Data Publishing and Data Citation - Are We There Yet?

Jens Klump | Science Leader Earth Science Informatics

7 December 2017

Why am I interested in data sharing?

I am not a sociologist.

- I am a geochemist, my field of research is Earth Science Informatics.
- Research data infrastructures have been part of my work since 1999.
- When I switched from marine geology to limnology I was puzzled by the difference in attitudes towards data sharing.
- Over the years I made some observations in this respect.

Why do communities behave so differently?

Image: Jens Klump (CC-BY)

South Atlantic (Namibia)

Image: Jens Klump (CC-BY)

Lake Baikal (Russia)

Is this a generational thing?

- Are “Digital Natives” more open to data sharing?
- The study “Researchers of Tomorrow” found that PhD students do not share more data.
- PhD students seem to emulate their supervisor’s behaviour.
- Some say “digital natives” do not exist, the behavioural drivers are more general.

Structural barriers

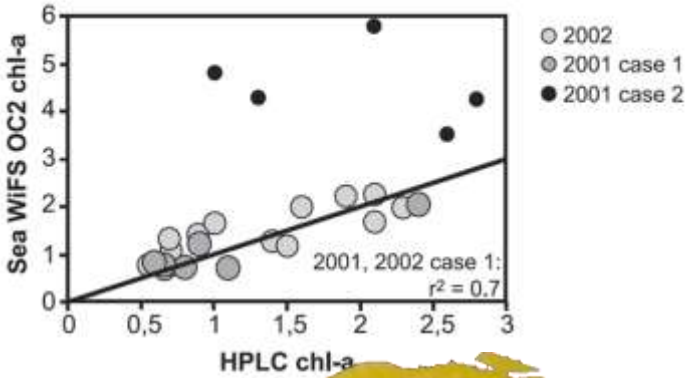
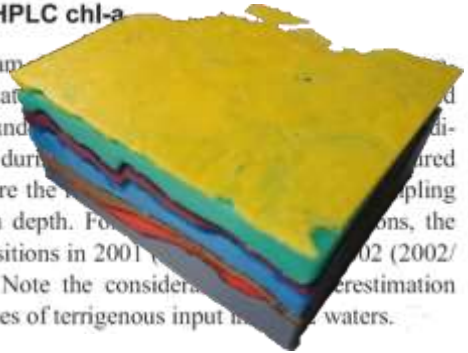
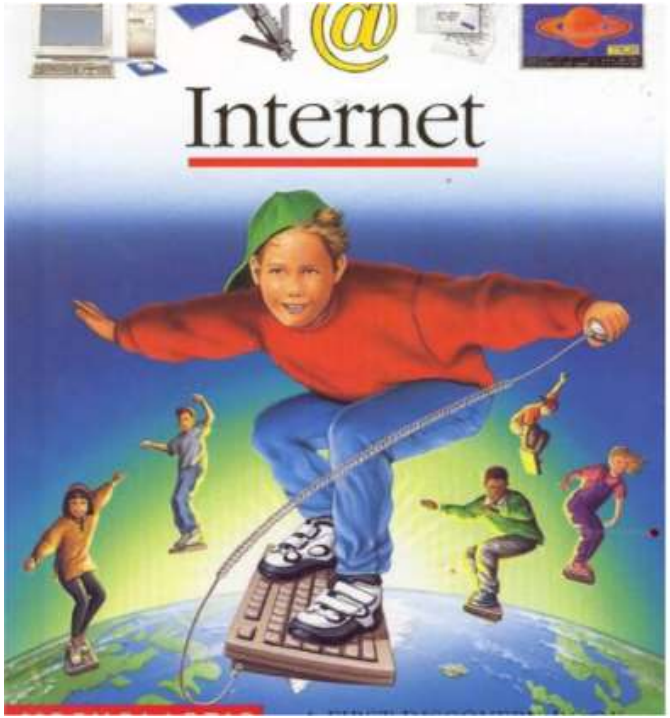


Fig. 2. The scattergram ... trations of chl-*a* calculat ... d determined from ground ... di- tions in Lake Baikal duri ... ured chlorophyll (HPLC) are the ... pling point from 5 to 30 m depth. Fo ... ons, the most cloud-free acquisitions in 2001 ... 02 (2002/ 07/20) were chosen. Note the consider ... erestimation caused by the influences of terrigenous input i ... waters.



- Structural barriers exist in journals.
- Many journals still emulate paper, data are limited to figures and tables.
- This does not allow the publication of large datasets or non-tabular data.

The internet will set us free?

