DATA SCIENCE JOURNAL

Callaghan, S. 2019. Research Data Publication: Moving Beyond the Metaphor. *Data Science Journal,* 18: 39, pp. 1–7. DOI: https://doi.org/10.5334/dsj-2019-039

ESSAY

Research Data Publication: Moving Beyond the Metaphor

Sarah Callaghan
STFC Rutherford Appleton Laboratory, GB
sorcha.ni@gmail.com

Metaphors are a quick and easy way of grasping (often complicated) concepts and ideas, but like any useful tools, they should be used carefully. There are as many arguments about how datasets are like cakes[1] as there are about how datasets aren't like cakes.[2]

It can be easy to categorise a dataset as being a special class of academic paper. Positively, this means that the tools and services for scholarly publication can be utilised to transmit and verify datasets, improving visibility, reproducibility, and attribution for the dataset creators. Negatively, if a dataset doesn't fit within the criteria to meet the "academic publication" mould (e.g. because it is being continually versioned and updated, or it is still being collected and will be for decades) it might be considered to be of less value to the community.

It is often said that "all models are wrong, but some are useful" (Box, 1979). Hence we need to determine the usefulness and limits of models and metaphors, especially when trying to develop new processes and systems.

This paper further develops the metaphors for data outlined in Parsons and Fox (2013), and gives real world examples of the metaphors from scientific data stored in the Centre for Environmental Data Analysis (CEDA) – a discipline-specific environmental data repository, and the processes that created the datasets.

Keywords: data publication; data management; data variability; data types

Introduction

The process of research is not only about learning and discovering, but also about sharing these discoveries with others, so that society as a whole can benefit from the efforts put in by the individual. When it comes to complex academic concepts, the choice of words for how a concept is described can make a difference to how well it is understood by others,[3] especially when moving between research domains.

Hence we make such use of metaphors and analogies when it comes to describing complex concepts. Tying a concept (for example, quantum superposition) to a real world "thing" (for example, Schrödinger's cat in a box) allows people unfamiliar with the original concept to connect it with something they have experience of, and provides a foundation that can be elaborated on. If, upon further examination, it is found that the analogy gets stretched beyond all reason, then that is acceptable, as long as those using it don't simply rely on it as an article of blind faith. Analogies and metaphors require critical thinking.

Scientific concepts are formulated in human language, and are intended to be processed by the human brain (even if that brain needs to be highly trained before it can properly grasp the concepts being described). Scientific data, on the other hand, is more often than not designed to be machine consumable (as well as

[1] They're a structured object created of raw materials, requiring set processes to create, and are generally more palatable and usable if presented in an appropriate way.

[2] Datasets aren't necessarily physical objects, consumption of a dataset doesn't mean the dataset no longer exists, datasets can be transported from one place to another via the internet.

[3] For an excellent discussion of the concept of polysemous terms (where the same term can have multiple meanings – i.e. duck – referring to an action, or duck – referring to aquatic birds) please see Joe Hourcle's Ignite talk at https://www.youtube.com/watch?v=oSWyg_RbqG8.

predominantly machine produced). Measurements are often not useful without the context surrounding them. It is one thing to know that a particular river level rose by 10 cm. It is only by knowing where this happened, how high the river was to begin with, and how high the rise would have to be at that location to flood the houses built there, that we are able to put the data into context, and make it useful.

Yet we still need that data. If a homeowner who got flooded wished to claim on their insurance for flood repairs, having that data and context available means they'd have proof that it was river flooding that caused the damage, rather than a burst pipe.

We also need to have the research data that underpins key research findings available and understandable, both for reproducibility and to prevent fraud/misuse. Making data usable by others takes effort and time and is often unrewarded by the current system for gaining academic credit.

Metaphors and Analogies – definitions

At this point, it is prudent to define exactly what is meant in this essay when it the words "metaphor" and "analogy" are used.

