"Once upon a time..."

Mike Taylor (Digital Science, UK)

Representations of data have evolved considerably in the last 20 years: in terms of the scale of the data we have, the ways in which we can communicate and the expectations of our audiences. But as our tools evolve, so we have to evolve our understanding of what we're doing, the limitations, strengths or weaknesses and how we can use visualization to support our insights and narratives.

Introduction

A long, long time ago, the options for visualizing data were extremely limited: you were basically limited to one of three options: Excel, using very specialized software (SPSS for us social scientists) or drawing the graphics – if not by hand, then using something dire, like MS Paint. As if that wasn't limiting enough, the main way of research communication (in fact *all* written communication) was via paper, and much of that was black and white[i]. Even when electronic distribution became more common, legacy systems were mostly still in that greyscale mode, and we'd often have to have pained conversations with authors about whether the *dashed* lines in a chart were sufficiently different from the *dotted* lines and why (when a 10% grey tint was obviously quite different from a 15% grey tint on the lab's laser printer) the graphs weren't going to work in a book or on a screen limited to 256 shades of grey. With such constraints, the text had to do the heavy-lifting of conveying the story behind the research: data visualization was a thing to be approached with caution and to be used only when unavoidable.

I wish I was joking: as someone who has spent much of their life supporting scientists to communicate their findings (when not actually doing research myself), I'm still triggered by the memory of picking up a freshly printed book on audio technology to find that the oscilloscope plots were essentially illegible. That, dear reader, was in 2000. It feels like a long, long time ago.

Characters, plots, themes

When we sit down to tell a fairy story, everyone knows that the wolf is bad, that the forest is dark and that bears are domestically minded, but quite possessive about their porridge: much of the background is already known to the audience. And when we seek to produce an engaging narrative about our research findings, we also need to make sure that everyone is on the same page.

The field that I work in (and research) is usually known as altmetrics (plural, with a small 'a', the company I work at is Altmetric), which is a shortened form of 'alternative metrics'. To make a sweeping generalization, altmetrics is the study of links between research and the broader internet (Figure 1). Think citations in Wikipedia or in a government-published policy document, links on Twitter and Facebook, mentions in news outlets and blogs: all of which link to a journal article, or a clinical trial, or a book.

But I hear a question: "what are they an alternative to?" And that is a very reasonable question, the answer to which has shifted over the last decade.

Ten years ago, the folk who had the idea of altmetrics would have probably answered by positioning altmetrics as an alternative to citation-based metrics like the Journal Impact Factor, or the H-Index. The original manifesto that described the ambitions of altmetrics uses quasi-political terms and is accessible here (https://www.altmetrics.org). I strongly recommend it: some of the ambition has come to pass, much hasn't and some hopes weren't backed up the data.

Nowadays, I think most people would say that 'altmetrics' provide an *alternative view* into understanding 'impact'.

One of the nice things about working in this field is that most people know something of what we're talking about; and one of the less nice things is that people often have a set of views that are located in their personal experience, or prejudices from the past. Wikipedia is regarded as a low-quality information source by some, despite its frequent use of specialized editors for scientific subjects, its formal rules on citation, its status as one of the world's top sites and the way in which its data is used to enrich search results in Google and others.[ii]

And for people with an academic background, there's also the occasional belief that citations are an objective measure of impact without bias or perspective. A *very* common phenomenon is that researchers in other fields are unfamiliar that this – the field of altmetrics – is

[i]Up until the late 1990s , I worked for a publishing company that levied a colour charge on anyone who wished to go beyond black and white; from memory that base price was £85 for the first image – several hundred pounds in contemporary pricing.

[ii]When my son started university 2 years ago, he was asked to do some preparatory reading: this included many Wikipedia pages, and several YouTube videos (as well as books).