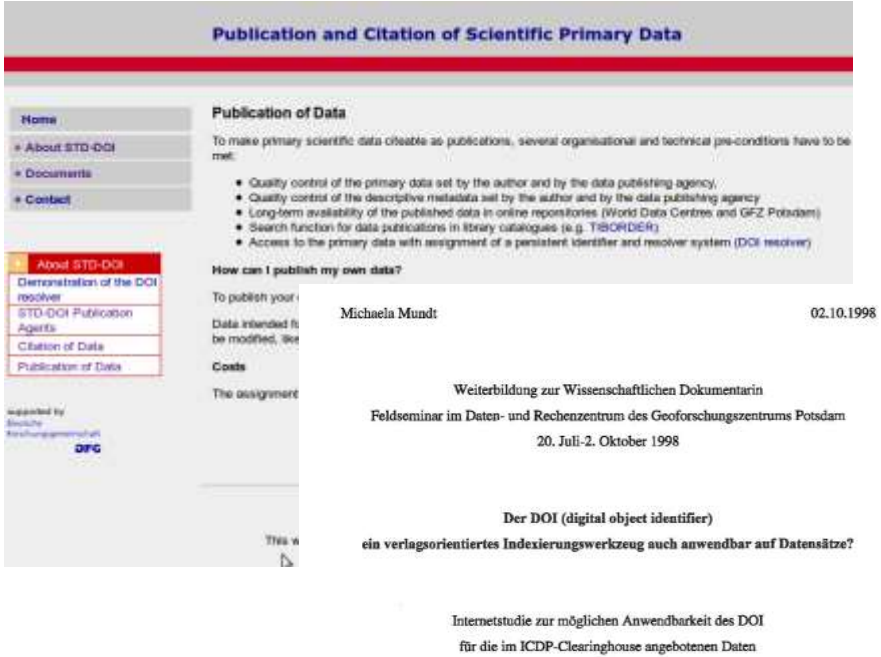


- The internet was invented to facilitate information exchange between researchers at CERN.
- It was expected that the emerging internet would broaden access to knowledge.
- Alas, it did not happen as expected.

Open Access to Data

- In 2003, the signatories of the “Berlin Declaration for Open Access to Knowledge in the Sciences and Humanities” called for open access not only to literature but also to data.
- In 2006 the OECD followed with a “Recommendation of the Council concerning Access to Research Data from Public Funding”
- More policies followed since.

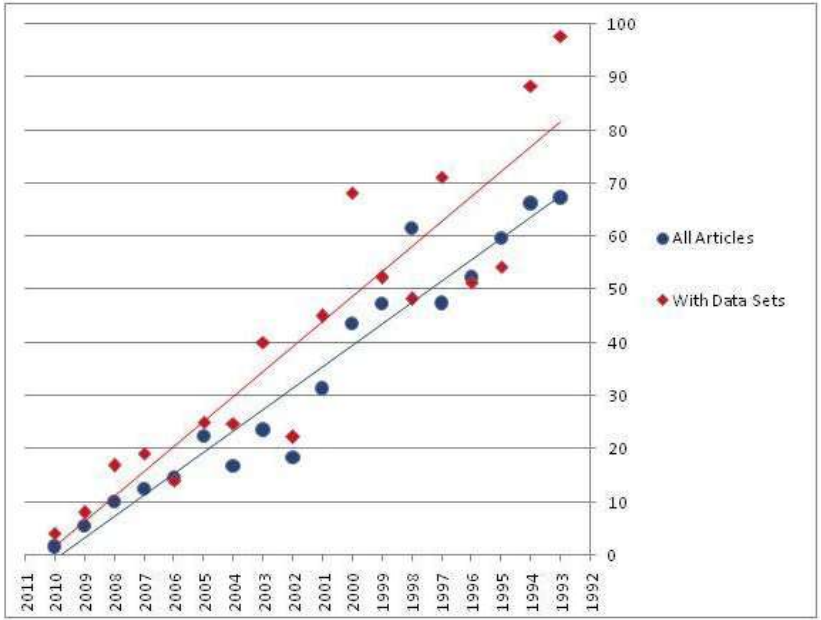
DOI for data publishing and citation



- The currency of science is the citation.
- Citation should be an incentive to publish data.
- Using DOI for data publications was seen as a way to treat data publications in the same way as classical publications and make them citeable.

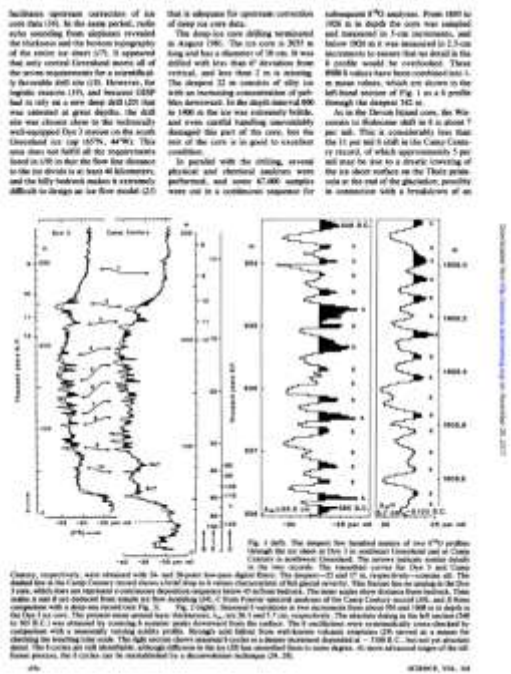
Is data citation an incentive?

