GBIF
www.gbif.org

Recommended practices for citation of data published through the GBIF network

Version 1.0

May 2012

Suggested citation

GBIF (2012). Recommended practices for citation of the data published through the GBIF Network. Version 1.0 (Authored by Vishwas Chavan), Copenhagen: Global Biodiversity Information Facility. Pp.12, ISBN: 87-92020-36-4. Accessible at http://links.gbif.org/gbif_best_practice_data_citation_en_v1

ISBN: 87-92020-36-4
Persistent URI: http://links.gbif.org/gbif_best_practice_data_citation_en_v1
Language: English

Document Control:

Version	Description	Date of release	Author(s)
1.0	Review, edits and final styling	11 May 2012	Vishwas Chavan

About GBIF

The Global Biodiversity Information Facility (GBIF) was established as a global mega-science initiative to address one of the great challenges of the 21st century – harnessing knowledge of the Earth's biological diversity. GBIF envisions 'a world in which biodiversity information is freely and universally available for science, society, and a sustainable future'. GBIF's mission is to be the foremost global resource for biodiversity information, and engender smart solutions for environmental and human well-being[1]. To achieve this mission, GBIF encourages a wide variety of data publishers across the globe to discover and publish data through its network.

[1] GBIF (2011). GBIF Strategic Plan 2012-16: Seizing the future. Copenhagen: Global Biodiversity Information Facility. 7pp. ISBN: 87-92020-18-6. Accessible at http://links.gbif.org/sp2012_2016.pdf

Table of Contents

Executive summary

Currently, we lack consistency in data citation. It is difficult or impossible, given existing citation practices, to identify who originally created or added value to a datum. This issue is further compounded when aggregator and information systems or networks facilitate discovery and access to the data federated through distributed publishers.

The GBIF Data Publishing Framework Task Group established in 2009, recommended that GBIF institutionalize a 'data citation mechanism' and establish a 'data citation service' facilitating deep data citation, and registration and resolving of citations (Moritz et.al, 2011).

As an early uptake of this recommendation, GBIF in consultation with a group of experts has come up with recommended practices for citing biodiversity data. This document recommends a set of styles for (a) Publisher-based citations, and (b) Query-based citations. *The recommended sets of styles for publisher-based citations are for immediate uptake by data publishers, data owners, data custodians, and data aggregators.*

A query-based citation requires implementation of a data citation service. GBIF is exploring various options for implementing such a service, and this document does not describe the technical approaches involved. Progress on the data citation service will be reported in future GBIF communications and eventual publication in the form of a follow-up guide.

Until query-based citation is implemented, users of GBIF-mediated data are requested to use the recommended citation displayed when downloading data from the GBIF Data Portal.

Introduction

A lack of incentives is one of several recognized impediments to increased publishing of biodiversity data. In order to address this, and to offer solutions, GBIF constituted the Data Publishing Framework Task Group (DPF TG) in 2009. The group recommended three mechanisms, viz. the Data Paper, deep data citation, and data usage index (Moritz et.al, 2011). Regarding data citation, it recommended that GBIF introduce a data citation mechanism and establish a data citation service facilitating deep-data citation, and registration and resolving of citations.

As an early uptake of this recommendation, GBIF in consultation with group of experts has come up with recommended practices for citing biodiversity data. This document recommends set of styles for (a) Publisher-based citations, and (b) Query-based citations. *The recommended sets of styles for publisher-based citations are for immediate uptake by data publishers, data owners, data custodians, and data aggregators.*

A query-based citation requires implementation of a 'data citation service'. GBIF is currently exploring various options to implement such a service, and the proposed technical approach is not described in this document.

Styles for data citation: Why?

A data citation that can recognize the efforts of all relevant players in the data life cycle is essential. While emphasizing the urgency of well-structured data citations, GBIF DPF TG commented that for the purposes of accountability and citations (attribution), all contributors of data to any aggregation should be identified and acknowledged. Individuals or institutions responsible for primary data have an obligation to make these ownership statements available to the aggregators, and the aggregators have an obligation to use them. In any data aggregation chain the aggregator at each level is responsible for identifying of data sources from previous levels of aggregation, and of the contributors involved (Moritz et.al. 2011).