From Lexico, a metaphor is defined as: "A figure of speech in which a word or phrase is applied to an object or action to which it is not literally applicable",[4] while an analogy is "A comparison between one thing and another, typically for the purpose of explanation or clarification."[5] Metaphors are also defined as "A thing regarded as representative or symbolic of something else", while analogies are "A thing which is comparable to something else in significant respects."

In this essay, I lean towards the latter definition of analogy as being a more useful mental model for the further discussion, as we are directly comparing data and its methods of production, dissemination, etc. with other real world methods for doing the same things, but with other outputs. For this reason, the text speaks more of analogies than metaphors, while the title pays homage to the essay (Parsons & Fox, 2013) that inspired this work.

Metaphors and Analogies – discussion

> "No one metaphor satisfies enough key data system attributes and multiple metaphors need to co-exist in support of a healthy data ecosystem." (Parsons & Fox, 2013)

Data publication as a metaphor has been addressed extensively in (Parsons & Fox, 2013), leading to the quote above. But before we dive into examples of metaphor and analogy in the data domain, it is helpful to review what they mean.

From (Gentner & Jeziorski, 1993):

> *'Analogy can he viewed as a kind of highly selective similarity. In processing analogy, people implicitly focus on certain kinds of commonalities and ignore others. Imagine a bright student reading the analogy "a cell*[6] *is like a factory." She is unlikely to decide that cells are buildings made of brick and steel. Instead she might guess that, like a factory, a cell takes in resources to keep itself operating and to generate its products. This focus on common relational abstractions is what makes analogy illuminating.'* (Gentner & Jeziorski, 1993) p 448

This action of focussing on some commonalities and ignoring others is crucial when using analogies to illustrate scientific concepts. We can produce an analogy that "a dataset is like a book". Commonalities include that both contain information, in a structured and formatted way, which is consumable by a user, and both are the product of sustained effort, potentially from a wide range of actors. The differences between them make it just as easy to say "a dataset is not like a book", in that a dataset can be constantly changing; may not be a physical, but a virtual object; mostly isn't designed for humans to read unassisted;[7] and often what people consider to be "a dataset" isn't a self-contained unit (as it requires extra information and metadata to make it understandable and usable). This metadata is indeed part of the dataset, but it is often stored in different formats and locations to the data, and hence can be assembled differently by different users.

[4] https://www.lexico.com/en/definition/metaphor.

[5] https://www.lexico.com/en/definition/analogy.

[6] The word cell itself is an analogy, originally defined as: "small room for a monk or a nun in a monastic establishment; a hermit's dwelling" (c. 1300), from Latin cella "small room, store room, hut," (https://www.etymonline.com/word/cell).

[7] Yes, before a human can ingest the information in a book, they have to learn to read the language it's written in, and some may require assistance to read it, e.g. reading glasses, audio format, etc.

Obviously, it is possible to push analogies too far, and have them break. This is more likely to happen when users of the analogy don't have a good understanding of each of the two things being compared. In the (Gentner & Jeziorski, 1993) quote above, if the student didn't have any other concept of what a cell was, she could easily imagine that they were tiny buildings made of bricks and steel, and the analogy used would do nothing to correct that misapprehension.

It is also important to remember that analogy is not causation – if two phenomena are analogous, it does not imply that one causes the other. Most of the time this is obvious, but it must be explicitly kept in mind for those times when it is not.

A similar, and entertaining blog post on data using an air transportation analogy is given by (Lusoli, 2017).

Types of analogy and real world scientific examples

Data Publication

Data publication, as an analogy, came about as a result of several drivers, including:

- researchers being required to publish as many works as possible in as many high impact journals as possible
- dataset creators wanting to be given recognition for their work, and their efforts to make the data findable, accessible, interoperable and reusable
- repository managers and data curators wanting to quantify the impact of the data in their archives, and get credit for their efforts for making it usable by the wider community

This resulted in pressure to squeeze all research outputs into shapes that resemble publications, hence the proliferation of the data journal, a place where researchers can publish a paper about their dataset, linked via permanent identifier to the dataset itself (stored in a trustworthy repository). The data paper then can be cited and used as a proxy for the dataset when reporting the importance and impact of the researcher's work.