- Current strategies focus on data citation as an incentive for data publication.
- Analysis of citation rate shows that publications with openly available data are cited more frequently and over a longer period of time.
- It takes a long time for this effect to show noticeable effects.
- This is not a strong incentive.



Sears (2011)

Do we cite data?



- Dansgaard, W., Clausen, H. B., Gundestrup, N., Hammer, C. U., Johnsen, S. F., Kristinsdottir, P. M., & Reeh, N. (1982). A New Greenland Deep Ice Core. *Science*, **218**(4579), 1273–1277.
- I often used the data of Dansgaard et al. (1982) as a reference curve.
- Did I cite the paper or the data?
- Where is the intellectual merit?

Empty archives

Image: Library of Congress, Prints & Photographs Division, MD-1111-77 (public domain)

- Since 2005 approx. 3.5 Million datasets have been registered through DataCite.
- CrossRef has published more than 90 million DOI during the same period.
- Compared to the number of publications, the number of data publications is still very small.
- Are we getting the incentives right?

Are we getting the incentives right?

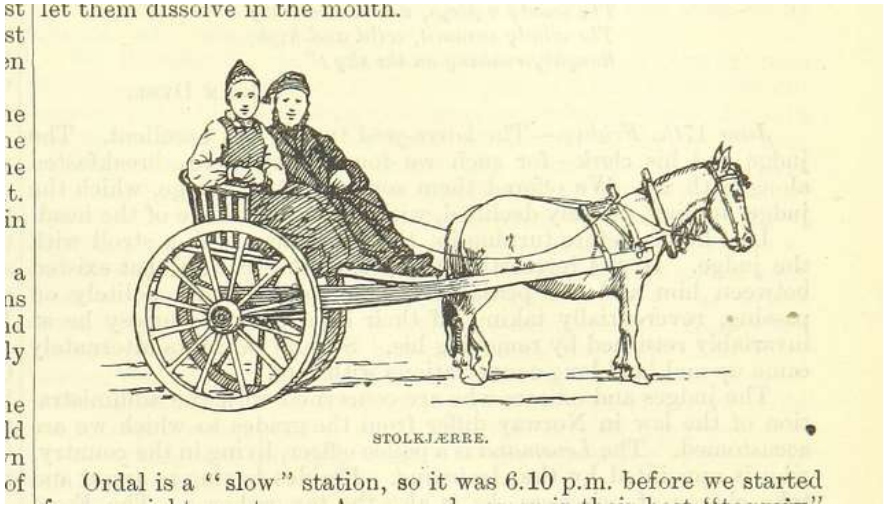


Image: British Library (public domain)

- If the desired norm is to share data, how do we motivate compliant behaviour?
- No norm is effective without enforcement measures but the academic system offers little leverage.
- Most effective, in this case, are community norms.
- "Carrot and stick" will not work because the horse is not harnessed to the cart.

Are we getting the incentives right?

- The situation may be quite different to the horse harnessed to the cart.
- In this situation, “carrot and stick” as means to motivate compliance do not work.
- The animals roaming the plains might not even be interested in what we are doing.
- *We have to find better strategies.*
- *We have to understand the social drivers.*

Image: ETH Zürich Library (public domain)

Gift culture in science

- Gift culture is a mode of exchange where valuables are not traded or sold, but given without an explicit agreement for immediate or future rewards.
- Scholarship is characterised by a gift culture in which members of the community make each other precious gifts.
- Putting data on the internet without being able to expect a gift in return is not an incentive in this model of scholarly culture.

Social capital

- The American definition of social capital refers to the networks of relationships among people who live and work in a particular society. This is not what I mean here.
- The European definition of social capital refers to it as a facet of social status.

