

# Putting Principles in Practice: Can a Publisher Implement Data Citation?

Anita de Waard

VP Research Data Collaborations

[a.dewaard@elsevier.com](mailto:a.dewaard@elsevier.com)



*Elsevier*

Research Data Services

<http://researchdata.elsevier.com/>

# Can a publisher bring the Data Citation Principles into practice?

1. 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.
2. 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.
3. Evidence: Where a specific claim rests upon data, the corresponding data citation should be provided.
4. Unique Identification: A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. Access: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. Persistence: Metadata describing the data, and unique identifiers should persist, even beyond the lifespan of the data they describe.
7. Versioning and granularity: Data citations should facilitate identification and access to different versions and/or subsets of data. Citations should include sufficient detail to verifiably link the citing work to the portion and version of data cited.
8. 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.

1. 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.

BDDIC working draft #2 x (1 unread) - antawaard x Susanna-Assunta Sanson x Pilot\_Task Force List - G x Ordered list of Themes E x Task Forces Master List x Spectral modelling near t x

www.sciencedirect.com/science/article/pii/S1386142513009098

Download PDF Export More options... Search ScienceDirect Advanced search

The estimation and prediction of atmospheric CO<sub>2</sub> values and associated global climate change require a good understanding of global CO<sub>2</sub> concentrations [4] and [5]. In-situ CO<sub>2</sub> measurements and sampling

Administration) carbon cycle GHG cooperative air sampling network<sup>1</sup>, CDIAC (Carbon Dioxide Information Analysis Center). From these networks it was found that the annual average concentrations of CO<sub>2</sub> rose from 315.98 ppmv in 1959 to 385.34 ppmv in 2008 giving an annual average growth rate of 1.4 ppmv per year [1]. These fixed measurements are complemented with ship and aircraft observations. A majority of these observatories are over US and European land surfaces and a minority in the southern

Satellite measurements have the potential to significantly reduce these shortcomings. Space based instruments can provide data, which are essential for estimation of the GHGs and their variability study, over a vast region on continuous spatial and temporal intervals. There are a few satellite based

<http://www.sciencedirect.com/science/article/pii/S1386142513009098>

[12] and [13], and the Infrared Atmospheric Sounding Interferometer (IASI) [14]. These instruments perform CO<sub>2</sub> sensitive measurements in the thermal infrared (TIR) spectral region.

The information from TIR nadir measurements is limited to upper and middle tropospheric CO<sub>2</sub>, while the near-infrared (NIR) nadir measurements are sensitive over all altitudes [15]. In order to get information about regional CO<sub>2</sub> source and sinks it is important to get sensitive measurements near the earth's surface. Therefore, to retrieve total CO<sub>2</sub> columns NIR measurement is necessary. In this regard the satellite instruments that measure reflected solar radiation covering the important absorption bands of CO<sub>2</sub> in the NIR spectral region gain significance.

SCIAMACHY (SCanning Imaging Absorption spectroMeter for Atmospheric CHartography) [16] onboard ENVISAT, launched in 2002 [17] was the first CO<sub>2</sub> dedicated space mission that gave data for almost a decade.<sup>2</sup> It covered non-continuously the spectral range of 0.24–2.38 μm. There were eight channels and

2. 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.

Download PDF Export citation Jump to references More options...

Search ScienceDirect Search

Physics Reports <http://dx.doi.org/10.1016/j.physrep.2004.07.002>

## The Durham HepData Project

REACTION DATABASE • DATA REVIEWS • PARTON DISTRIBUTION FUNCTION SERVER • OTHER HEP RESOURCES

### Reaction Database Full Record Display

View [short record](#) or as: [plain text](#), [AIDA](#), [PyROOT](#), [YODA](#), [ROOT](#), [mpl](#) or [ScaVis](#)

#### ACHARD 2004 — Studies of hadronic event structure in $e^+ e^-$ annihilation from 30-GeV to 209-GeV with the L3 detector

Experiment: [CERN-LEP-L3 \(L3\)](#)  
Published in [PREP. 399,71](#) (DOI:10.1016/j.physrep.2004.07.002)  
Preprinted as [CERN-PH-EP/2004-024](#)  
Record in: [INSPIRE](#)

CERN-LEP. Comprehensive study of hadronic event shapes and distributions in  $E^+ E^-$  interactions from collision energies from 91 to 209 GeV. These data update and supersede many of the L3 results published previously.. This section contains the 2,3,4 and 5 jet fractions for the JADE, Durham(KT) and Cambridge algorithms as a function of their respective jet resolution parameters (YCUT). This section contains the distributions of the event shape variables THRUST, Heavy Jet

<http://www.sciencedirect.com/science/article/pii/S0370157304002753>

### 3. Evidence: Where a specific claim rests upon data, the corresponding data citation should be provided.

## *Presenting Supplementary Material at the relevant location*

Download PDF Export citation Jump to references More options...

