Research Trends

Volume 1
Issue 38 *Back to the future: 10 years of Scopus and beyond*

Article 6

9-1-2014

A quick look at references to research data repositories

Sarah Huggett
Elsevier

Recommended Citation

Huggett, Sarah (2014) "A quick look at references to research data repositories," *Research Trends*: Vol. 1 : Iss. 38 , Article 6.
Available at: https://www.researchtrends.com/researchtrends/vol1/iss38/6

Section 5:
Behind the data

A quick look at references to research data repositories

Sarah Huggett

While published papers are one of the most visible outputs of the research process, in a way they are only the tip of the iceberg: the research workflow is composed of much more than meets the eye of the external observer (see Figure 1).

Most scholarly research uses data in one guise or another, and recently there have been calls for data to become more systematically visible research output rather than remain a background variable of academic endeavors. For instance, the Force 11 community movement, which aims to support the advancement of scholarly communications, has issued eight Data Citation Principles that stress the importance of data being "considered legitimate, citable products of research" (1). These principles highlight main citation issues, such as access, unique identification, and interoperability and flexibility. Research Trends' curiosity was piqued: could there be a way to estimate the visibility of research data in the published literature?

Methodology

Researchers may make data available in data repositories, and authors may subsequently reference these data in their scholarly outputs. So how could these data citations be analyzed?

One of the challenges mentioned by Force11 is unique identification: researchers may refer to datasets they cite by various names; however, the web addresses of the repositories in which the data reside can be used as reliable identifiers. So first, a list of data repositories was needed; this was extracted from databib (a website describing itself as "a searchable catalog / registry / directory bibliography of research data repositories") in June 2014. This yielded 971 results (see examples in text box) of data repositories in various fields, countries, and of various sizes. Notably nearly half of the listed repositories originate from the USA (see Figure 2).

Examples of data repositories from the databib list:

- 1000 Genomes (Thousand Genomes) (A deep catalog of human genetic variation)
- DataONE (Data Observation Network for Earth)
- Dryad
- Flybase
- Freebase
- Marine Geoscience Data System
- Ontario Data Documentation, Extraction Service and Infrastructure (ODESI)
- Sloan Digital Sky Survey
- TreeBASE
- World Data Center
- WormBase

Second, papers citing these repositories websites needed to be identified. The Scopus advanced search function allows searching the reference fields of papers for websites, which was done for all URLs on the databib list, truncating the addresses and using wildcards as appropriate. The records of the papers identified as containing the URLs in their reference lists were then extracted.

There are two main potential caveats to this approach:

- If an author fails to include the website to the references or mentions the website in the full text but not the references, their papers will not be retrieved by this search method.

- Some of the websites listed by databib are more than just data repositories. If a researcher references the website with a purpose other than data citation, then their paper will still be retrieved by this search method.

Results

This analysis returned 178,909 1996-2014 documents, with a whopping 19% annual growth (CAGR) between 2009-2013, leading to over 30,000 papers in 2013. Most of the documents are articles (113,618 articles with 24% 2009-2013 CAGR), conference papers (37,410 conference papers with 7% 2009-2013 CAGR), and reviews (19,334 reviews with 16% 2009-2013 CAGR) (see Figure 3).

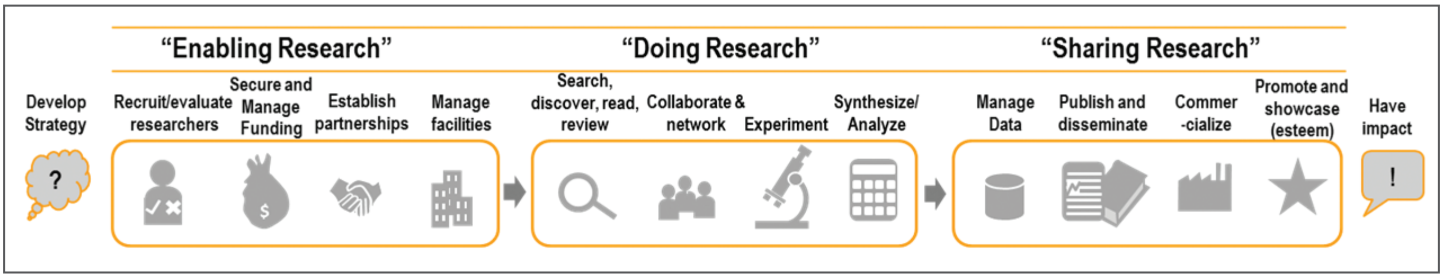

Figure 1: Researcher Workflow. Source: Elsevier's Response to HEFCE's call for evidence: independent review of the role of metrics in research assessment