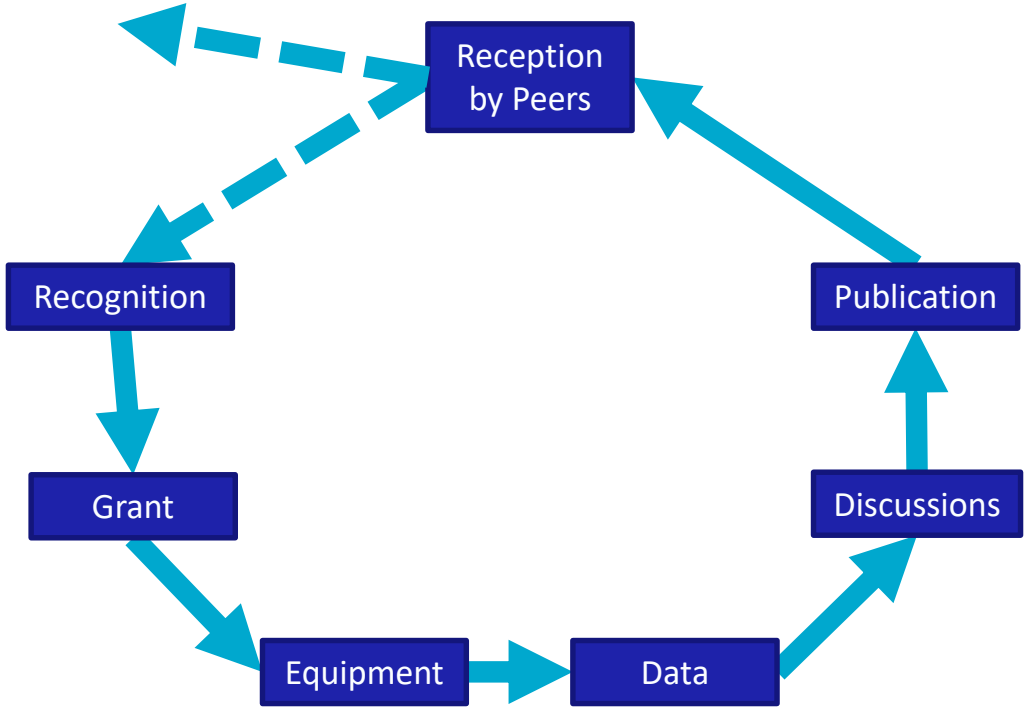
Bourdieu (1983) defines social capital as the means of an individual to influence social transactions and rise in social rank.

Social capital is based on material and symbolic exchange relationships. This exchange maintains, or even strengthens, relationships between individuals.

Data as social capital

- In the context of a scholarly reputation economy, data can be seen as a form of social capital.
- Sharing data with peers adds power to the network of obligations, expectations and trustworthiness of social structures among peers.
- Putting data on the internet without being able to expect a gain in scholarly reputation is not an incentive in this model of scholarly culture.

Reputation economy



After: Latour & Woolgar, 1982

Distinction gain vs. cooperation gain

Image: Nature (C)

Image: J Klump (CC-BY)

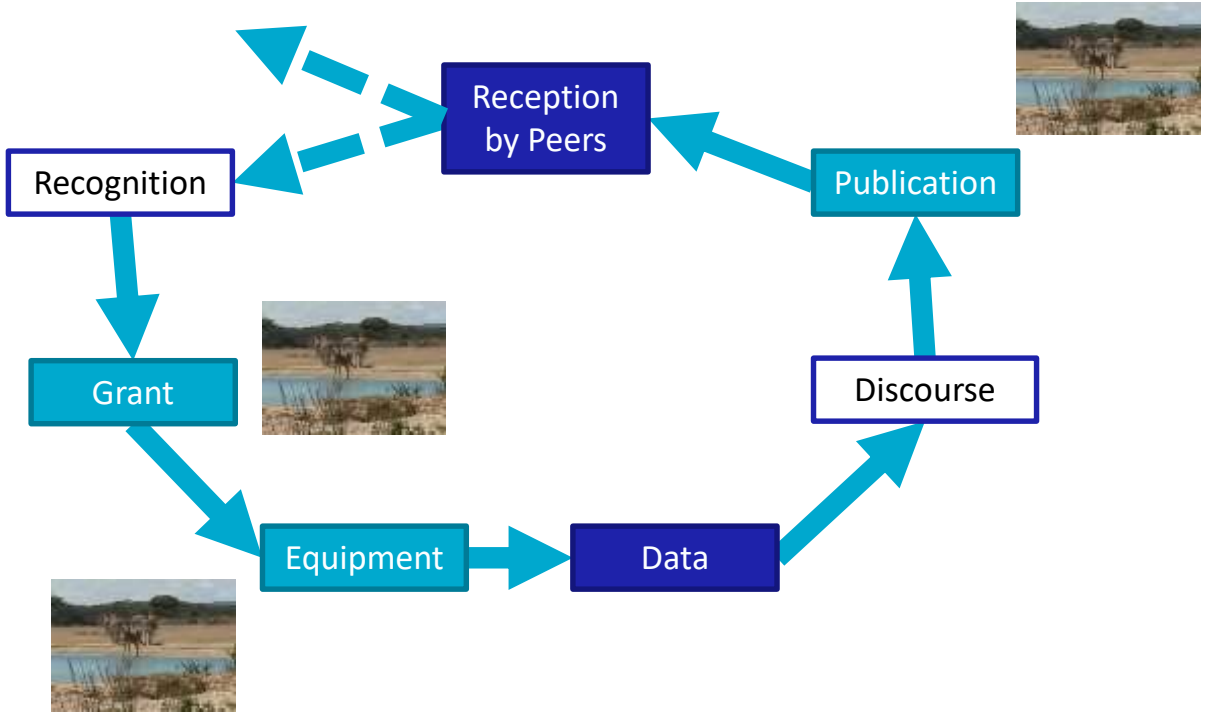
- Research is competitive but is also becoming more and more a collaborative exercise.
- Some projects are too big to be tackled by individuals, e.g. high-energy physics, ocean drilling, human genome, …
- Sometimes cooperation is necessary to gain and maintain distinction.
- Here, cooperation is enforced by strong social norms.

