Open Data Management in Agriculture and Nutrition

*This e-learning course is the result of a collaboration between **GODAN Action** partners, including **Wageningen Environmental Research** (WUR), **AgroKnow**, **AidData**, **the Food and Agriculture Organization of the United Nations** (FAO), **the Global Forum on Agricultural Research** (GFAR), and **the Institute of Development Studies** (IDS), **the Land Portal**, **the Open Data Institute** (ODI) and **the Technical Centre for Agriculture and Rural Cooperation** (CTA).*

GODAN Action is a three-year project UK's Department for International Development to enable data users, producers and intermediaries to engage effectively with open data and maximise its potential for impact in the agriculture and nutrition sectors. In particular we work to strengthen capacity, to promote common standards and best practice and to improve how we measure impact. [www.godan.info]

UNIT 2: USING OPEN DATA

LESSON 2.5: REFERENCING DATA

Photo by Neil Palmer (CIAT) licensed under CC BY-SA 2.0

Aims and learning outcomes

The sharing of scientific research is a long established process dating back to the 1660s with the creation of the Royal Society in England. The publication of journal papers allowed scientists to lay claim to their discoveries while sharing findings with others. Centuries later, our ability to share outputs now reaches far beyond just the publication, into the data collected as part of the research process. This lesson looks at how similar principles are now applied to both discover and reference scientific data.

After studying this lesson, you should be able to:

- explain the importance of citation
- list the key features that citation provides and explain each
- understand the value of adding persistent identifiers in the data exchange workflow
- identify existing good practices for the use of persistent identifiers.

Contents

List of figures

1. Introduction

The sharing of scientific research is a long established process dating back to the 1660s with the creation of the Royal Society in England. The publication of journal papers allowed scientists to lay claim to their discoveries while sharing findings with others. Many centuries later and the publication of scientific papers is not just about laying claim to findings but also used as a measure of performance.

The classic mode of distributing scientific results is through publication in professional journals. Professional journals are peer reviewed and heavily quality controlled, meaning that the amount of work necessary to be published is not insignificant.

The reward for such work is the inclusion in the 'citation index'. The index is used as a performance evaluation for scientists. This means that the effort of publication is rewarded both within the community as well as an aspect of career and promotion.

Citations thus serve a number of key purposes:

- claiming ownership
- providing impact
- consistent referencing
- ease of discovery.

While the first two are specifically benefits for the author only, the last two help with reuse. Just las at the end of this lesson, references help connect people to evidence or background material. References contain author names, place of publication and year of publication at a minimum. These factors help both indicate impact, but also ease discovery, allowing users to find the relevant issue of a journal and the page number of the article.

Taken together these four aspects mean that scientists are incentivised and rewarded to publish high-quality research. The citation provides an important trust mechanism that shows readers that you are a responsible author and have cited reputable work in your domain.

2. The Case for sharing scientific data

Publication of research papers is an important part of the scholarly method. More important than personal impact, the scholarly method is designed to advance the teaching, research and practice of a given scholarly or academic field of study.

Sharing scientific research is critical to advance teaching, research and practice. It allows scientists in a field of study to validate and build upon the research of others.

Validation and expansion of research findings may require access to scarce resources, such a telescopes or chemical compounds, however more recently many experiments have used data as an input. Thus there have been calls for data sharing as part of the scientific publishing process.

In principle, scientists are prepared to provide data, but the necessary extra work that they must perform in processing, context documentation and quality assurance for that data is often neither appreciated nor acknowledged (Brase *et al.*, 2015). Conversely evidence suggests that those who make data available received more citations than similar studies for which the data was not made available, suggesting that data can increase impact (Piwowar, 2013).

The importance of sharing research data has meant that many parties have now put in place policies and mandates that enforce the publication of research data.

Many of the biggest journals and publishers now have data sharing requirements that must be met. You can read more about some of these here:

- Science's Data Deposition Policy[1]; scroll down to the 'Data and Materials Availability after Publication' section
- Springer Nature's Data Sharing Policies[2]; as a specific journal example see *Nature*'s policy on Availability of data, materials, and methods[3]
- Wiley's Data Sharing Service
- American Geophysical Union's Publication Data Policy[4]
- Sage's Replication policy[5].