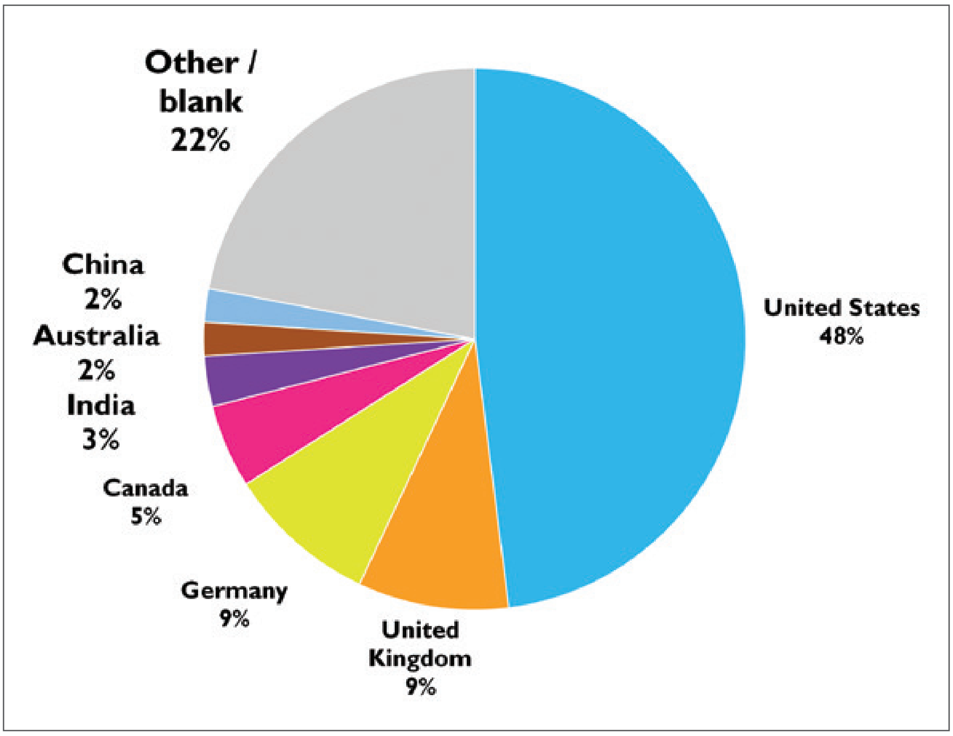

Figure 2: Geographical distribution of data repositories. Source: SciVal

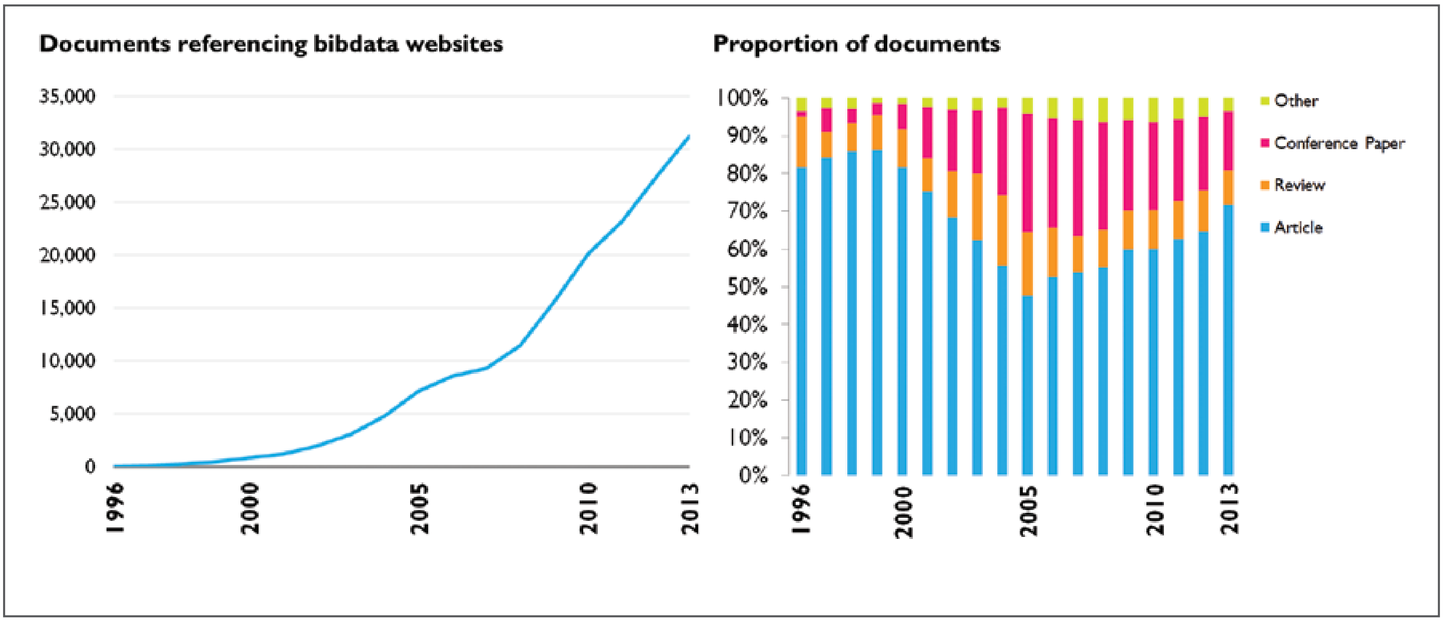

Figure 3: 1996-2013 documents citing bibdata websites. Source: Scopus

These documents received 1,879,964 citations, and a word cloud of 2013 papers' document titles (see Figure 4) shows the preponderance of health-related topics.