Waiting at the watering hole

- Sometimes waiting at the watering hole can a successful strategy.
- The art is to identify suitable watering holes.
- Which resources do researchers need to access for their distinction gain?
- This is not only an opportunity to coerce compliant behaviour but also to develop better services for researchers.

Image: Jens Klump (CC-BY)

Reputation Economy



The Role of the Funders

- Funders can set the norms for data publication through funding rules.
- Top-up funding may be given to cover the cost of data management.
- Not all funders are willing to police their data publication guidelines.

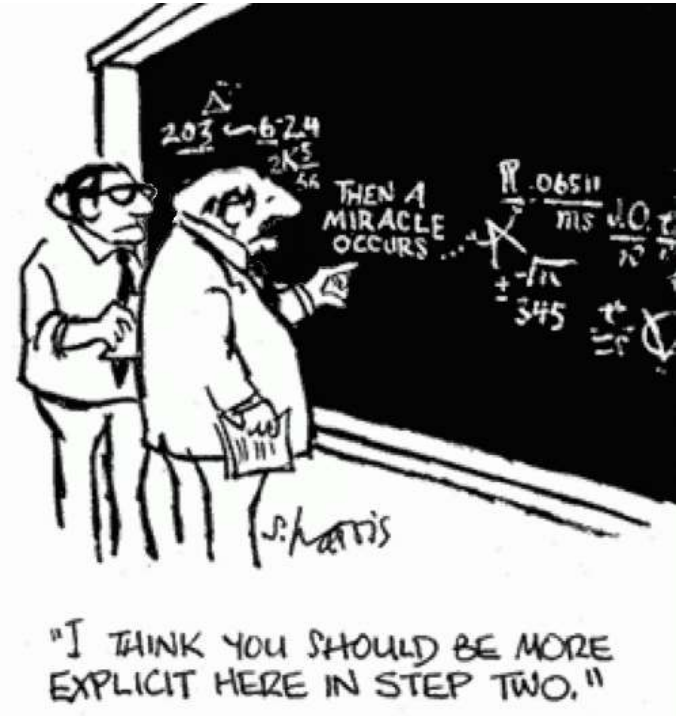
The Role of the Infrastructures

- Research is becoming more collaborative and infrastructures have an important role.
- Infrastructures are in a strong position to enforce data policies.
- Infrastructures should become more aware of their roles in the data lifecycle.

The Role of the Journals

- Journals have a central role in the scholarly discourse.
- As a matter of quality, journal papers should always come with “proof”.
- Journals are starting to demand that data accompanying a publication is deposited in a trustworthy data repository.
- Data citation is still not common practice.

