ARTICLE

Dimensions: Bringing down barriers between scientometricians and data

Christian Herzog, Daniel Hook, and Stacy Konkiel

Digital Science

Keywords: bibliometrics, scientometrics, Dimensions, indicator development

ABSTRACT

Until recently, comprehensive scientometrics data has been made available only in siloed, subscription-based tools that are inaccessible to researchers who lack institutional support and resources. As a result of limited data access, research evaluation practices have focused upon basic indicators that only take publications and their citation rates into account. This has blocked innovation on many fronts. Dimensions is a database that links and contextualizes different research information objects. It brings together data describing and linking awarded grants, clinical trials, patents, and policy documents, as well as altmetric information, alongside traditional publications and citations data. This article describes the approach that Digital Science is taking to support the scientometric community, together with the various Dimensions tools available to researchers who wish to use Dimensions data in their research at no cost.

1. AN INTRODUCTION TO DIMENSIONS DATA

We (Digital Science) are honored to contribute to this special issue of *Quantitative Science Studies* by summarizing what Dimensions can offer to the scientometrics research community. We would like to state upfront and as clearly as possible:

Dimensions is available at no cost for scientometric research purposes.

To register for no-cost access, simply fill out the form at https://dimensions.ai/data_access.

For those readers with a little more time, we would also like to provide a conceptual and technical introduction to Dimensions in a bit more detail, answering questions such as the following: What is Dimensions? Why did we develop Dimensions? What sets Dimensions apart from other similar databases? How does Digital Science see collaboration with the scientometrics research community evolving today and in the future?

Dimensions is a database of linked information that describes the research life cycle more completely than any similar system to date. It encompasses awarded grants, publications and their citations, clinical trials, patents, and policy papers.

Dimensions was intentionally built to help to contextualize the research, discovery, and evaluation environments (Hook, Porter, & Herzog, 2018). One of several motivations for Digital Science to build Dimensions was to widen the reach of research evaluators beyond publication citation analysis (Figure 1). At the same time, we wanted to offer the community the capacity to perform bibliometric analyses that evolve both academic discussions and

an open access journal

Citation: Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387–395. https://doi.org/10.1162/qss_a_00020

DOI: https://doi.org/10.1162/qss_a_00020

Received: 21 June 2019
Accepted: 05 December 2019

Corresponding Author:
Stacy Konkiel
s.konkiel@digital-science.com

Handling Editors:
Ludo Waltman and Vincent Larivière

The MIT Press

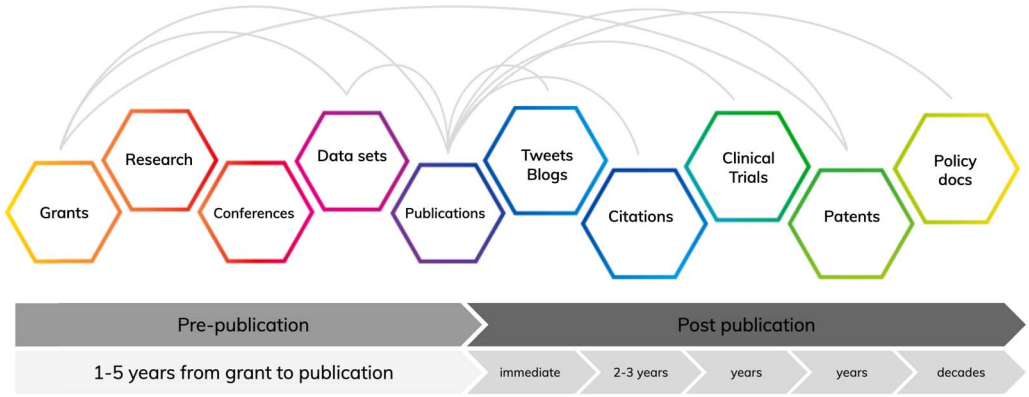

Figure 1. An illustration of the scholarly communication life cycle, from the moment that research is funded to its downstream impacts upon patents and public policy.

practical impacts beyond the limitations of databases that focused solely on a subset of the researcher workflow.