Search ScienceDirect Search

☒ Show thumbnails in outline

Table 1

4. Results

4.1. Taxonomy

4.2. Taphonomy

reptiles (Squamata, mainly snakes and lizards), bony fish (Osteichthyes) and birds. The data in Table 1 shows that numbers of specimens are generally low and widely distributed among the samples indicating dispersion rather than concentration of the remains. Using Green's coefficient of dispersion (see Krebs, 1989: Equation (4.25)) shows values of 0.09 in Level Q-4 and 0.02 in Level Q-5 which indicate random rather than aggregated or uniform distribution of the remains among the samples.

Inline Supplementary Table S1

Table S1. Counts of specimens by taxonomic categories.

| Locus | Basket/level | Bird | Fish |
|-------|--------------|------|------|
| 106   | 3            | 1    | 27   |
| 112   | 1            | —    | 5    |
| 157   | 1            | —    | 49   |
|       | 2            | 2    | 80   |
| 33    | 15           | —    | 7    |
| 46    | 16           | 2    | 3    |
|       | 17           | 13   | 21   |
| 55    | 6            | 6    | 19   |

http://dx.doi.org/10.1016/j.jas.2012.07.001

Get rights and content

Bibliographic information

Citing and recommended articles

Recommended articles

Design choices in imaging speech compr...  
2012, NeuroImage

Show more information

- Supplementary material inserted at the place of reference/citation
- Put material into the right context
- Make it easier for readers to find
- Initially in closed text-box, action to open

# *Small side note:*

## *Taking evidence a step further:*

### Cortex Registered Report:

- Two-step submission process:
  - Method and proposed analysis are submitted for pre-registration
  - Paper is conditionally accepted
  - Research is executed
  - Full paper submitted, accepted provided that protocol is followed
- All experimental data made available Open Access

Featured in The Guardian “Confronting the 'sloppiness' that pervades science”, <http://bit.ly/1aUAY7f>

4. Unique Identification: A data citation should include a persistent method for identification that is **machine actionable**, globally unique, and widely used by a community.

## Interlinking Articles and Data through accession numbers

*Enabling one-click access to relevant primary data*

### 2. Methods

#### 2.1. The isolate *P. pentosaceus* and dextranucrase production

The isolate *P. pentosaceus* (Genbank Accession Number [EU569832](#)) was screened from the soil sample collected from a sugarcane field of Assam (near Guwahati), a well-known hotspot for biodiversity (Thakur et al., 2007). The isolate was maintained in modified MRS agar medium (Goyal and Katiyar, 1996) at 4 °C and subcultured every 2 weeks. The protocol for production and purification of dextranucrase from the isolate *P. pentosaceus* was followed as described by Purama and Goyal (2008).

- Author-tagged
- Captured in article XML
- Linked to data repository from the

[http://public.lanl.gov/herbertv/papers/Papers/2014/IDCC2014\\_vandesompel.pdf](http://public.lanl.gov/herbertv/papers/Papers/2014/IDCC2014_vandesompel.pdf)

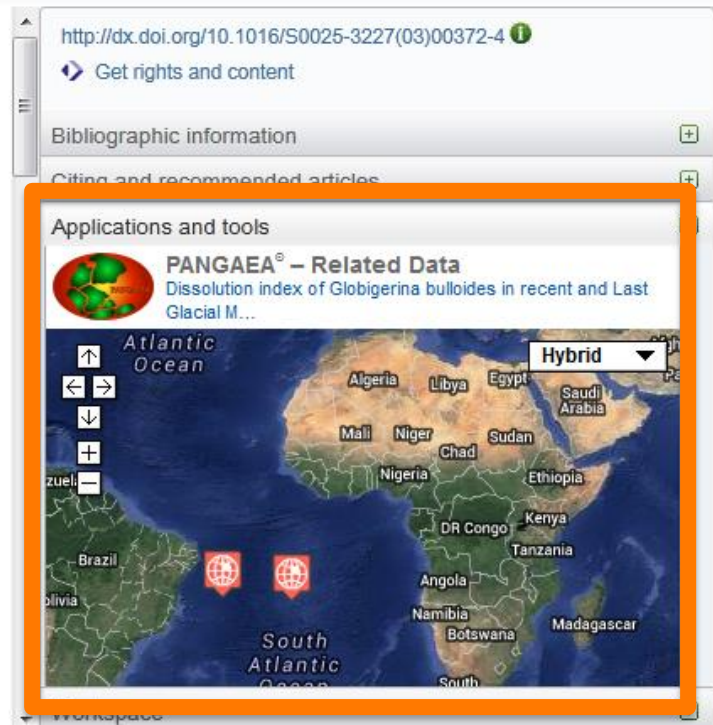
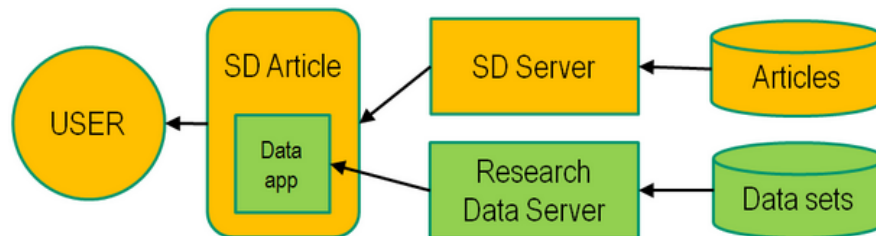
article on ScienceDirect