Reproducible Science



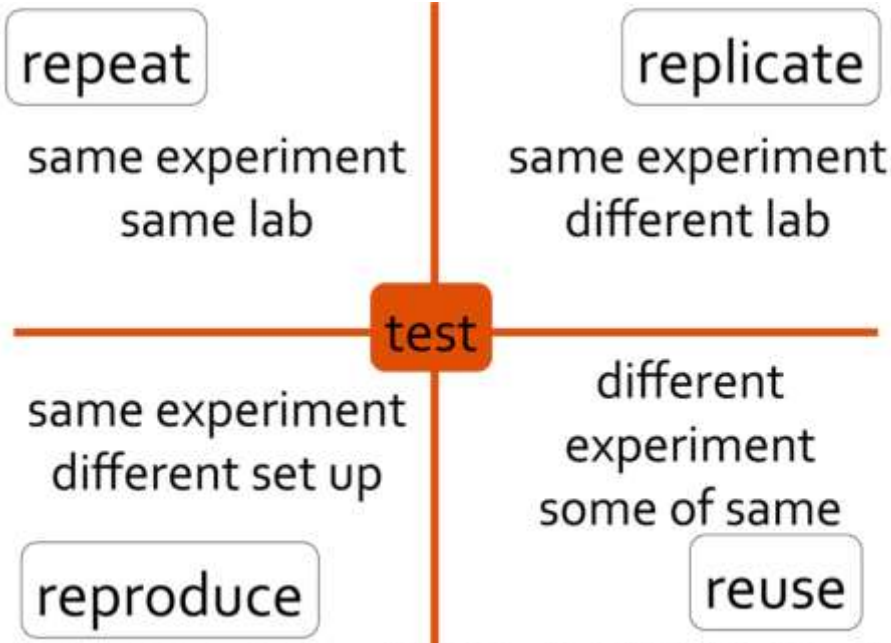
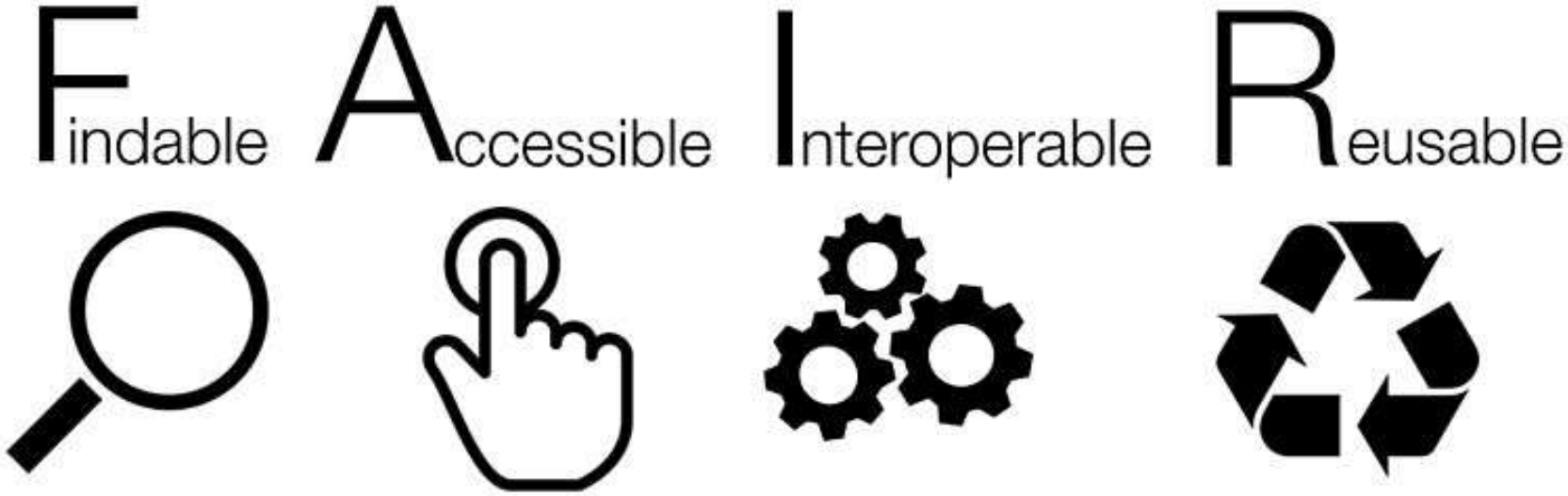


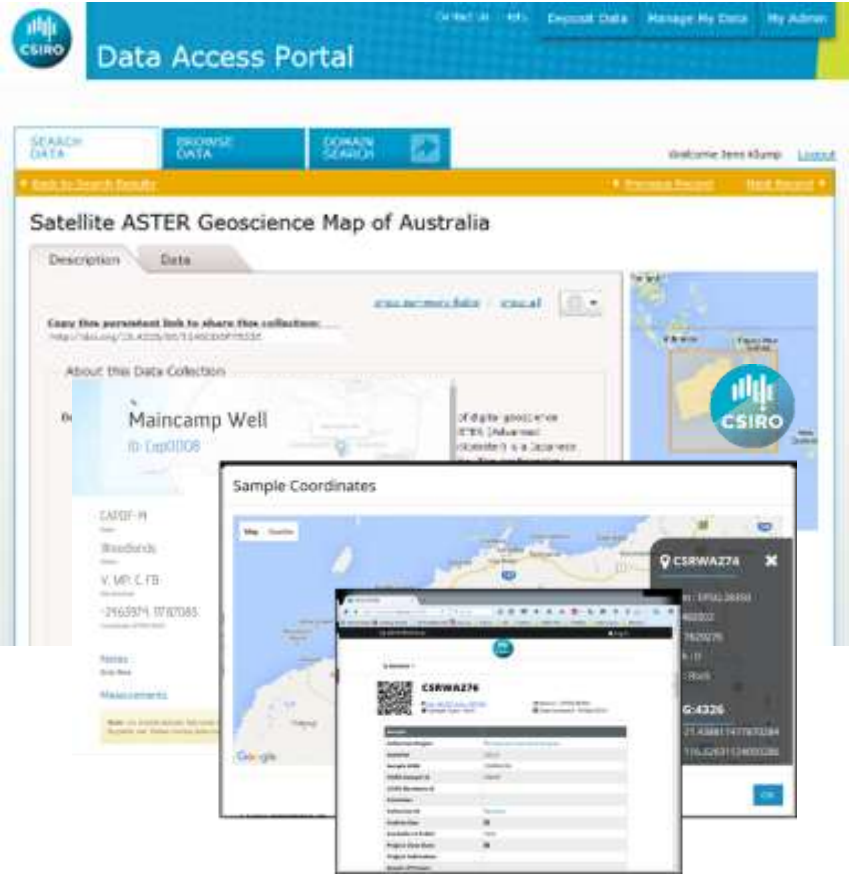
Figure by Carole Goble adapted from Drummond C, Replicability is not Reproducibility: Nor is it Good Science, online and Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

FAIR Data Principles



F - Findable

- Data and metadata are assigned a globally unique and persistent identifier.
- Data are described with rich metadata.
- Data/Metadata are registered or indexed in a searchable resource.

