes, Pensoft developed a specific set of applications designed to meet the needs of the biodiversity community. Most of these were implemented in the Biodiversity Data Journal and its associated ARPHA Writing Tool (AWT):

- Import of primary biodiversity data from Darwin Core compliant spreadsheets, or manually via a Darwin Core editor, into manuscripts and their consequent publication in a structured and downloadable format (Smith et al. 2013).
- Direct online import of Darwin Core compliant primary biodiversity data from GBIF, Barcode of Life, iDigBio, and PlutoF into manuscripts through web services and their consequent publication in a structured and downloadable format (Senderov et al. 2016).
- Import of multiple occurrence records of voucher specimens associated with a particular Barcode Index Number (BIN) (Ratnasingham and Hebert 2013) from the Barcode of Life.
- Automated generation of data paper manuscripts from Ecological Metadata Language (EML) metadata files stored at GBIF Integrated Publishing Toolkit (GBIF IPT), DataONE, and the Long Term Ecological Research Network (LTER) (Senderov et al. 2016, see also Pensoft's blog for details).
- Automated export of the occurrence data published in BDJ into Darwin Core Archive (DwC-A) format (Wieczorek et al. 2012) and its consequent ingestion by GBIF. The DwC-A is freely available for download from each article's webpage that contains occurrence data.
- Automated export of the taxonomic treatments published in BDJ into Darwin Core Archive. The DwC-A is freely available for download from each article that contains taxonomic treatments data.
- Novel article types in the ARPHA Writing Tool and its associated journals (Biodiversity Data Journal, Research Ideas and Outcomes (RIO Journal), and One Ecosystem): Monitoring Schema, IUCN Red List compliant Species Conservation Profile (Cardoso et al. 2016), IUCN Global Invasive Species Database (GISD) compliant Alien Species Profile, Single-media Publication, Data Management Plan, Research Idea, Grant Proposal, and others.
- Nomenclatural acts modelled and developed in BDJ as different types of taxonomic treatments for plant taxonomy.
- Markup and display of biological collection codes against the Global Registry of Biological Repositories (GRBIO) vocabulary (Schindel et al. 2016).
- Workflow integration with the GBIF Integrated Publishing Toolkit (IPT) for deposition, publication, and permanent linking between data and articles, of primary biodiversity data (species-by-occurrence records), checklists and their associated metadata (Chavan and Penev 2011).
- Workflow integration with the Dryad Data Repository for deposition, publication, and permanent linking between data and articles, of datasets other than primary biodiversity data (e.g., ecological observations, environmental data, genome data and other data types) (see Pensoft blog for details).

- Automated archiving of all articles published in Pensoft's journals in the Biodiversity Literature Respository (BLR) of Zenodo on the day of publication.

Best practice recommendations

- For any form of data publishing, follow the FAIR Data Publishing Principles (Wilkinson et al. 2016).
- Follow the Joint Declaration of Data Citation Principles for citation of data in scholarly articles (Altman et al. 2015).
- Deposition of data in an established international repository is always to be preferred to supplementary files published on a journal's website.
- Smaller data files, especially those directly underpinning an article, should also be deposited at a data repository and linked from the article. We recommended, however these to be published also as supplementary file(s) to the related article, to ensure an additional joint preservation and presentation of the article together with its associated data.
- If a specialized and well establisdhed repository for a given kind of data exists, it should be preferred over non-specialized ones (see also section "Data Deposition in Open Repositories" below for finer detail), for example:
    - Primary biodiversity data (species-by-occurrence) records should be deposited through the GBIF IPT.
    - Sample-based biodiversity data (e.g., species abundances from monitoring or inventory studies) should be deposited through the GBIF IPT.
    - Genomic data should be deposited at any of the three INSDC repositories (GenBank, European, Nucleotide Archive, ENA and the DNA Databank of Japan, DDBI) either directly or via an affiliated repository, e.g. Barcode of Life Data Systems (BOLD).
    - Barcoding and metabarcoding data should be deposited at the Barcode of Life Data Systems (BOLD) or PlutoF.
    - Metagenomic data should be deposited at EBI Metagenomics
    - Protein sequence data should be deposited at UniProtKB.
    - X-ray microtomography (micro-CT) scans should be deposited at Morphosource.
    - Phylogenetic data should be deposited at TreeBASE.
- Heterogeneous datasets, or data packages containing various data types should be deposited in generalist repositories, for example Dryad Data Repository, Zenodo, Dataverse, or in another appropriate repository.
- Repositories not mentioned above or in the "Data Deposition in Open Repositories" section below, may be used at the discretion of the author, if they provide long-term preservation of various data types, persistent identifiers to datasets, discoverability, open access to the data, and well proven sustainablility record.
- Digital Object Identifiers (DOIs) or other persistent identifiers (e.g., "stable URIs") to the data deposited in repositories, as well as the name of the repository, **should always be published** in the paper using or describing that data resource.