At the time of writing, the Dimensions database included over 105 million publications and their citations, along with content types that illustrate the larger information life cycle: from funding of an idea (via grants data for 5 million funded projects), to the eventual publications that result from such support, to the impact of the publications (illustrated through the 1.1 billion citations to 100 million research outputs and altmetrics for 11 million research outputs, respectively), to the artifacts of the real-world application of research (more than 497,000 clinical trials, 39 million patents, and 434,000 policy documents). Taken together, these data points can paint a richer, fuller picture of research than was previously available (Figure 2).

Dimensions sources data from a number of organizations. Indices such as Crossref and PubMed Central serve as a "backbone" for publication data. To this backbone is grafted data derived from full-text access to more than 75 million of the 105 million books and articles. These full-text publications are mined to enhance metadata, citations, funding acknowledgements, and

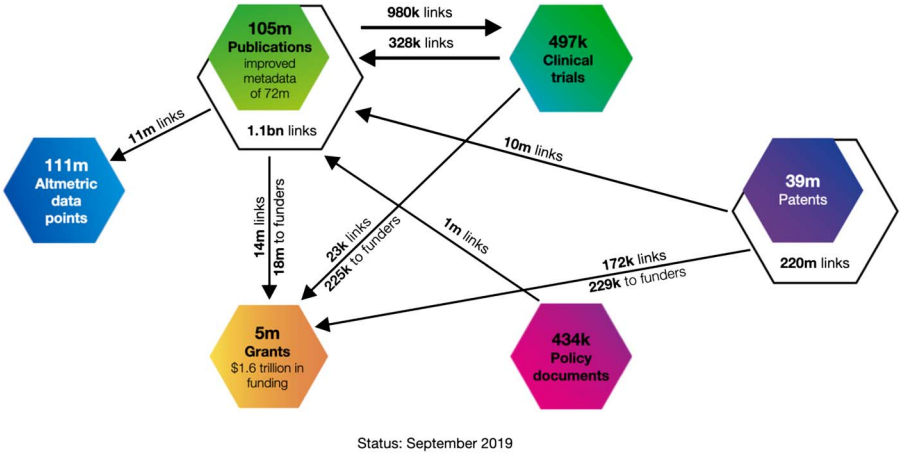

Figure 2. An illustration of the linkages between altmetric data, citations, publications, grants, clinical trials, patents, and public policy documents in Dimensions.

so on. In addition, initiatives and resources, such as OpenCitations[1] and I4OC[2], clinical trial registries, funding agencies, and openly available public policy data, supplement the richness of the Dimensions database. Finally, Digital Science companies, such as Altmetric and IFI Claims, contribute further data to create the Dimensions database. We have linked and "harmonized" millions of records, using open standards such as GRID[3] and ORCID,[4] to make the data as interoperable as possible. A fuller description of the sources and process is given in Hook, Porter, and Herzog (2018).

The Dimensions data can be accessed in a number of ways: from online interfaces that offer rich, contextual search and data visualization (Dimensions, Dimensions Plus, and Dimensions Analytics), to powerful APIs that allow users to search and aggregate programmatically across the entire Dimensions database with precision and retrieve indicators for millions of publications, grants, etc. (Dimensions API and Dimensions Metrics API), to bulk data access that allows "power users" and research teams to perform high-powered analyses across the entire Dimensions database, at scale.

As Dimensions is a relatively young database, there are some known caveats to its data and what can be done with it. Dimensions does not yet index the entirety of the research ecosystem and may never do so. However, our aim is to provide researchers and others in the ecosystem with a dependable, identifier-driven, verifiable index of objects that are important to them in support of the widest variety of use cases.

We have recently added preprints as a content type in Dimensions and are still working to ensure that the methodology that we use to integrate these records into the core data set makes sense for all the relevant use cases, from personal (e.g., creating CVs), to institutional (e.g., collaboration benchmarking), to publisher (e.g., assessing feasibility of new journals), and bibliometric or scientometrics (e.g., calculation of global citation benchmarks).