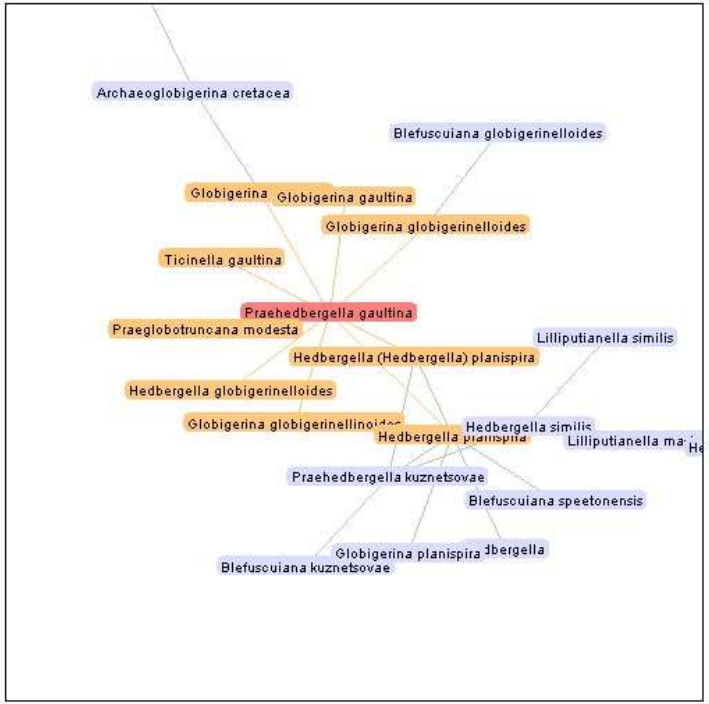


A - Accessible

- Making data open using a standardised protocol.
- Sometimes there can be good reasons why data cannot be made open (privacy, national security, commercial, cultural).
- Be transparent about the reasons for restricting access.

I - Interoperable

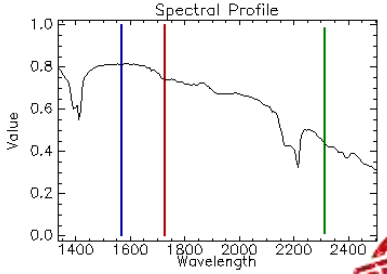
- Use community agreed formats, language and vocabularies.
- Link to related information using identifiers.
- This should include cross-linking between literature, data, and samples.



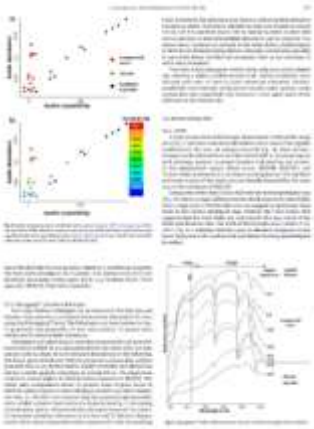
Linking Samples with Data and Publications

Specimen (Rock Store)

Spectrum
(Data Access Portal)



Publication



CERTIFIED ORIGINAL

CERTIFIED ORIGINAL

cross-reference

cross-reference

R - Reusable

- Maintain the initial richness of the data.
- Supply a machine readable licence and provenance information.
- Use discipline-specific data and metadata standards to give rich contextual information with the data.

PANGAEA®
Data Publisher for Earth & Environmental Science

Not logged in (log in or sign up)

Always quote citation when using data

Data Description

Citation: König-Langlo, G; Gernandt, H (2008): 426 ozonesonde profiles from Georg-Forster-Station. *Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research, Bremerhaven*, doi:10.1594/PANGAEA.547963, *Supplement to:* **König-Langlo, Gert; Gernandt, Hartwig (2009):** Compilation of ozonesonde profiles from the Antarctic Georg-Forster-Station from 1985 to 1992. *Earth System Science Data*, **1(1)**, 1-5, doi:10.5194/essd-1-1-2009

Abstract: On 22 May 1985 the first balloon-borne ozonesonde was successfully launched by the staff of Georg-Forster-Station (70°46' S, 11°41' E). The following weekly ozone soundings mark the beginning of the continuous investigation of Germany to study the vertical ozone distribution in the southern hemisphere.

In 1985 these ozone soundings have been the only record showing the change of vertical ozone distribution in the southern polar stratosphere in September and October. The regular ozone soundings from 1985 until 1992 are a valuable reference data set since the chemical ozone loss became a significant feature in the southern polar stratosphere.