See <http://www.elsevier.com/databaselinking>

5. Access: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

*Integrating (meta)data into the article page view*

- Supplementary data at PANGAEA
- Bidirectional links between PANGAEA <> ScienceDirect
- Data visualized next to the article



See <http://www.elsevier.com/databaselinking>



6. Persistence: Metadata describing the data, and unique identifiers should persist, even beyond the lifespan of the data they describe.

### Requirements for the PID/HTTP Bridge

- **Support for resource versioning**, discovery of versions, access to versions to reflect that resources used or created during the research process are increasingly dynamic

#### FORCE11 Data Citation Principles:

*(7) **Specificity and Verifiability:** ... Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, **version** and/or granular portion of data retrieved subsequently is the same as was originally cited.*



7. Versioning and granularity: Data citations should facilitate identification and access to different versions and/or subsets of data. Citations should include sufficient detail to verifiably link the citing work to the portion and version of data cited.

- Discussion with Dave DeRoure:
  - How do you reference a [Research Object](#)?
  - Is that a good way to describe an experiment?
  - (Should we start a Force11 WG on it?)
- Discussion with David Rosenthal:
  - Are DOIs really the best identifiers for datasets?
  - Perhaps URI's (that can have a hierarchical structure, cf. DNS) are a better identifier mechanism?
- Requirement from Herbert van den Sompel: [make it machine-actionable](#)!
- Question to you all:
  - What is a dataset? (Cf. David Minor: what is an object?)

8. 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.


### How data and articles are linked

There are several ways in which we support interlinking of articles and data:

- **Referencing data in your article through tagging identifiers or accession numbers:** If your article contains relevant unique identifiers or accession numbers linking to information on genes, proteins, diseases, etc. or structures deposited in public databases, and you would like your article to link to that data, please identify these entities in the following way:

*database abbreviation: data identifier*

For example, "*PDB: 1TUP*" to identify the protein with accession number "*1TUP*" in the Protein Data Bank (PDB). Please bear in mind that an error in a letter or number will result in a dead link in the article. Database abbreviations and further examples can be found in the listing of [supported databases](#).

- **Data DOI's:** Elsevier supports [Data DOI's](#)  as persistent identifiers for scientific data. If you include a data DOI in your article, it will automatically turn into a link to your data on ScienceDirect.
- **Linked data repository banners on ScienceDirect:** Elsevier collaborates with selected data repositories to show banner links next to relevant articles on ScienceDirect. This linking system requires that the data repository maintains accurate records of associations between articles and data sets. What you need to do as an author to support this type of linking depends on the data repository; see links to more information in the [supported databases](#) section.
- **Data visualization and integration applications:** In close collaboration with selected data repositories, Elsevier has developed a number of data-integration and visualization applications that are shown next to the article on ScienceDirect, e.g. the [Protein Viewer](#) (with PDB), the [PANGAEA](#) data visualization tool, and the [Genome Viewer](#) (with NCBI). These applications build further on tagged entities or banner links to visualize data and integrate it into the online reading experience.

<http://www.elsevier.com/about/content-innovation/database-linking>

# Can a publisher bring the Data Citation Principles into practice?

1. 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.
2. 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.
3. Evidence: Where a specific claim rests upon data, the corresponding data citation should be provided.
4. Unique Identification: A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. Access: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. Persistence: Metadata describing the data, and unique identifiers should persist, even beyond the lifespan of the data they describe.
7. Versioning and granularity: Data citations should facilitate identification and access to different versions and/or subsets of data. Citations should include sufficient detail to verifiably link the citing work to the portion and version of data cited.
8. 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.