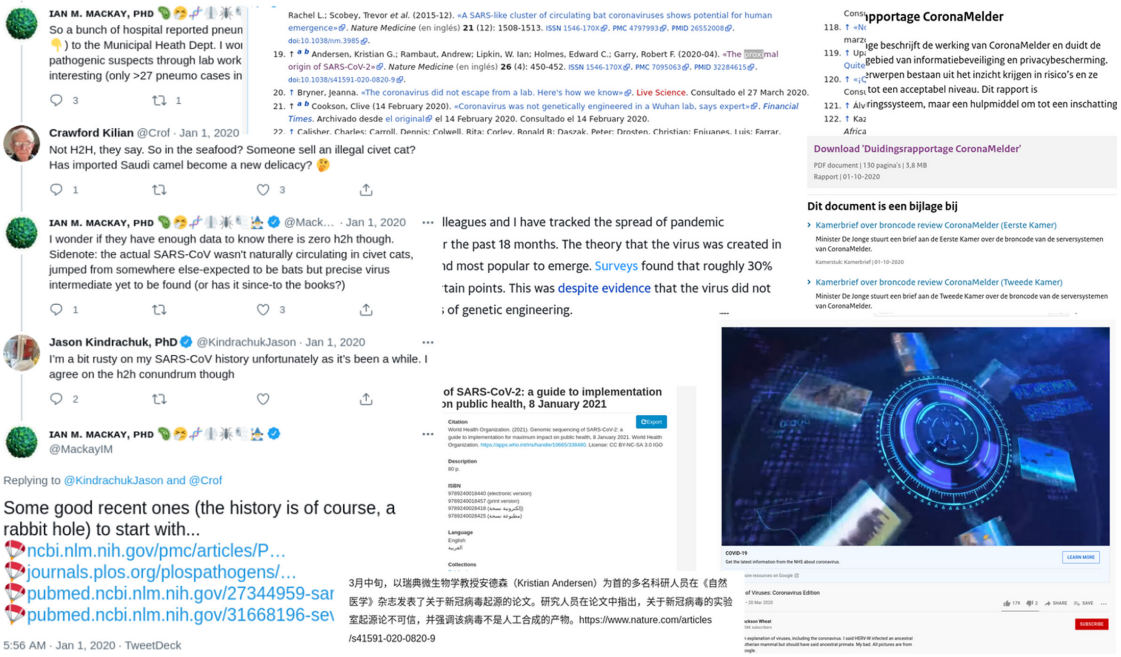

Figure 1. Examples of mentions, links and citations found across the internet. The company, Altmetric, searches the internet for mentions, links and citations to research; the research field, altmetrics, studies the meaning, frequency and nature of those links.

an area of academic study, with journals, conferences, books, grants, etc.

In short: in conversations with academics, we're rarely on the same page, and one of the first tasks that I have to do is to make sure that our visualizations and narratives are rooted in a common belief structure. The way that I choose to start my storytelling is by rooting it within a familiar background, using terms and approaches that will make the audience receptive (Figure 2).

Understanding our audience, remaining authentic to our characters and retaining a sense of story development are essential skills for the storyteller.

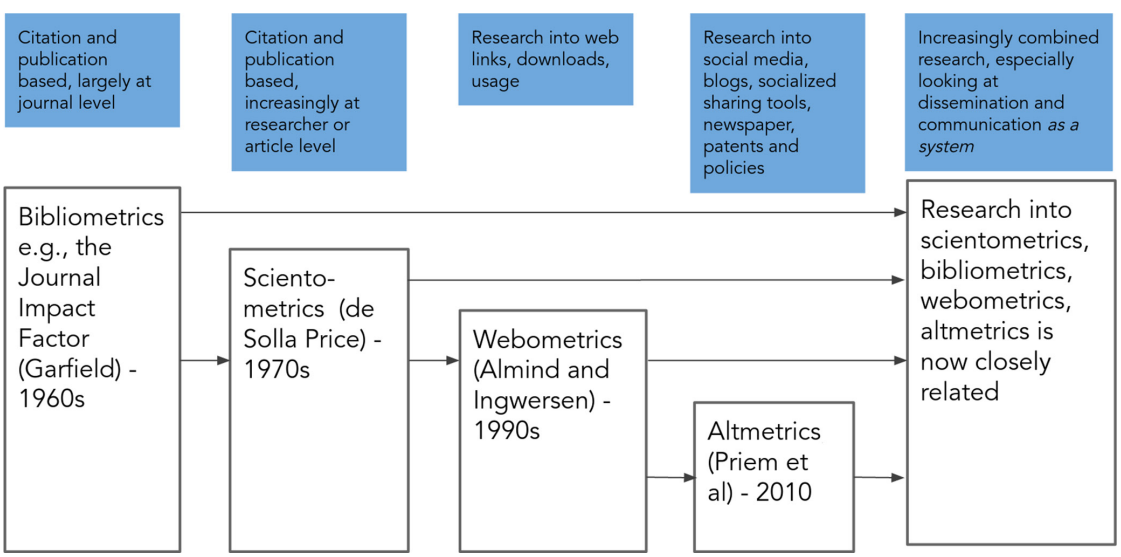

Figure 2. The relationship between altmetrics and other fields of research, along with key influencers and the periods in which they dominated their fields.