The balloon-borne soundings were performed at the upper air sounding facility of the neighbouring station Novolazarevskaya, just 2 km apart from Georg-Forster-Station. Till 1992, ozone soundings were taken without interruption. Afterwards, the ozone sounding program was moved to Neumayer-Station (70°39' S, 8°15' W) 750 km further west.

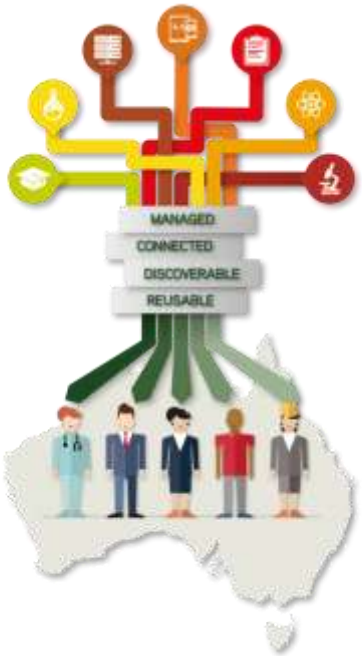
Project(s): **Meteorological Long-Term Observations @ AWI** (AWI_Meteo)

Coverage: *Latitude:* -70.770000 * *Longitude:* 11.830000

Date/Time Start: 1985-05-22T00:00:00 * *Date/Time End:* 1992-01-29T00:00:00

Event(s): **GF_8771** * *Latitude:* -70.770000 * *Longitude:* 11.830000 * *Date/Time:* 1985-05-22T00:00:00 * *Location:* Antarctic * *Campaign:* Ozone_studies_1985-1992 * *Basis:* Georg Forster Station * *Device:* Radiosonde (RADIO)

Open Research



- Research is producing larger and more complex data than ever before.
- These data outputs should be effectively managed and shared.
- Better data:
  - better described
  - more connected
  - more integrated and organised
  - more accessible
  - more easily used for new purposes
- Better data allows new questions to be answered, larger issues to be investigated, and data landscapes to be explored.

So, are we there yet?

- Data citation and data publication has come a long way.
- Compared to the total volume of publications, the number of data publications is still small.
- Initiatives such as Open Research and FAIR Data work to integrate data into the scholarly record.
- Achieving a change of culture around data requires us to understand the fundamental social drivers in the research communities.
- Equipped with understanding the drivers for change, all stakeholders must work together to implement this change.

Thank you

Mineral Resources
Jens Klump
Science Leader Earth Science Informatics

t +61 8 6236 8828
e jens.klump@csiro.au
w http://people.csiro.au/Jens-Klump

References

- Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. In R. Kreckel (Ed. & Trans.), Soziale Ungleichheiten (Vol. Special Volume 2). Göttingen, Germany. Retrieved from http://unirot.blogsport.de/images/bourdieukapital.pdf
- British Library, HEFCE, & JISC. (2012). Researchers of Tomorrow - The research behaviour of Generation Y doctoral students (p. 85). London, United Kingdom: JISC. Retrieved from http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow
- Drummond, C. (2009). Replicability is not Reproducibility: Nor is it Good Science. Presented at the 26th International Conference on Machine Learning (ICML 2009), Montréal, QB: International Machine Learning Society (IMLS). Retrieved from http://www.csi.uottawa.ca/~cdrummon/pubs/ICMLws09.pdf
- Hagstrom, W. O. (1982). Gift giving as an organising principle in science. In B. Barnes & D. Edge (Eds.), Science in Context: Readings in the Sociology of Science (pp. 21–34). Milton Keynes, United Kingdom: The Open University Press.
- Klump, J. (2017). Data as Social Capital and the Gift Culture in Research. Data Science Journal, 16(14), 1–8. https://doi.org/10.5334/dsj-2017-014
- Latour, B., & Woolgar, S. (1982). The cycle of credibility. In B. Barnes & D. Edge (Eds.), Science in Context: Readings in the Sociology of Science (pp. 35–43). Milton Keynes, United Kingdom: The Open University Press.
- Mauss, M. (2011). The Gift: Forms and Functions of Exchange in Archaic Societies. (I. Cunnison, Trans.). Mansfield Centre, CT: Martino Fine Books. Retrieved from https://libcom.org/files/Mauss%20-%20The%20Gift.pdf
- Mundt, M. (1998). Der DOI (digital object identifier) ein verlagsorientiertes Indexierungswerkzeug auch anwendbar auf Datensätze? (Semesterarbeit) (p. 19). Potsdam, Germany: Fachhochschule Potsdam. Retrieved from http://dx.doi.org/10.2312/GFZ.misc.370184
- Peng, R. D. (2011). Reproducible Research in Computational Science. Science, 334(6060), 1226–1227. https://doi.org/10.1126/science.1213847
- Sears, J. R. (2011). Data Sharing Effect on Article Citation Rate in Paleoceanography. EOS, Transactions, American Geophysical Union, 92(53, Fall Meet. Supp.), IN53B–1628. http://adsabs.harvard.edu/abs/2011AGUFMIN53B1628S