In the normal course of running, we add new grant and publication data every few months, and have plans to add data from new patent jurisdictions and policy makers over time. Additionally, we continue to work with publishers to improve the coverage and quality of their data in Dimensions.

Over time, we hope that Dimensions will become a data set that truly represents research across regions and disciplines. To that end, we continue to increase our coverage of non-English language content, as well as working to include more outputs that represent humanities and social sciences research. We are also working toward integrating more monograph content, which we know to be so much more important than journal articles in many disciplines.

Dimensions data can only be as "open" as the data sources that we draw upon allow it to be: We work closely with our content providers to allow data to be used in as many contexts as possible. However, much of the data comes with some restrictions, and this can have an effect upon data sharing. For example, applications such as open source analytics dashboards or archiving of significant segments of Dimensions data in open data repositories for scientometrics studies are use cases that we cannot currently support without discussion. We consider each request carefully and work with all parties involved (the user, the data provider, and our product and sales teams) to come to an agreeable solution.

Dimensions data is constantly improving: Since Dimensions launched in 2018, we have received valuable feedback from the scientometrics community on the quality of the Dimensions data, including our field classification data (Bornmann, 2018; Orduña-Malea &

[1] https://opencitations.net/
[2] https://i4oc.org/
[3] https://www.grid.ac
[4] https://orcid.org/

Delgado-López-Cózar, 2018). This has helped us to improve Dimensions, and we welcome further constructive feedback from the community to this effect.

Above all else, we aim for Dimensions to be an innovative and comprehensive data source and a responsive and thoughtful member of the research community.

2. THE PHILOSOPHY DRIVING DIMENSIONS

The Digital Science team developed Dimensions with three major questions in mind:

1. How can we make publication and citation data more accessible, both broadly for the benefit of the academic community and specifically for the benefit of researchers who focus on research on research?
2. How can we move beyond simply lamenting the incorrect use of prevailing research impact metrics (e.g., Journal Impact Factor) to instead provide a broader, richer, and more connected view of the research life cycle and all its impacts?
3. How can we ensure a "division of powers" in indicator and metric development, by offering, for free, aggregated and curated data to the research community to develop new indicators, rather than developing proprietary metrics?

To achieve our first goal, we decided to do two things: first, to make a version of Dimensions[5] that focuses on publication and citation data freely available for personal use, specifically without the need to register; and second, to make access to the entirety of Dimensions data available at no cost for scientometric research purposes, in a wide variety of formats, to enable large-scale analysis on scholarly communication trends.

To support our goal of providing a broader, richer view of the research life cycle and all its impacts, we have invested (and are investing still) in a diverse set of efforts to broaden the scope of Dimensions beyond existing bibliometric data. For example, we continue to work to integrate more grants, patents, clinical trials, and policy documents, not only aggregating millions of previously siloed records but also creating links between these records based on increasing occurrence of persistent identifiers, as well as AI-based techniques, and by mining relationships referred to in full text. This makes Dimensions an increasingly interesting data set that is ripe for use in scientometric analysis and gaining evaluative insight. Exploration of the deep and multifaceted linkages already provides a colorful view of the research process and the results that it produces (Bode et al., 2018). Moreover, we have decided to take an "inclusive" approach to the publications we index in Dimensions. We believe that Dimensions should be a comprehensive data source, not a judgment call, and so we index as broad a swath of content as possible and have developed a number of features (e.g., the Dimensions API, journal list filters that limit search results to journals that appear in sources such as Pubmed or the 2015 Australian ERA[6] journal list) that allow users to filter and select the data that is most relevant to their specific needs.

To support the third goal that we have highlighted here, we have involved scientometricians in the development of Dimensions since the very beginning, as part of a larger group of 100+ development partners that also include research organizations and funders.