Like a kid in an eye-candy shop

The first time my eyes were opened to the power of visualization was shortly after the introduction of the term 'altmetrics'. At the time, I was working at Elsevier Labs, an R&D group, and we'd gotten hold of some web data on research usage. Another colleague had access to Tableau. This, as it turned out, was an intoxicating mix: he called me into the office. "Hey," he said, "want to know what your data looks like?"

I sat at his desk, expecting to see a huge dump of tab-delineated data. Instead, he proudly showed me this multi-dimensional, technicolour exploding star. He grabbed a slider: it shrank and grew before my eyes. I think he could even make it spin. I was utterly gobsmacked, astounded, blown away. It was as if we had just moved from 1950s *Dr Who* – black and white, creaking sets, rubbery tentacles – into the latest *Star Wars*. I gripped the desk. Eventually I breathed. Then I sat back.

It looked beautiful, but what did it mean?

The more we talked, the more I realized that there was danger here, as well as opportunity. My colleague had loaded the data, fiddled around, produced a number of visualizations and called me in as I passed. The data wasn't a thing he needed to understand, it was enough to know that he could do something – *anything* – with it.

If you go to watch a blockbuster film, there might be 20 minutes of visual brilliance set within an otherwise unforgettable film. And that's probably not a bad ratio, as long as those 20 minutes fit into the rest of the story. If you have a powerful visualization: that's great, as long as it's rooted in the research, as long as it fulfils a need, as long as your audience is engaged in the wider plot.

Visualizations don't have to be all-singing, all-dancing to have an impact. In Figure 3, I've used Google Data Studio to convey an extremely important (and quite challenging) idea: that different persona on Twitter, get retweeted (and tend to show a higher tendency towards viral messaging) than others. Doctors and oncologists are retweeted more often than patients (when discussing research), and people involved in pharmaceutical science rather less. Twitter bots (Auto_Tweets in this chart) are useful, but rarely retweeted. The global population has a ratio of 1.2 retweets to every original tweet.

Nothing loses an audience faster than a spectacular display that means nothing. Be ruthless and self-critical. If you can't build a story from a visualization, then you need to start again.

Inherited knowledge

"No man is an island,
Entire of itself,
Every man is a piece of the continent,
A part of the main." – *John Donne*

Visualizations are representations of data (just as data are representations of some notion of the real world), and representations function like metaphors. And like any metaphor, there's a sense in which we use learning from one area to understand another. Visualizations are,