The data paper should provide comprehensive and full (human-readable) metadata in a peer-reviewed and quality controlled way, while also ensuring that the dataset it describes has also been checked from the point of view of understandability and usability.

A real-world example of a dataset that has been published in a data journal is the Global Broadcast Service (GBS) datasets (Callaghan et al., 2013), measurements from a radio propagation dataset investigating how rain and clouds impact signal levels from a geosynchronous satellite beacon at radio frequencies of 20.7 GHz. The data streams linked to the paper, and which the paper describes in detail, are the result of a definite, discrete experiment, resulting in a well-defined, discrete and fully completed dataset, which will not change in the future. The dataset has been through two levels of quality assurance: the first was performed on ingestion into CEDA,[8] where the file formats were standardised and metadata was checked and completed. The second level of quality assurance was performed as part of the scientific peer review process carried out when the data paper and dataset were submitted to the *Geoscience Data Journal* for review and publication.

As this dataset is complete, well-documented and quality assured, it can be considered to be a first-class, reference-able, scientific artefact. There are other peer-reviewed journal articles which use the GBS data as the basis for their results, see for example (Callaghan et al., 2008).[9] However, datasets can be discrete, complete, well-defined and permanently available and citeable without the need for the proxy of a data paper, or any other publication attached to them. This is of particular value when it comes to publishing negative results, or data that don't support the hypothesis they were collected to verify, but may be useful for testing other hypotheses.

These types of datasets are possibly the closest thing we have to the "dataset as a book" analogy, and therefore are the easiest to fit into the data publication mould. Unfortunately, many other datasets do not fit in with this shape. Many datasets are dynamic, and are modified or added to as time progresses. Then there are issues with granularity – some researchers may only need a subset of a larger dataset for their work, but need to accurately and permanently identify that subset. Citing at the level of every one of the subsets results in reference lists that are long and unwieldy, and can make it difficult to find the subset required in a long list of very similarly named datasets.

[8] Centre for Environmental Data Analysis, http://www.ceda.ac.uk.

[9] Though it is worth noting that these papers were written before data citation and publication were common, or even implemented, and so the link between the dataset and the paper using it is not made explicit or permanent.

For text based items, such as books and articles, tools exist to compare text from one instance of an article to another, allowing the reader to be sure that the contents of two instances are the same, regardless of the format they are in (for example, an article in hard copy in a journal as compared with a pdf). We currently do not have a way of evaluating the scientific equivalence of datasets regardless of their format. The ease with which it's possible to modify datasets (and not track the changes made) also means that it can be very hard to tell which dataset is the canonical, original version, or even what the differences are.

Data publication can work very well as an analogy, but users must be aware that it really is only applicable to the subset of datasets which can be made complete, well-documented, well-defined, discrete and quality controlled.

Users should also be aware that there are different publishing authorities (and review criteria) for the data set and the data paper. This can put a false primacy on the data paper and undercut the perceived value of the underlying data and their creators and curators. When users are accustomed to having the paper as the primary (and only) object of authority, the dataset may then be downgraded as to merely being "supporting information" when in fact it is the article that provides the supporting information about the data.

Big Iron (industrialised data production)

Big Iron, as defined in (Parsons & Fox, 2013) typically deals with massive volumes of data that are relatively homogenous and well defined but highly dynamic and with high throughput. It is an industrialised process, relying on large, sophisticated, well-controlled, technical infrastructures, often requiring supercomputing centres, dedicated networks, substantial budgets, and specialized interfaces.

An example of this is the data from the Large Hadron Collider, CERN, but in the Earth Sciences, the Coupled Model Intercomparison Projects (CMIP) are another.

The Intergovernmental Panel on Climate Change[10] (IPCC) regularly issues Assessment Reports, detailing the current state of the art of climate models, and their predictions for future climate change. These reports are supported by the data from the climate model runs performed as part of CMIP. Each CMIP is an international collaboration, where climate modelling centres around the world run the same experiments on their different climate models, collect and document the data in standard ways and make it all available for the wider community to use, via custom built web portals.