Conclusion

The visibility of research data as estimated by references to data repositories in the published literature has seen strong growth in recent years. The topics covered by these papers are preponderantly centered on health-related issues. This topical issue is seeing initiatives aiming to further integrate research data into the more traditional outputs of research that are scholarly communications (1). There are still challenges ahead, in particular regarding unique identification and meta-data integration, which would allow more rigorous and accurate bibliometrics analyses. Nevertheless, with current computational storage capacities and increasing demand from the research community, the future of research data currently appears full of potential promises.

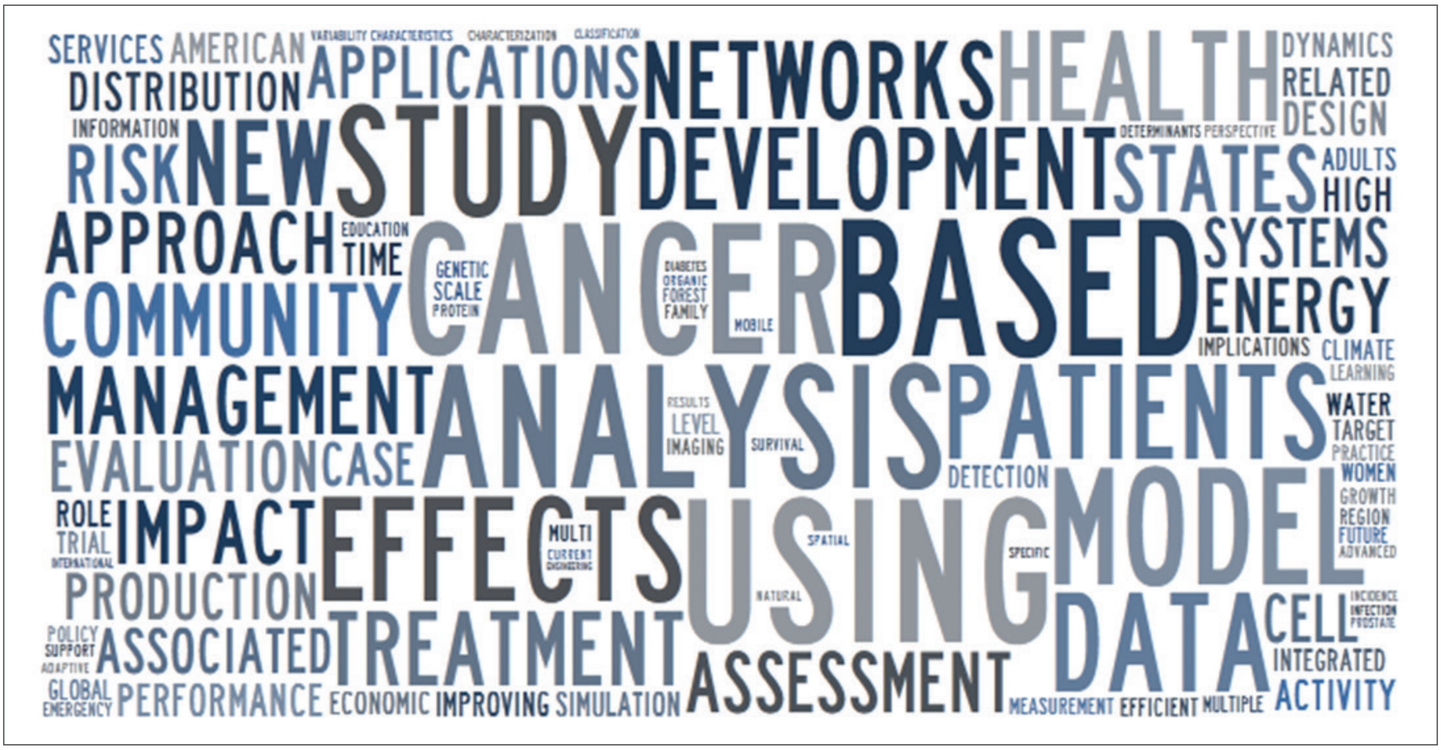

Figure 4: Word cloud of words of titles of 2013 documents citing bibdata websites. Source: Scopus and tagxedo

References:

1. Force 11 - "Joint Declaration of Data Citation Principles", accessed at https://www.force11.org/datacitation in August 2014.