When we saw the first winds of change in the community regarding the perception of research evaluation, and the rethinking of incentives brought about by DORA and the Open

[5] https://app.dimensions.ai
[6] https://www.arc.gov.au/excellence-research-australia

Data movement, we knew that Dimensions should not be about rankings or the creation of new metrics. We believe that it is the role of Digital Science as a data provider to continually add relevant data sets to Dimensions and to enhance those data by establishing more connections between the different types of research objects in Dimensions. We felt strongly that these data should be provided at no cost to the scientometrics research community. We also believe that we should not develop indicators, because to do so would stop us from being a neutral party when it comes to supporting the indicators of others. Indicators, in our opinion, should be developed and owned by the research community to ensure a multifaceted, rich, open, and neutral perspective on the research process and its results.

In June 2018, we launched a year-long pilot phase to provide no-cost Dimensions access for scientometrics research. This pilot allowed us to test our approach, streamline the application and contract process, and optimize our internal review and approval workflows to provide quick, easy, no-cost access to Dimensions data. It also allowed us to launch a Scientometrics User Group for researchers, to provide resources and support to those using the data. As of September 2019, we have granted no-cost access to more than 100 researchers and research teams and have a growing community of more than 200 User Group members. There are still many more steps to be taken in support of our goal, one being this communication in *Quantitative Science Studies* to reach a broader audience with our offer of no-cost data access for scientometricians.

We are now entering the next chapter of the Dimensions Scientometrics User Group by announcing a partnership with the International Society for Scientometrics and Informetrics (ISSI). ISSI members pursuing personal, noncommercial scientometrics research projects will have the ability to receive no-cost Dimensions access as a benefit of their ISSI membership. In turn, Dimensions will work with ISSI to scale and broaden access to the data to a much greater number of scientometricians worldwide.

3. USING DIMENSIONS DATA FOR LARGE-SCALE, REPRODUCIBLE ANALYSES

We created Dimensions to allow for the kinds of large-scale analyses that scientometrics researchers and research analysts typically pursue. In fact, we developed the Dimensions API specifically to meet this use case—the API not only allows data retrieval but also provides a business logic layer[7] that allows researchers to truly "work" with the data at scale, such as by developing composite indicators.

Digital Science also works toward supporting reproducible scientometrics research (Herzog, Hook, & Adie, 2018), which is why all Dimensions no-cost agreements allow for scientometrics researchers to retain a copy of their data to reproduce their analyses. However, although we would also like to fully support open data by allowing researchers to archive their scientometric data in open repositories such as Figshare, we are ourselves restricted by the same legal constraints that make bulk data access relatively rare.

4. REQUIREMENTS FOR FREE DIMENSIONS DATA ACCESS

As part of our commitment to openness and accessibility, we made the Dimensions publications core search freely available to everyone[8]. The functionality in the free Dimensions product has been described as at least comparable with subscription offerings, in terms of the scope

[7] We developed a simple, custom language for working with the Dimensions API, called *Dimensions Search Language*, which has a number of business logic functions embedded. To learn more, visit our API documentation https://docs.dimensions.ai/dsl/index.html.

[8] See https://app.dimensions.ai to use the free version of Dimensions

of publications indexed, keyword search functionalities, and contextual information provided (Harzing, 2019; Thelwall, 2018; Visser, van Eck, & Waltman, 2019). Additionally, Dimensions includes

- Full-text search capabilities for more than 72 million of the 105 million publications records in the system (and title and abstract search for all articles);
- Search results that include detailed author affiliation data and lists of citing articles; and
- Supporting grant information, altmetric data, citing patents, clinical trials, and policy documents as well as links to ancillary data.

Overall, the free version of Dimensions offers the full context of all indexed publications in one place. The full version of Dimensions differs from the free version in the number of analyses that a user can perform, access to the API, and ability to perform searches across our clinical trial, grants, patents, and public policy indices. Users who need access to large-scale publications data or patent, grants, and other data for noncommercial analysis can apply for no-cost access to the full version of Dimensions.

Given that no-cost Dimensions data access for scientometric research purposes is part of our overall philosophy, we have developed a quick and lean process that allows us to quickly provide access to Dimensions data. The only condition for approval is that the data and tools are only to be used for the agreed upon purpose.