CMIP5, the most recent complete CMIP, resulted in datasets totalling over 2 PB of data. As this data is the foundation for the IPCC assessment and recommendations, it is vital that the data is stored and archived properly.[11] Dealing with these data volumes requires not only custom built infrastructure, but also standards for file and metadata formats (e.g. NetCDF, CF Conventions, CMOR, etc.). Collecting the metadata describing the experiments that were run to create the datasets alone took several weeks' worth of effort, and several years of effort to design and build the CMIP5 questionnaire which collected the metadata (Guilyardi et al., 2013).

The industrialised production of data is likely to increase over the next years, given the increased ability of researchers to create and manage big data. The opposite of this analogy is also valid in many cases, as described in the next section.

Artist's studio (small scale data production, unique and non-standard output)

Similar to Big Iron, this analogy focusses on the method of production of a dataset, rather than the dataset itself. The artist studio analogy covers the long tail of data produced by small groups or even single researchers, working in relative isolation.

Artist studios generally produce one-of-a-kind pieces, which may have standard shapes and forms (e.g. oil paintings) but may equally come in non-standard shapes, sizes and materials (e.g. sculptures, video and audio installations, performance art etc.) The aim is to produce something of use/interest to a consumer, even if they are part of a limited domain. Similarly, it's often not easy, or even possible to share the outputs of the studio (it is possible to make copies/prints of paintings, and smaller models of sculptures, but other objects of art, like Damien Hirst's famous shark in formaldehyde (Hirst, 1991) are nearly impossible to reproduce[12]).

[10] http://www.ipcc.ch/.

[11] This is done by the Earth System Grid Federation (ESGF) Peer-to-Peer (P2P) collaboration and enterprise system, which develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data. https://esgf-node.llnl.gov/projects/esgf-llnl/.

[12] Sharing this work using half a goldfish in a plastic bag doesn't capture the full nature of the piece, even though it would be easier to transport.

Datasets produced by small research groups follow this analogy. The emphasis is on the production of the finished product, sometimes with the supporting documentation and metadata being neglected, due to lack of time, effort and potentially interest on the part of the creator. If the dataset is only aimed at a small user group, then the metadata is provided as jargon, or users are simply assumed to have a sufficient level of background knowledge. Sharing the data is often not considered, as for the researchers, holding the only copy of the data makes it more valuable, and therefore more likely that they'll receive extra funding.

An example "artist studio" is the Chilbolton Facility for Atmospheric and Radio Research (CFARR).[13] It is a small facility, located in Hampshire, UK, with approximately 6 permanent staff, who collectively build, maintain and run a selection of meteorological and radio research instruments. In recent years, the focus of the facility has been on collaborations with other research groups in universities and other research centres. Previously the facility had been more focussed on radio research, and as such had developed its own data format for the instruments it built, rather than tying in with existing community standards. Similarly, the data was stored on a variety of servers, with a bespoke tape backup system.

When CFARR's funding structure changed, pressure was put on the staff to archive all new data and the majority of existing data in CEDA. This made it easier for the facility staff, in that they no longer needed to maintain servers or the backup system, but it made things harder in that effort was needed to convert the data files to netCDF, and to collect and agree on the metadata that should accompany them. The culture change to move from the artist studio model to a more standardised and collaborative model took effort and time, and should not be underestimated.

Science Support

Science support is what data repositories such as CEDA[14] do on an operational, everyday basis. Even though we're not directly (or physically) embedded in a research organisation,[15] we interact with researchers and research centres on a regular basis to ensure that the processes for data ingestion are carried out smoothly and efficiently.

For data centres embedded in a research centre, data management can be seen as a component of the broader "science support" infrastructure of the lab or the project, equivalent to facilities management, field logistics, administrative support, systems administration, equipment development, etc. In our case, CEDA concentrates on data management, and providing services to make it and use of data easier for the researcher.