One problem with many of these policies is that the data is always accompanying a journal article. The data itself is not a first-class research object that can be referenced.

In 2004 the German National Library of Science and Technology (TIB) assigned its first Digital Object Identifier (DOI) to scientific data in order to make scientific datasets citable research outputs.

The DOI is a persistent identifier used to uniquely identify objects, standardised by the International Organisation for Standardisation (ISO). The DOI aims to provide a 'resolvable' identifier to a digital object, such as URL. Unlike an ISBN and ISRC, which are just identifiers, a DOI is actionable and interoperable.

[1] http://www.sciencemag.org/authors/science-editorial-policies

[2] http://www.springernature.com/gp/authors/research-data-policy/?countryChanged=true

[3] http://www.nature.com/authors/policies/availability.html

[4] https://authorservices.wiley.com/author-resources/Journal-Authors/licensing-open-access/open-access/data-sharing.html

[5] https://us.sagepub.com/en-us/nam/journal/big-data-society#ARTICLETYPES

3. Persistent data identifiers on the web

Actionable identifiers was a concept first outlined by Tim Berners-Lee as part of the 5-stars schema for linked open data.

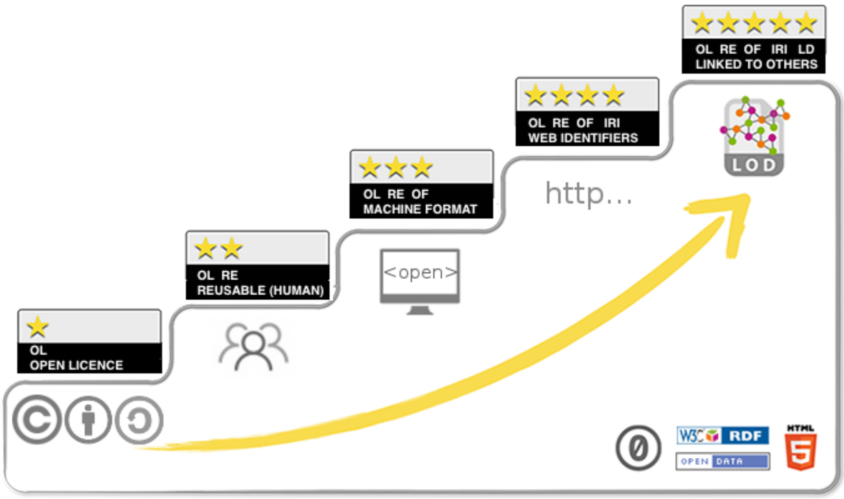

Figure 1 The 5-star schema for linked open data

This schema, outlines the process of making data open and linked on the web. As part of the schema the fourth step relates to identifiers. The general principal of the 4th step is to use web-based identifiers (URIs) to denote things. This way people can find out more information about your things and point to them through the use of a resolvable URI.

Unlike books, which use an ISBN as an identifier, a web-based identifier does not require a catalog through which that identifier can be resolved. However web-based identifiers can be created by anyone, on any domain, and thus do not hold the same level of authority.

The idea of the digital object identifier was to provide an authoritative and resolvable identifier for digital objects. Essentially a DOI is a persistent web-based identifier that points to the digital object on the web, regardless of its location. While anyone can put a digital object on the web, DOIs are only available for objects that have been approved as high quality. Essentially the DOI is attempting to replicate the same impact that a journal citation index does, just on the web (Data Citation Synthesis Group 2014).