However, Roberts and Chavan (2008) observe that it is difficult or impossible, given existing citation practices, to identify who originally created or added value to a datum. For data to be citable, it is necessary that they can be referred to in a consistent way (Klump et. al., 2006). However, citing digital data poses challenges that do not apply to citing printed publications. Unlike print publications where the article is frozen in time once published, digital data can continue to be edited, altered and updated. More than one publisher, aggregator, information system or network can publish the same data record or dataset.

Evolution of the data citation styles

These and several other challenges related to digital data citation in general, and that of the biodiversity data accessible through GBIF network in particular, have been investigated in consultation with several professional societies (DataCite and CODATA) and research groups interested in data citation.

Considering the complexity of the GBIF network and data flow, it was felt that a two-pronged strategy should be adopted, involving (a) set of data citation styles for immediate uptake by data publishers and data users, and (b) implementing a data citation mechanism and services. This document deals only with the first part of the strategy.

Key considerations

In devising the proposed set of styles for data citation, the following aspects of the GBIF network and the nature of digital biodiversity data were taken into consideration.

1. Several individuals, from the creator/collector of the data to the publisher and aggregators, play vital roles in the data life cycle, and each needs to be adequately recognized, attributed or credited.
2. Datasets are created, maintained and published by individuals, groups, institutions, consortiums and information systems or networks.
3. A dataset can be created and published by the same entity, or it can be created by one entity but published by another entity.
4. Even when data are published, their quality will be updated or the quality of the record can be improved.
5. Each player involved in the data life cycle merits recognition for his or her role.
6. Datasets are often dynamic in nature, as they continue to undergo revision, alteration, editing, additions, deletion, annotation and other value additions.
7. The same data record or dataset can be accessible through more than one access point, such as through the data publisher, data aggregators, and information system or network and discovery infrastructures.
8. When data are accessed through aggregators, information system or network portals and discovery infrastructures, the resultant datasets are a combination of data records contributed by multiple data publishers through multiple data resources.
9. Data citation practices should keep the advantages of print citations, but be distinguishable from them. They should add other components enabled and required by the digital form and systematic nature of datasets, and they should be consistent with most approaches.
10. Data citation practices need to be flexible enough to accommodate deep citations, versioning and any amount of additional information of interest to archivers, producers, distributors, publishers or others without losing functionality.

Recommended styles for data citation

Two sets of styles for data citations are recommended:

- Publisher-based citations (for immediate uptake and use in metadata)
- Query-based citations (for future uptake, pending technical services being developed by GBIF)

Publisher-based citations:

Publisher-based citations are authored or described by the owner, custodian or publisher of the dataset. The unit for citation is the dataset. All those responsible for the dataset development should be individually cited, and the role played by each contributor can be recorded in the citation. Elements that form the citation practice are listed in Table 1.

Table 1. Elements that constitute publisher-based citation.

Element	Definition	Style to be used in citation
Publisher	Original publisher of the dataset. This can be an individual, group of individuals, institute or consortium	In the case of an individual: <Family Name Initials> In case of individuals: <individual 1>, <individual..n> In case of institute, consortium: <official name of the institute/consortium>
Year	Year when dataset was first published	YYYY e.g. 1990, 2010 etc.
(Year)	Dataset is published only one time	(YYYY) e.g. (1990) – dataset was published for the first and last time in the year 1990. This means dataset has remained static, without any changes.
(Year -)	Dataset has undergone changes since it was first published	(YYYY -) e.g. (2005 -) – dataset was published in 2005 for the first time, however since then has undergone changes, edits etc.
Title	Title of the dataset	<Official product name or title of the dataset> e.g. Birds of Northeast India, or Database of Bombay Natural History Society.
Number of records	Total number of records in the dataset. In the case of a dynamic dataset, record numbers could should be at the time of last release or last version	Numeric value e.g. 10023 records.
Modes of publishing	How dataset has been published, e.g. online, offline (CDROM, DVD,	Published <mode of publishing>, e.g. published online

	etc.)	
Primary access point	If dataset is published online then access point where it is originally published	<URI>, e.g. http://www.spandan-bip.org
Released on	Date on which dataset was first released	<released on DD/MM/YYYY>, e.g. released on 19/05/1988
Persistent identifier	Persistent identifier such as DoI, LSID, GUID, Handle that either resolve to one of the following – metadata, citation, Data Paper	<Persistent identifier>, e.g. doi:89.92020-20-91
Version no. or last updated or last released	In case of a dynamic dataset, mention either (a) current version number or (b) date of last update and/or date of last release	<ver. 2.2> or <last updated on DD/MM/YYYY> or last <released DD/MM/YYYY>
Contributor	Applicable in case of datasets owned by institutes and/or consortium. All individuals who were responsible for dataset with their role in paranthesis	<contributed by contributor1 (role), contributor2 (role)…., contributor n(role)>, e.g. contributed by Hill JP (data collector), Maklani DP (metadata author), Xu BM (digitizer)