Different data centres will have different ways of providing science support to their core user base. For example, an institutional data repository, responsible for all the data being produced by a university will have datasets which are non-standardised and are usually geared towards a specific set of intended uses and local reuse in conjunction with other local data. In terms of the "artist studio" analogy, an institutional repository is like an art gallery or museum, where different datasets will have different data management requirements. By contrast CEDA, which has multiple PB of data in the archives, must standardise in terms of file formats, metadata models etc., hence moving towards a more "Big Iron" analogy.

In common with institutional repositories, CEDA also focuses on managing data (and sometimes merging datasets to create more useful resources) in order to meet the needs of our user community, which is international in scope and covers a wide range of users, from schoolchildren, to policy makers, to field researchers and theoreticians.

Map Making

Map making as an analogy refers to the final representation of the data, and the process of putting the data into a context, primarily geographical. Maps also help to define the boundaries of what is known, and what isn't. Though data presented in this way tend to be fixed in time, maps are useful for showing dynamical datasets, or time slices through complex multidimensional processes, e.g. the four dimensional structures of clouds/rain changing in time.

The results of map making, the maps themselves, are datasets in their own right, and so need to be treated in the same way as other datasets with regard to preservation, metadata etc. The act of plotting some parameter on a geographical map results in a well-standardised structure for intercomparison and visualisation.

---

[13] https://www.stfc.ac.uk/research/environment/chilbolton-facility-for-atmospheric-and-radio-research-cfarr/.

[14] Centre for Environmental Data Analysis http://www.ceda.ac.uk.

[15] Other NERC datacentres are wholly embedded in their research centre – for example the National Geological Data Centre, within the British Geological Society. https://www.bgs.ac.uk/services/ngdc/.

Linked Data

The "data" in Linked Data are defined extremely broadly and are envisioned as small, independent things with specific names (URIs) interconnected through defined semantic relationships (predicates) using model and language standards (e.g. the Resource Description Framework, RDF). It has a major emphasis on Open Data, as linked data focuses on enabling the interoperability of data and capitalising on the interconnected nature of the Internet.

Linked data isn't commonly used for dealing with scientific data, but instead, is predominantly used in our metadata, where we have complete focus on preservation, curation and quality, unlike other linked datasets available elsewhere. Using linked data for metadata structures does require standardisation and agreement on the formal semantics and ontologies.

Linked data is very flexible, and lends itself well to distributed and interdisciplinary connections, provided the formal semantics can be agreed to be applicable across multiple domains. Linked data as a concept unfortunately hasn't fully permeated the research environment as yet – many scientific researchers don't understand the semantics (and have little interest in them).

Linked data is often used as a support structure for Big Iron.

The Cloud: "x as a service"

There is an argument that the mechanisms for data publication should be invisible, and data should be accessible and understandable without any prior knowledge. Cloud services such as Dropbox allow users to store their data, and access them from any web browser, or mobile app, provided they have an internet connection.

"Data as a service" ties in with "software as a service", in that the users only take the data they need at any given moment, and in some cases may not even download it, instead using dedicated computing resources elsewhere to perform the manipulations needed on the data.

An example of this is JASMIN,[16] a system that provides petascale storage and cloud computing for big data challenges in environmental science. JASMIN provides flexible data access to users, allowing them to collaborate in self-managing group workspaces. JASMIN brings compute and data together to enable models and algorithms to be evaluated alongside curated archive data, and for data to be shared and evaluated before being deposited in the permanent archive.

Data, in this context, aren't the fixed and complete products described in other analogies, but instead are more fluid and dynamic. Still, once the datasets are deposited in the permanent archive, they become fixed products, and are citeable and publishable.

Providing significant resources for data manipulation is undoubtedly useful, but the focus with this system is on the service, not necessarily on the data. The data however, is the backbone of the system – there is no point having the service without the data and the users who want to analyse it.

