Measuring the impact of research data in 4 points

1. Understanding the concept of the impact of research data

The data originating from scientific work, or **research data** (see CoopIST sheet Online initiation to research data and their management (PDF)), are products of research, much like publications or patents.

Scientific data can be re-used in order to conduct new research, or to compile and build larger data collections. In any case, it is useful to monitor the use of these data and to assess their impact over time in terms of benefits to scientists, policymakers, economic actors and, more broadly, civil society.

Impact measurement requires prior specification of the assessed elements, measurement criteria, and measurement indicators and tools. The calculated indicators have to be interpreted carefully, by taking the limitations of these choices into account (see CoopIST section Evaluating publications (in French)).

Datasets that are collected from the field, from the laboratory or during surveys can be subject to impact measurements when they enter into a **publication process**.

The publication process can be a traditional one (scientific journal):

- case of a dataset inserted into an article in the form of tables or graphs, or attached as an annexure to the article in a dedicated section (Supplementary Material), or mentioned in the article with a link to a site external to the journal where the dataset can be accessed;
- case of a dataset whose description is the topic of a special article, or **data paper** (see CoopIST sheet Writing and publishing a data paper in a scientific journal (in French)).

The publication process may be unique:

- case of a **database** made publicly available on a website;
- case of a **textual data file** (an Excel spreadsheet, for example) stored and accessible at a data repository.

In all cases, the dataset has to fulfil certain conditions in order for its impact to be measurable. It has to be:

- **accessible** so that it can be viewed, downloaded, used, or obtained on request from the author;
- **described precisely** so that it can be interpreted and used. Recommendations such as those of the DataCite international consortium should be followed to describe scientific data;

- **identified unequivocally** so that it can be cited, via the allocation of a permanent digital identifier such as the DOI (Digital Object Identifier) (see CoopIST sheet Citing a scientific dataset (in French));
- **clearly localized** so that it can found in cyberspace, for example.

2. Measuring the impact of data through the number of citations received

An indicator of a dataset's impact can be the **number of citations** it receives from scientific publications or other datasets. The fact that a dataset can be referenced and cited in a publication makes it easier to locate it and also helps quantitative publications (bibliometric tools) take it into account.

When the dataset has been the subject of a data paper published in a traditional or specialized journal (data journal), the measurement of its impact through the number of citations is identical to that of a traditional publication (see CoopIST sheet Main indicators of impact associated with scientific publications (in French)).

However, the practice of citing a dataset has not yet been systematically and consistently adopted by scientific journals and their commercial publishers. This can make it difficult to locate datasets associated with a publication.

3. Measuring the impact of data in cyberspace

A dataset's impact on the internet can be assessed by measuring the actions and interactions that this dataset generates on social networks, blogs and microblogs, and in digital media. These quantitative measurements of social activity on the Internet are called **altmetrics** (see CoopIST sheet Becoming familiar with altmetrics, alternative impact measurements of a publication (in French)). They complement traditional impact measurements such as the number of citations received.

These alternative measurements are based on a wide range of sources and actions (accessing, viewing, downloading, sharing, commenting, etc.) and offer varied and immediate indicators. Examples:

- number of times the page describing the dataset has been viewed;
- number of online accesses to the dataset file or the number of downloads of the file;
- number of times the dataset has been bookmarked on shared-bookmarks sites;
- number of times the dataset was referenced in a personal online bibliographic database of a researcher (for example, with the Mendeley site and software);
- number of online comments generated by the dataset on social media, blogs and microblogging sites, and in the online press.

4. Discovering tools for measuring the impact of data

The measurements associated with datasets have been inspired by those for scientific publications. Impact measurement models specific to scientific data are being explored by database producers and by information providers.

The following examples give an idea of the different types of information sources and the main indicators in use today.

Data Citation Index

Data Citation Index (DCI, Clarivate Analytics) is a paid database that indexes a selection of online data repositories, datasets and data from studies. DCI provides two types of impact indicators:

- **Usage Count:** an indicator showing the number of times the reference to a collection of data (Repository, Data Set or Data Study) was saved or exported from the DCI database, or the link to the source data was clicked by the user;
- **Times Cited:** number of times a database referenced in the DCI database was cited by other publications or other datasets in one of the Clarivate Analytics databases (Web of Science, Biosis, Medline, etc.).

Google Scholar

The specialized Google Scholar search engine (http://scholar.google.com/) indexes scientific literature on the internet. After a search on title, author name, journal title, etc., Google Scholar displays the number of citations received by the publication concerned:

- **Cited by**: a search on the metadata (title, author, year of publication, source) of a data paper allows, if Google Scholar has indexed it as a journal article, to view the number of times the article was cited and to access each of the identified and listed citations.