- Exceptional cases when publication of data is not possible, or some of the data remain closed or obfuscated, should be discussed with the publisher in advance. In such cases, the authors should provide an open statement explaining why restrictions in open data publishing are needed to be put in force. The author's statement should be published together with the article.

How to Cite Data

This section originates from a draft set of Data Citation Best Practice Guidelines that has been developed for publication by David Shotton, with assistance from colleagues at Dryad and elsewhere, and from earlier papers concerning data citation mechanisms (Altman and King 2007, Green 2009, Penev et al. 2009a). It also encompasses the latest international efforts to standardise the data and software citation mechanisms carried out within the CODATA, FORCE11 and RDA networks (CODATA/ITSCI 2013, Starr et al. 2015, Rauber et al. 2016, Smith et al. 2016).

The well-established norm for citing genetic data, for example, is that one simply cites the GenBank identifier (accession number) in the text. Similar usage is also commonplace for items in other bioinformatics databases. The latest developments in the implementation of the data citation principles, however, strongly recommend references to data to be included in the reference lists, similarly to literature references (Rauber et al. 2016). The following guidelines apply to more heterogeneous research data published in other institutional or subject-specific data repositories frequently described in related journal articles or data papers (see below). They are intended to permit data citations to be treated as "first class" citation objects on a par with bibliographic citations, and to enable them to be more easily harvested from reference lists, so that those who have made the effort to publish their research data might more easily be ascribed academic credit for their work through the normal mechanisms of citation recognition.

For such data in data repositories, each published data package and each published data file should always be associated with a persistent unique identifier. A Digital Object Identifier (DOI) issued by DataCite, or CrossRef, should be used wherever possible. If this is not possible, the identifier should be one issued by the data repository or database, and should be in the form of a persistent and resolvable URL. As an example, the use of DOIs in the Dryad Data Repository is explained on the Dryad wiki.

Data citations may relate either to the author's own data, or to data created and published by others ("third-party data"). In the former case, the dataset may have been previously published, or may be published for the first time in association with the article that is now citing it. All these types of data should, for consistency, be cited in the same manner.

Best practice recommendations

As is the norm when citing another research article, any citation of a data publication, including a citation of one's own data, should always have two components:

- An **in-text citation statement** containing an **in-text reference pointer** that directs the reader to a formal data reference in the paper's reference list.
- A formal **data reference** within the article's reference list.

We recommend that the in-text citation statement also contains a separate citation of the research article in which the data were first described, if such an article exists, with its own in-text reference pointer to a formal article reference in the paper's reference list, unless the paper being authored is the one providing that first description of the data. If the in-text citation statement includes the DOI for the data (a strongly desirable practice), this DOI should always be presented as a dereferenceable URI, as shown below. Further to this, both DataCite and CrossRef recommend displaying DOIs within references as full URLs, which serve a similar function as a journal volume, issue and page number do for a printed article, and also give the combined advantages of linked access and the assurance of persistence (Edmunds et al. 2012, Ball and Duke 2015).

For example, Dryad recommends to cite always both the article in association with which data were published and the data themselves (Fig. 1).

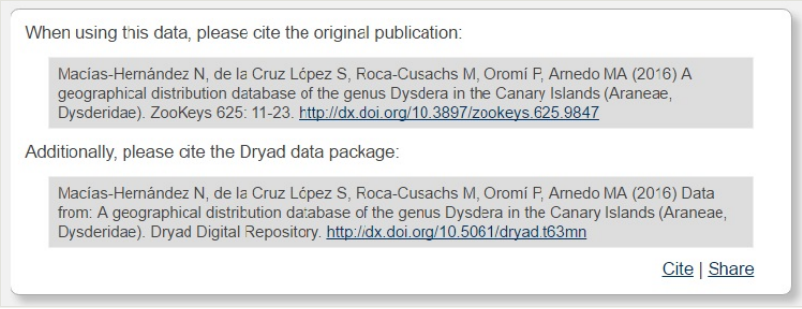

Figure 1.

Recommendation of Dryad to cite both the original article in association with which the data were published and the data themselves.

The data reference in the article's reference list should contain the minimal components recommended by the FORCE11 Data Citation Synthesis Group (Martone, M (Ed.) 2014) and corresponding to the data citation principles 2 (Attribution and credit), 4 (Unique Identifier (e.g., DOI, Handle), 5 (Access to humans and machines), 6 (Persistence) and 7 (Version and granularity):

- Author(s)
- Year
- Dataset Title
- Data Repository or Archive
- Global Persistent Identifier
- Version, or Subset, and/or Access Date

These components should be presented in whatever format and punctuation style the journal specifies for its references.

The following example demonstrates in general terms what is required.

In-text citation:

“This paper uses data from the [*name*] data repository at https://doi.org/***** (Jones *et al.* 2008a), first described in Jones *et al.* 2008b. “

Data reference and article reference in reference list:

Jones A, Bloggs B, Smith C (2008a).