Depending on the permutation and combination of publisher(s) and the nature of release, a dataset can be cited using one of the six styles as listed in Table 2.

Table 2. Styles for publisher-based citations.

	Complete formulation	Short formulation
Style 1	**Publisher (individual) with one-time release of dataset**	
	Publisher (YEAR), <Title of data resource>, <total no. of records>, published <modes of publishing>, <Primary access point>, released on<release date>, <Persistent Identifier>.	Publisher (YEAR), <Persistent Identifier>.
Style 2	**Publisher (individual) with frequent update or release of dataset**	
	Publisher (YEAR). <Title of data resource>, <total no. of records>, published <modes of publishing>, <Primary access point>, first released on<release date>, <current version no. or last updated/released on (date)>, <Persistent Identifier>.	Publisher (Year first published/released -). <Version no., or last updated/released on (date)>, Persistent Identifier.
Style 3	**Publisher (group of individuals) with one time release of dataset**	
	Publisher 1, ….. and Publisher n (YEAR). <Title of data resource>, <total no. of records>, published <modes of publishing>, <Primary access point>, released on <release date>, <Persistent Identifier>.	Publisher 1 et.al. (YEAR). Persistent Identifier.
Style 4	**Publisher (group of individuals) with frequent update or release of dataset**	
	Publisher 1, ….. and Publisher n <YEAR). <Title of data resource>, <total no. of records>, published <modes of publishing>, <Primary access point>, first released on<release date>, <current version no. or last updated/released on (date)>, <Persistent Identifier>.	Publisher 1 et.al. <YEAR (Year first published/released -)>. <Version no., or last updated/released on (date)>, Persistent Identifier.
Style 5	**Institute/consortium (multiple contributors) with one time release of dataset**	
	<Publisher as Institution / Research Group /	<Publisher as Institution /

	Consortium> (YEAR), <Title of data resource>, <total no. of records>, <Contributed by contributor 1(role), contributor 2 (role)….. contributor n(role)>, <published (modes of publishing)>, <Primary access point>, released on<release date>, <Persistent Identifier>.	Research Group / Consortium> (YEAR), <Persistent Identifier>
Style 6	**Institute/consortium (multiple contributors) with frequent update or release of dataset.**	
	<Publisher as Institution / Research Group / Consortium> <YEAR (Year first published / released -)>, <Title of data resource>, <total no. of records>, <Contributed by contributor 1(role), contributor 2 (role)….. contributor n(role)>, <published (modes of publishing)>, <Primary access point>,<Version no., or last updated/released on (date)>, <Persistent Identifier>.	<Publisher as Institution / Research Group / Consortium> <YEAR (Year first published / released -)>, <Version no., or last updated/released on (date)>, <Persistent Identifier>

Table 3. Hypothetical examples of styles for publisher-based citations.