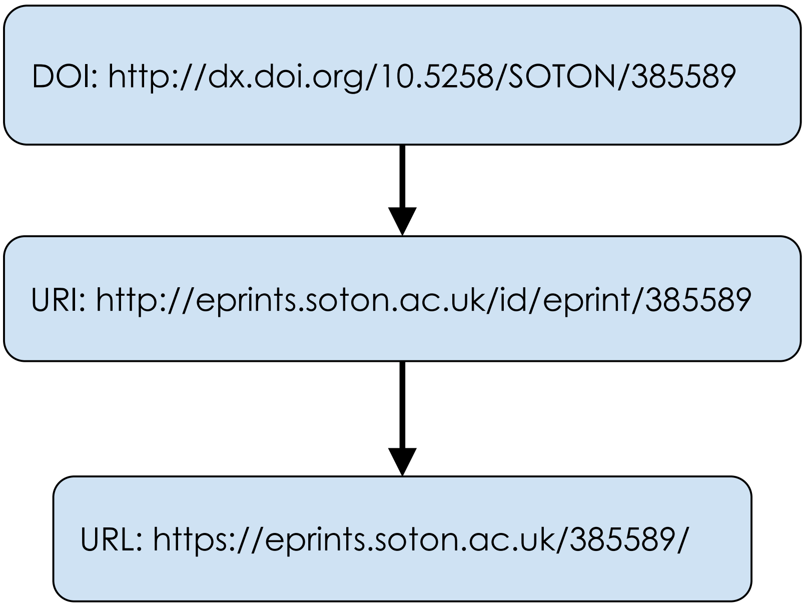
DOI: http://dx.doi.org/10.5258/SOTON/385589
URI: http://eprints.soton.ac.uk/id/eprint/385589
URL: https://eprints.soton.ac.uk/385589/

Figure 2 Example DOI for data

Figure 2 shows how a DOI references a dataset on the web. In this instance the dataset is available from an institutional repository. The repository provides both a unique identifier for the dataset (URI) as well as a human readable web page about the dataset (URL). The DOI is simply another identifier for the dataset which points to the actual URI. The purpose of the DOI is to provide a persistent and authoritative identifier for any digital object regardless of the location of the object on the web.

In addition to providing a resolution engine for these persistent identifiers, the DOI web service is also able to directly provide metadata about the digital object through content negotiation, as shown in Figure 3.

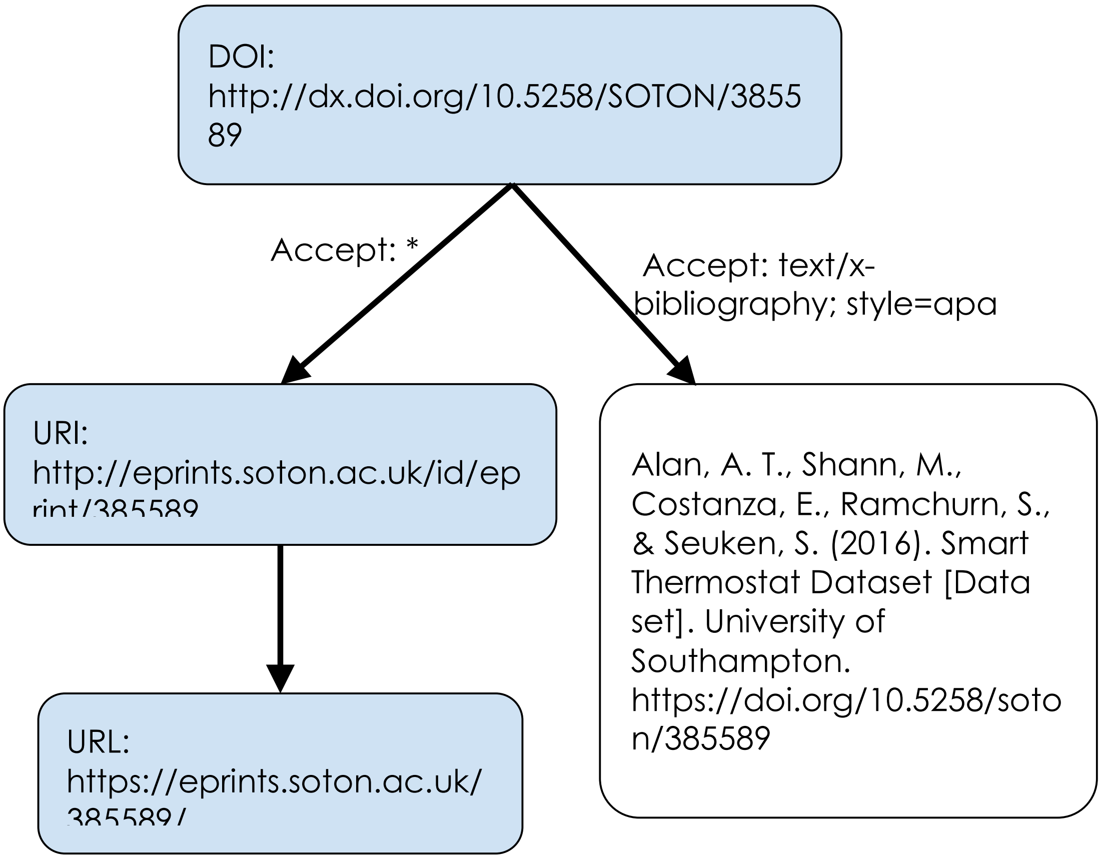
DOI: http://dx.doi.org/10.5258/SOTON/385589
Accept: *
Accept: text/x-bibliography; style=apa
URI: http://eprints.soton.ac.uk/id/eprint/385589
Alan, A. T., Shann, M., Costanza, E., Ramchurn, S., & Seuken, S. (2016). Smart Thermostat Dataset [Data set]. University of Southampton. https://doi.org/10.5258/soton/385589
URL: https://eprints.soton.ac.uk/385589/