Conclusions

It goes without saying that all analogies are wrong, but some are useful, and hence should come with a health warning – especially when following an analogy to the furthest reaches of its logic can result in sheer absurdity.[17] When dealing with data, just like in life, there is no all-encompassing analogy for what we do. Instead, metaphors and analogies should be used in ways to illuminate and clarify, but we should always remember that metaphors are useful tools for thinking about things, but can also limit how we think about things. (Ball, 2011, Parsons and Fox, 2013). Pushing an analogy so far that it breaks can be a useful process, in that it helps determine the limits of understanding, especially as part of an ongoing conversation. Finally, for this essay, the author would like to leave the reader with some very appropriate words from (Polya, 1954, page 15):

> "And remember, do not neglect vague analogies. But if you wish them respectable, try to clarify them."

[16] http://www.jasmin.ac.uk/.

[17] For example, the quote misattributed to Albert Einstein "You see, wire telegraph is a kind of a very, very long cat. You pull his tail in New York and his head is meowing in Los Angeles. Do you understand this? And radio operates exactly the same way: you send signals here, they receive them there. The only difference is that there is no cat." https://quoteinvestigator.com/2012/02/24/telegraph-cat/.

Acknowledgements

I would like to thank Charlotte Pascoe and Bryan Lawrence for their helpful comments on earlier drafts of this essay, and the reviewers for their guidance and useful suggestions. Further thanks should go to all my colleagues at the Centre for Environmental Data Analysis and in the other NERC Environmental Data Centres for their help, support and collaboration over the years.

Competing Interests

The author has no competing interests to declare.

Author Information

During the time of writing this essay Sarah Callaghan was Programme Manager for the Centre for Environmental Data Analysis at STFC Rutherford Appleton Laboratory, and was Editor-in-Chief for the Data Science Journal. She moved to a new role as Editor-in-Chief at Cell Press on the 1st May 2019.

References

Ball, P. 2011. A metaphor too far. *Nature*. DOI: https://doi.org/10.1038/news.2011.115

Box, GEP. 1979. "Robustness in the strategy of scientific model building". In Launer, RL and Wilkinson, GN (eds.), *Robustness in Statistics*, 201–236. Academic Press. DOI: https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Callaghan, SA, Boyes, B, Couchman, A, Waight, J, Walden, CJ and **Ventouras, S.** 2008. An investigation of site diversity and comparison with ITU-R recommendations. *Radio Science*, 43(4). DOI: https://doi.org/10.1029/2007RS003793

Callaghan, SA, Waight, J, Agnew, JL, Walden, CJ, Wrench, CL and **Ventouras, S.** 2013. The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK. *Geoscience Data Journal*, (August 2003). DOI: https://doi.org/10.1002/gdj3.2

Gentner, D and **Jeziorski, M.** 1993. The shift from metaphor to analogy in Western science.

Guilyardi, E, Balaji, V, Lawrence, B, Callaghan, S, Deluca, C, Denvil, S, Lautenschlager, M, Morgan, M, Murphy, S and **Taylor, KE.** 2013. Documenting Climate Models and Their Simulations. *Bull. Amer. Meteor. Soc.*, 94: 623–627. DOI: https://doi.org/10.1175/BAMS-D-11-00035.1

Hirst, D. 1991. "The Physical Impossibility of Death in the Mind of Someone Living". *2170 × 5420 × 1800 mm, Glass, painted steel, silicone, monofilament, shark and formaldehyde solution.*

Lusoli, W. 2017. "The open research data (air)space." https://www.linkedin.com/pulse/open-research-data-airspace-wainer-lusoli/.

Parsons, MA and **Fox, PA.** 2013. Is Data Publication the Right Metaphor? *Data Science Journal*, 12: WDS32–WDS46. DOI: https://doi.org/10.2481/dsj.WDS-042

Polya, G. 1954. Mathematics and Plausible Reasoning. In *Induction and Analogy in Mathematics*. Princeton, NJ: Princeton Univ. Press.

How to cite this article: Callaghan, S. 2019. Research Data Publication: Moving Beyond the Metaphor. *Data Science Journal,* 18: 39, pp. 1–7. DOI: https://doi.org/10.5334/dsj-2019-039

Submitted: 14 March 2018 **Accepted:** 23 July 2019 **Published:** 14 August 2019