	Complete formulation	Short formulation
Style 1	**Publisher (individual) with one-time release of dataset**	
	Chavan, V. S. (1996). Amphibians of the west coast of India. 1223 records, published online, http://www.vishwaschavan.in/indfauna/amphibians_west_coast/, released on 12 June 1998, doi: 10.5284/1000164.	Chavan, V. S. (1996), doi: 10.5284/1000164.
Style 2	**Publisher (individual) with frequent update or release of dataset**	
	Johnson, D. K. (2002 -). Observational dataset of the mammals of South Africa, 32001 records, Online http://www.satol.ac.za/mammalsdb/, 01/10/ 2002, version 1.2 (last updated on 01/01/2012), doi: 10.1000/123.	Johnson, D. K. (2002-). Version 1.2 (last updated on 01/01/2012), doi: 10.1000/123.
Style 3	**Publisher (group of individuals) with one time release of dataset**	
	Patterson, D.H., Villadsen, M. B., Arino, M. P., Christensen, B. K., Chopde, D. D., Subramanian, Q. T., Khan, M. K. (2001). MEDUSA: Molluscs of the Indian Ocean, 31092 record, Online, http://www.incois.gov.in/MEDUSA/, released on 31/07/ 2001, doi: 10.incois/2000654.	Patterson et.al. (2001), doi: 10.incois/2000654.
Style 4	**Publisher (group of individuals) with frequent update or release of dataset**	
	Hauser, C.H., Berendsohn, D. P., Edwards, P. H., Ratnamsinghe, D. H. Q., Gaikwad, J. B., Ranganathan, S. D. (1999 -). Database on inventory of Australian traditional medicinal plants, 233678 records, Online, http://www.ala.gov.au/medicinal_plants/, released on 15/08/1999, version 2.2 (last updated on 21/01/2002), doi: 15.ala/3000678.	Hauser et. al. (1999 -), version 2.2 (last updated on 21/01/2002), doi: 15.ala/3000678.
Style 5	**Institute/consortium (multiple contributors) with one time release of dataset**	
	Bombay Natural History Society (2011). Literature based species occurrence data of birds of northeast India. 2400 records, Contributed by Narwade, S. (Principal Investigator, Content Provider, Metadata Provider), Karla, M. (Processor), Varier, D. (Custodian, Metadata Provider), Jagdish, R. (Processor), Satpute, S. (Custodian), Khan, N. (Custodian), Talukdar, G. (Publisher), Mathur, V. B. (Publisher), Vasudevan, K. (Publisher), Pundir, D. K.	Bombay Natural History Society (2011), doi: 10.3897/ibif.150.2011.

	(Publihser), Chavan, V. (Metadata Provider, Editor), and R. Sood (Programmer), Online, http://ibif.gov.in:8080/ipt/resource.do?rBNHS-NEW, released on 01/09/2011, doi: 10.3897/ibif.150.2011.	
Style 6	**Institute/consortium (multiple contributors) with frequent update or release of dataset.**	
	National Centre for Biodiversity Informatics (2001 -). Indfauna: electronic catalogue of known Indian fauna, 201097 records, Contributed by Narwade, S. (Principal Investigator, Content Provider, Metadata Provider), Karla, M. (Processor), Varier, D. (Custodian, Metadata Provider), Jagdish, R. (Processor), Satpute, S. (Custodian), Khan, N. (Custodian), Talukdar, G. (Publisher), Mathur, V. B. (Publisher), Vasudevan, K. (Publisher), Pundir, D. K. (Publihser), Chavan, V. (Metadata Provider, Editor), Wildlife Trust India (Content Provider), and R. Sood (Programmer), Online, http://www.ncbi.org.in/indfauna/, Version 3.0 (last updated on 01/10/ 2011), doi: 10.3897/ncbi.ncl.2001	National Center fro Biodiversity Informatics (2001 -), Version 3.0 (last updated on 01/10/2011), doi: 10.3897/ncbi.ncl.2001

Query-based citation:

Query-based citations are composite, cascading citations to be used to cite the data retrieved and/or downloaded from the access point of publishers, aggregators, information systems or networks or discovery platforms such as GBIF. When available, they will be used to cite the final dataset retrieved, as well as for analysis and interpretation purposes.

Query-based citation consists of two parts: (a) details of access points (access point name, URI, no. of records, no. of datasets, retrieval date and time); and (b) persistent identifiers of the publisher-driven citation of the datasets that contributed to the resultant (retrieved) dataset. Table 3 describes the key elements of the query-based citation (other than those described in Table 1).

Table 3. Elements used in query-based citation (other than those described in Table 1).