Figure 3 Using a DOI to retrieve a bibliography record

3.1. DOI Providers

There are a number of ways in which a DOI could be linked to a dataset:

- Figshare[6] will provide a DOI for any deposited work, which includes data.
- Zenodo[7] also provides DOIs for any kind of research output, including datasets.
- Dryad[8] provides DOIs for data submissions linked to papers (for a fee, which includes storage, curation, archiving and checks for best-practice

There are also a number of domain-specific repositories providing a DOI for data. The journal Scientific Data[9] maintains a good list of repositories you can look at.

Choosing a DOI provider will depend on particular circumstances. Both Figshare and Zenodo are both free-to-use services for example, whereas Dryad charges for the service they offer to cover ongoing storage and archival costs.

[6] https://figshare.com
[7] https://zenodo.org
[8] http://datadryad.org
[9] https://www.nature.com/sdata/policies/repositories

Dryad accepts data relating to publications; if it is not associated with a paper then they will not accept it. Figshare and Zenodo will accept any research output, whether linked to publication or not. In that sense Figshare and Zenodo are more broadly applicable to any research outputs.

Data may be more discoverable in Dryad or domain-specific repositories (DSRs) than in general purpose ones likes Figshare and Zenodo. Data is likely to be formatted in standard ways and more easily searched by online or other tools if they are archived in Dryad or DSRs. This is likely to encourage reuse.

Figshare is a for-profit commercial entity, Zenodo is run by CERN and was supported by the EU OpenAIRE project at one point, whilst Dryad is a not-for-profit entity supported by research grants and membership fees for organisations.

4. Discussion

Books and journal articles have long benefited from an infrastructure that makes them easy to cite, a key element in the process of research and academic discourse. Datacite (the group behind the DOI) believes that digital objects should be citable in the same way.

Datacite DOIs are designed to:

- support proper attribution and credit
- support collaboration and reuse of data
- enable reproducibility of findings
- foster faster and more efficient research progress, and
- provide the means to share data with future researchers

For researchers the ability to publish and cite digital objects and first class objects offers incentives to publish high-quality, high-impact open data. Platforms such as Figshare make it increasingly easier to publish such data and create persistent data identifiers.

References

- Brase, J., Sens, I. and Lautenschlager, M. 2015. The tenth anniversary of assigning DOI names to scientific data and a five year history of DataCite. *D-Lib magazine* 21 (1/2). http://dx.doi.org/10.1045/january2015-brase
- Data Citation Synthesis Group (2014). Joint Declaration of Data Citation Principles. Martone M. (ed.), FORCE11, San Diego, CA, USA. Available at https://www.force11.org/group/joint-declaration-data-citation-principles-final
- Piwowar, H.A. and Vision, T.J. 2013. Data reuse and the open data citation advantage. *PeerJ* 1 (2013): e175.