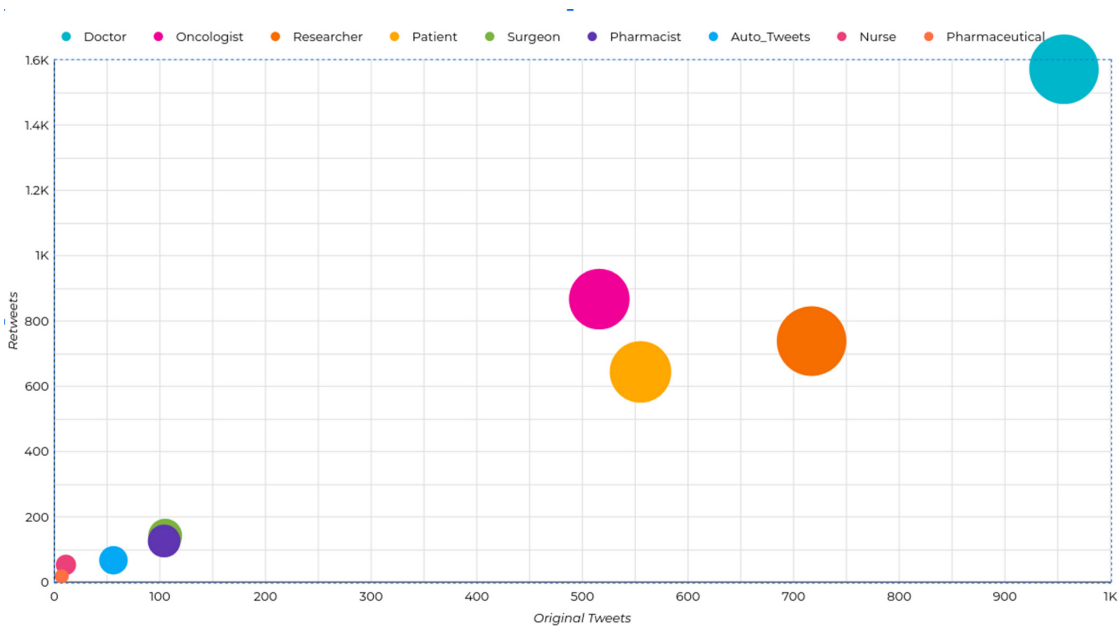

Figure 3. Different personas are retweeted at different rates (data shown for lung cancer chemotherapy research). Dot size represents the number of publications mentioned in tweets. Dashed line represents global relationship.

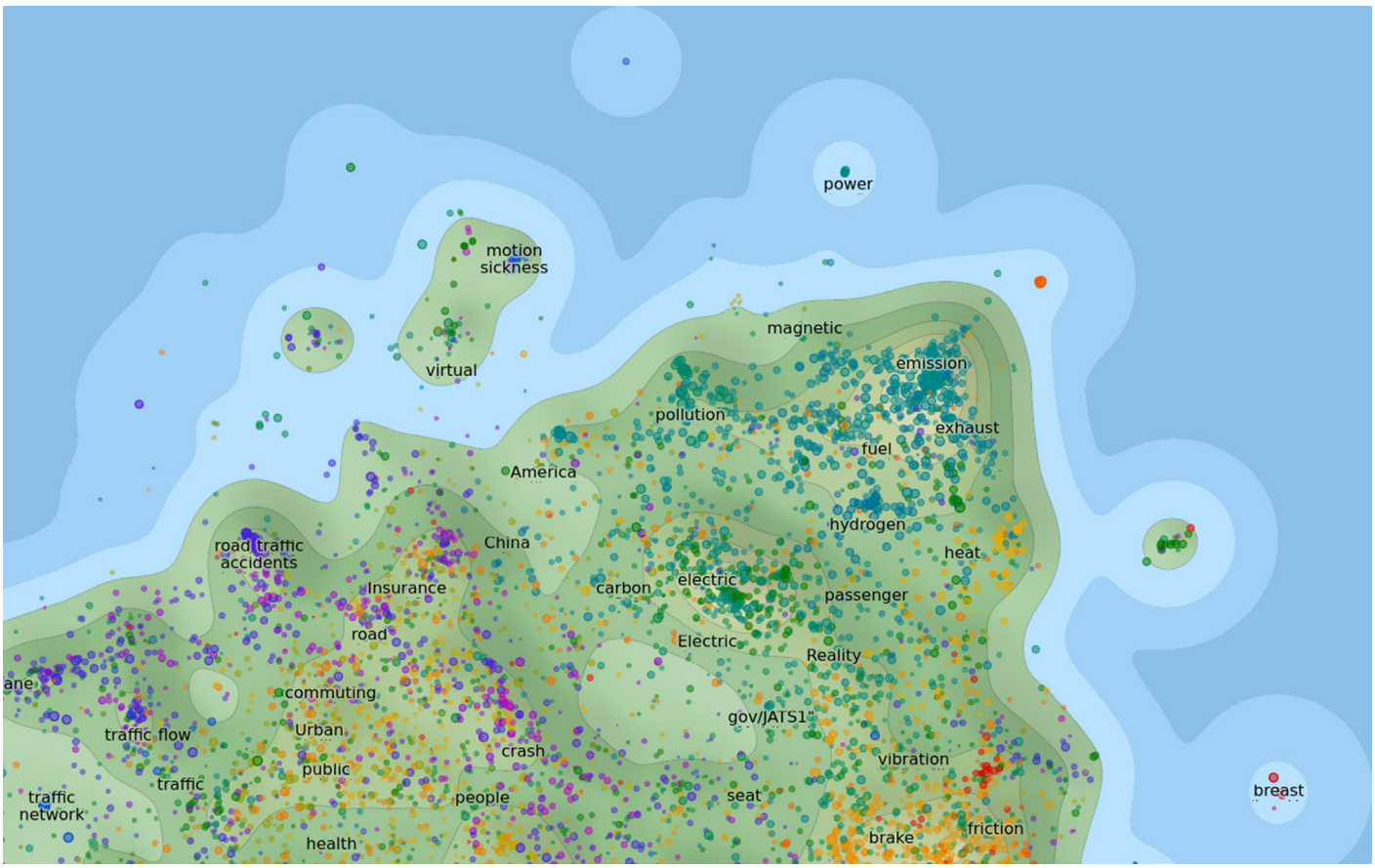

Figure 4. A Lingo4G representation of the Dimensions search "(self-driving OR autonomous) AND (cars OR automobiles)".

therefore, not neutral, their very form relies on inherited knowledge – insights learned in one world are brought into the new.