Element	Definition	Style to be used in citation
Access point URI	URI of the access point from where resultant dataset was accessed, retrieved and/or used	<uri of the access point>, e.g. http://data.gbif.org/
Search string	Final search string used to retrieve the data	<search string>, e.g. Panthera tigris and India
Accessed on	Date and time when data was retrieved	<accessed on DD/MM/YYYY at hh:mm:ss>, e.g accessed on 27/12/2002 at 23:58:02
No. of records searched, retrieved, used	No. of records searched retrieved and downloaded	<no. records>, e.g. 12000 records
No. of datasets	No. of datasets contributed to the resultant/retrieved data	<through no. of datasets>, e.g. 77 datasets.
Persistent identifier assigned to the resultant dataset	Persistent identifier (e.g. DoI or LSIFD, etc.) assigned to	<doi or ark or lsid> e.g. doi: 98.92029.20-87

	resultant data once the full length citation is registered	
Persistent identifier of contributing dataset (no. of contributed records)	Persistent identifier of the datasets that contributed to the search (no. of records each dataset contributed to the search)	ARK 88899-32 (3000 records), doi:88-92020-92-87 (507 records)…….n

Query-based citations can be documented (authored) using one of the two styles presented in Table 4. The choice of styles depends on whether a user dataset is retrieved from a single dataset or an access point that aggregates multiple datasets (e.g. GBIF Data Portal).

Table 4. Styles for query-based citations.

	Complete formulation	Short formulation
Style 1	**User-driven citation for subset from single dataset**	
	Access Point (YEAR), <search string>, <no.of records searched/retrieved/used><accessed on (date/time)>, <Persistent identifier assigned by User to the resultant data> through <persistent identifier of the source dataset>	Access point (YEAR), <Persistent Identifier of assigned by the user to the resultant dataset>.
Style 2	**User-driven citation for resultant data contributed by multiple datasets**	
	Access Point (YEAR), <search string>, <no.of records searched/retrieved/used><accessed on (date/time)>, <Persistent identifier assigned by User to the resultant data> through <persistent identifier of the source dataset 1> (no. of records contributed), <persistent identifier of the source dataset 2> (no. of records contributed),………., <persistent identifier of the source dataset n> (no. of records contributed),	Access point (YEAR), <Persistent Identifier of assigned by the user to the resultant dataset>.

Table 4. Hypothetical examples of styles for query-based citations.

	Complete formulation	Short formulation
Style 1	**User-driven citation for subset from singe dataset**	
	http://www.ncbi.org.in/indfauna/ (2012), Hornbills and India, 989 records, accessed on 12 January 2012: 22:10:10 hrs, user doi: 99.6672/100.324.2012, publisher doi: doi: 10.3897/ncbi.ncl.2001.	http://www.ncbi.org.in/indfauna/ (2012), doi: 10.3897/ncbi.ncl.2001.
Style 2	**User-driven citation for resultant data contributed by multiple datasets**	
	http://data.gbif.org (2012), Amphibians and World, 120790 records, accessed on 12 January 2012: 22:10:10 hrs, user doi: 99.6672/100.324.2012, contributed by 14 datasets, publisher identifiers: 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (1000 records), 10.3897/ncbi.ncl.2001 (22000 records),	http://data.gbif.org (2012), doi: 99.6672/100.324.2012.

	10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records), 10.3897/ncbi.ncl.2001 (12000 records)	

Uptake of Publisher-based citation styles

The recommended sets of styles for publisher-based citations are for immediate uptake by data publishers, data owners, data custodians, and data aggregators.

The GBIF Metadata Profile (as implemented through IPT 2.0.2 and DwC-archive) has elements called ‘citation’ and ‘persistent identifier for citation’.

How publishers can generate a citation
1. Choose one of the six recommended styles (see Table 2). *2. Write the descriptive citation text in the ‘citation’ element.* *3. At this stage, the citation text becomes part of the metadata document.* *4. The persistent identifier of the dataset (as assigned by the GBIF Registry) is linked with the citation text.*

Uptake of query-based citation styles:

Query-based citations will eventually be generated automatically through a cascading combination of publisher-based citations of datasets that contributed to the retrieved data. However, the majority of the GBIF-mediated datasets do not currently have enriched and complete publisher-based citations. Further, technical modifications are required in the GBIF portal to construct such query-based citations through a citation service. This function is currently in scoping stage, and the service is expected to be available in due course.

References

1. Klump et.al., 2006. Data Publication in the Open Access initiative. Data Science Journal, 5: 79-83.
2. Moritz T, Kirshnan S, Roberts D, Ingwersen P, Agosti D, Penev L, Cockerill M, and Chavan V. Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. BMC Bioinformatics 2011, 12 (Suppl 15) S1, doi: 10.1186/1471-2105-12-S15-S1.
3. Roberts D and Chavan V., 2008. Standard identifier can mobiles data and free time. Nature, 453: 449-450.