DataCite Statistics

The international **DataCite** consortium (http://www.datacite.org) facilitates access to and reuse of research data. Each member of the consortium is allowed to allocate permanent digital IDs of type DOI to digital resources such as datasets, databases, software, images, maps, etc.

DataCite Statistics (https://stats.datacite.org/) provides statistics related to the online use of digital resources, such as datasets, to which DataCite has assigned permanent digital IDs of type DOI. The type and number of uses on the Internet of these resources can be viewed by Allocator (DataCite member who has allocated DOIs), by Datacentre (the centre hosting the data with a DataCite DOI), and by Prefix (unique digital address or prefix assigned by DataCite to a data producer). This prefix, attached to the identifier of the digital resource, provides access to the resource.

- The **Registrations statistics** refer to datasets having a DataCite DOI which have been downloaded online by a user.
- The **Resolutions statistics** indicate how often a DOI has been used to access the associated digital resource.

Public scientific-data repositories

Data repositories offer researchers the possibility of submitting their data files online so that they can be accessed publicly on the Internet (see CoopIST sheet Making your scientific datasets public (in French)). The process of submitting files and entering metadata about the file's contents are based on the same principle as for publication in an open archive (see CoopIST sheet Submitting your publications to an open archive (in French)).

Some repositories have functions for sharing files over the Internet and for citation by users, and thus display related indicators. This is the case of the multidisciplinary **Figshare** repository (https://figshare.com/ – United Kingdom), which offers links to *Start the discussion*, as well as to *Share* the reference of a data file on social networks and to *Cite* it. Indicators are displayed on the description page of the stored data:

- **Views:** number of times the resource was viewed on Figshare;
- **Mentioned by:** this indicator is represented by a circle displaying the number of tweets, mentions in the media, on Facebook, Twitter, LinkedIn, Google+, etc.;
- **Downloads:** number of times the data file has been downloaded via Figshare;
- **Citations:** number of times the reference to the dataset has been cited.

Other data repositories provide more specific impact indicators. This is the case of the **Global Biodiversity Information Facility** (GBIF) (http://www.gbif.org/), which provides an opportunity for participating institutions from all around the world to submit data on biodiversity (animal and plant species) to a common repository and, in this way, make it publicly available. The GBIF repository provides usage statistics and citation metrics with descriptive results in line with the interests and activities of the GBIF community, such as:

- **GBIF Used, GBIF Cited, GBIF Discussed, GBIF Acknowledged, GBIF Mentioned** (on the GBIF website, see Resources section: subsection Relevance)
- **Annual number of peer-reviewed articles using GBIF-mediated data** (https://www.gbif.org/document/82873/gbif-science-review-2016)
- **Number of countries** with authors who used GBIF-mediated data in peer-reviewed papers (https://www.gbif.org/document/82873/gbif-science-review-2016).

Useful links

Ball A., Duke M. 2015. How to Track the Impact of Research Data with Metrics. DCC How-to Guide. Edinburgh: Digital Curation Centre (DCC), 16 p.
http://www.dcc.ac.uk/resources/how-guides/track-data-impact-metrics

Costas R., Meijer I., Zahedi Z., Wouters P. 2013. The Value of Research Data: Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report. Leiden, The Netherlands: CWTS. 46 p. http://www.knowledge-exchange.info/event/value-reseach-data-metrics

Global Biodiversity Information Facility [GBIF]. 2016. GBIF Science Review 2016: Compiling the year's research uses of data accessed through the Global Biodiversity Information Facility. Copenhagen (Denmark): GBIF. 40 p. https://www.gbif.org/document/82873/gbif-science-review-2016

Ingwersen P., Chavan V. 2011. Indicators for the Data Usage index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. BMC Bioinformatics, vol. 12, Suppl 15. 10 p. doi:10.1186/1471-2105-12-S15-S3
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S15-S3

Kratz J. E., Strasser C. 2015. Comment: Making data count. Scientific Data. 5 p. doi:10.1038/sdata.2015.39. https://www.nature.com/articles/sdata201539

Marie-Claude Deboin
Scientific and Technical Information Service (DIST), Cirad
Translated from the French by Kim Agrawal
27 June 2017

Informations
To cite this document:
Deboin, M.C. 2017. Measuring the impact of research data in 4 points. Montpellier (FRA): CIRAD, 4 p.
http://coop-ist.cirad.fr/impact-donnees