For example, in Figure 4, I used the excellent application 'Lingo4G' to visualize a Dimensions search on self-driving cars. Under the hood, Lingo4G extracts concepts, applies some complex algorithms to the co-occurrence and proximity of those concepts and creates a number of attractive visualizations; it makes for an engaging way to understand a research landscape.

In this figure, we can see that the current state of research literature has a number of clusters related to our search that in some way represent clusters of research; and these may be identified by the labels that the application automatically generates. These 'hills' are connected in ranges, they are separated by valleys. Surrounding the research topics are seas, and in those seas, there are islands. Some of the seas are deep, some are shallow. To what extent does this visual metaphor work – e.g., can we – conclude that a dot in a 'valley' is in some ways answering a gap in research – perhaps between 'carbon' and 'crash'? How do we explain the 'breast' archipelago on the leftright? Is 'power' really unconnected with the main areas of research?

If I am going to ask you – the audience – to invest time in understanding a visualization, I need to invest my own time in making sure that I understand it, and the underlying data. Typically, I will spend time annotating it, exploring the data, questioning my own assumptions about the relationships and being able to answer awkward questions from the floor.

Be aware of what your meanings might be taken from visualization: be super aware that the more striking a visualization is, the more likely it is to be subject to criticism, and shared widely (and out of context).

Pictures have perspective

It should go without saying that data is not neutral: books such as *Data Feminism* have done much to throw a light on the biases that are frequently incorporated into decisions. Whether it is drug testing on male humans (largely drawn from student populations), planning decisions (using data based on frequent drivers commuting to work) or research trends that only use data from English-speaking universities, we

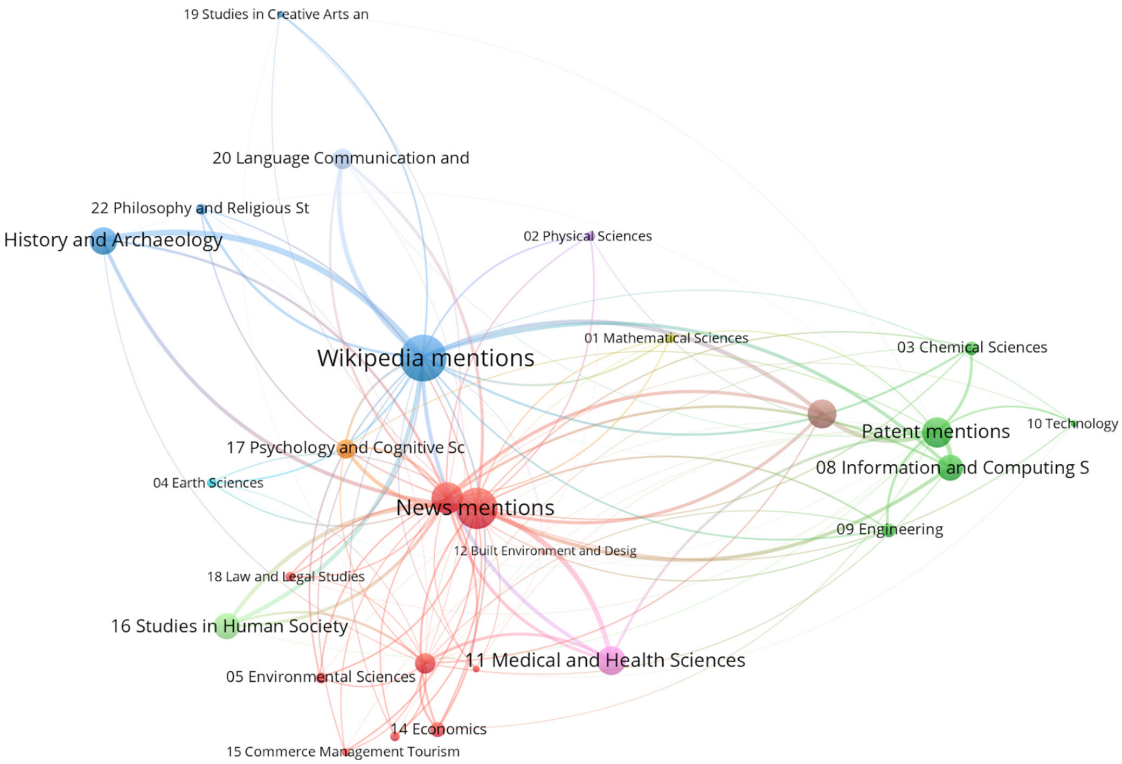

Figure 5. A network diagram exploring the relationship between altmetric citations and subject areas. VOS Viewer's export function has resulted in some cropped names and has omitted 'Blog mentions', being the red dot to the left and slightly above 'News mentions'.

need to be increasingly mindful of what we include and what we exclude, and how we approach these decisions.

Figure 5 shows a network graph that I constructed using VOS Viewer[iii] and Altmetric data, and it shows how different research fields get different rates of attention from various Internet sources. The background to this is that there are significant disciplinary differences between altmetric data, and I wanted to convey these rapidly, to a non-expert audience.

This apparently simple diagram has a number of assumptions:

- First, and most importantly, I excluded Twitter. It's a very large data source that would have dominated the rest of the chart, so I presented on that separately.
- Second, I excluded a number of smaller data sources; I felt they added noise and didn't help my message.
- Third, some data sources were excluded for technical reasons, relating to availability.
- Fourth, some otherwise interesting data sources are now excluded for privacy/commercial reasons. For example, LinkedIn no longer makes its data available.

There are other, broader assumptions relating to the overall notion of what scientific literature consists of. Researchers who publish in languages other than English are very often excluded from the common research infrastructure – whether for cultural reasons, gatekeeper bias, or financial reasons. A common observation among my colleagues who work internationally is that different cultures (whether related to academic discipline, language or country) value books more than research articles, and books are generally under-represented in most databases used for academic evaluation. The implication that follows is that the outputs of, e.g., a Russian academic specializing in fine art, a Brazilian sociologist, a French linguist or the vast majority of scientific production coming from countries such as Cuba literally *can't be counted* by people undertaking research evaluation and often can't be discovered by academics scanning the literature.

Any narrative that you create to support your audience's understanding of your findings will inevitably involve a series of selections. Providing that context, not glossing over it – or over-assuming

[iii]This excellent piece of software is available free from www.vosviewer.com and is often used to visualize citation data (https://app.dimensions.com has a free download function). Here, I constructed the network using Altmetric data export.

universal applicability – is essential to honest storytelling.

Summary

Data visualization is an increasingly common and frequently powerful tool for those of us who have to convey information about research and demonstrate trends across massive (and complex) datasets. However, with great power comes great responsibility: just because we can do something, that doesn't mean it's useful, necessary or intelligent. Data visualization tools don't mean we can forget about interpretation or describing data: they should support intelligent storytelling and enhance our abilities to communicate. ■

Further reading

- Chen, C. (2016) *CiteSpace: A Practical Guide for Mapping Scientific Literature*, Nova, New York.
- CiteSpace. http://cluster.cis.drexel.edu/~cchen/citespace/ - extraordinary visualization tool developed by Professor Chaomei Chen at Drexel. Many powerful images have been produced using CiteSpace; the application is rooted in good scientific research.
- Datawrapper. https://datawrapper.de – free data visualization application that supports Excel-type charts (and more), but more beautiful.
- D'Ignazio, C. and Klein, L. (2020) *Data Feminism*. MIT Press, Cambridge.
- Dimensions. https://app.dimensions.ai – massive research discovery database from Digital Science that offers free (non-commercial) data, with exports for VOS Viewer and CiteSpace.
- Google Data Studio. https://datastudio.google.com – free and powerful data visualization tool from Google, compatible with many data sources.
- Priem, J., Taraborelli, D., Groth, P., et al. (2010) *Altmetrics: A Manifesto.* https://altmetrics.org [Accessed 2 September 2021]
- VOS Viewer. https://www.vosviewer.com – stunning network visualization tool, not only from sources such as Dimensions, but can be used for bespoke data too.

Mike Taylor is Head of Data Insights at Digital Science, where he works with a variety of clients to surface insights drawn from Altmetric and Dimensions data. In his spare time, he is attempting to complete a PhD at the University of Wolverhampton and runs a theatre company in his home town of Oxford.