Any researcher can apply for no-cost Dimensions access for use of Dimensions data in their personal, noncommercial scientometrics research projects. Depending on researchers' needs, they can receive access to the Dimensions Analytics database, the Dimensions Analytics API, the Dimensions Metrics API, or a combination of the three tools.

Applied scientometric use cases (e.g., consulting, institutional analysis, and reporting) are not eligible for no-cost bulk data access, because to support a fair and sustainable Dimensions business model requires that organizations support our efforts to maintain and operate the platform with their license contribution for applied or commercial use.

Research groups conducting high-volume analyses may request access to Dimensions data in bulk; these requests are considered on a case-by-case basis.

Making Dimensions data available for noncommercial scientometrics research is aligned and engrained in the values that drive Digital Science overall, chief of which are community-mindedness and supporting open research. We intend to offer no-cost data access in perpetuity, barring restrictions from copyright holders that might impede our ability to allow no-cost access in the future. We want to make very clear we see making the data available in such a way as one of the core missions of Digital Science, even though we are organized as a commercial group, because Digital Science was founded upon and is driven by open research values.

5. WHO IS CURRENTLY USING DIMENSIONS DATA IN SCIENTOMETRICS RESEARCH?

To date, Dimensions has been contacted by scientometricians from diverse backgrounds who are studying a variety of research topics. These topics run the gamut: from how research funding affects publication rates in the developing world, to how diversity of researcher background can influence the quality of published research, to how national coauthorship trends can benefit scholarly impact and online attention for research. Researchers and analysts using Dimensions data include

- Michael Head, University of Southampton (UK), who has used Dimensions data as part of the Research Investments in Global Health study (Brown & Head, 2018), an ongoing academic analysis assessing levels of health-related research funding compared to factors such as the global and national burdens of disease;
- Janne Seppänen, University of Jyväskylä (FI), who is using Dimensions data to calculate an open, cocitation network citation rate percentile rank similar to the Relative Citation Ratio (an indicator developed by the U.S. National Institutes of Health [NIH] that can be used to understand the citation performance of an NIH-funded paper compared to its cocitation network) (Hutchins, Yuan, Anderson, & Santangelo, 2016; Seppänen, 2018);
- Pablo García-Sánchez and Manuel Jesús Cobo, University of Cadíz (ES), who are using interlinked publication, author, and citation data to map the impact of Andalusian universities (García-Sánchez & Cobo, 2018); and
- The teams behind VOSviewer[9] and CiteSpace[10], two noncommercial scientometric data visualization software packages that can create informative network graphs and other visualizations based on Dimensions publication data.

Beyond these examples, there are many more researchers interrogating our rich, linked data in innovative and creative ways. Although we are happy to support the use of Dimensions data in any kind of scientometric study, we are especially keen to offer Dimensions data and support to researchers who are developing and testing research impact indicators.

6. WHAT WE HOPE TO LEARN IN RETURN

Though we do expect to learn a great deal through partnering with the scientometrics community, it is not only what we want to learn that matters. Our hope is to establish a long-term open cooperation where we can play our part in supporting a vibrant research community that is developing increasingly powerful methods to understand and support researchers and research. We believe that we can provide access to interesting data that we aggregate and curate, and where scientometricians can seize the opportunity to work with these data to provide robust, community-owned indicators and tools to better map new frontiers of research.

To date, this cooperation has allowed the mutual sharing of our experiences and advice, which has already resulted in a number of new scientometrics publications and presentations, as well as improvements to Dimensions' data quality and coverage, improvements to our in-house machine learning techniques, and the release of several community-created, open source libraries for working with the Dimensions API. We are thankful for the feedback and hints we received so far, working together to improve Dimensions' data scope and quality. We look forward to continuing that journey!

At the end of the day—as Digital Science is made up of a diverse team of former scientists, librarians, publishers, hackers, and tinkerers—we are mostly just excited to see how the community uses Dimensions data to drive new discoveries and insights around scholarly communication that we could not have imagined ourselves.

7. FREE DATA ACCESS AND THE DIMENSIONS SCIENTOMETRICS USER GROUP

If you are interested in free Dimensions data access for your scientometric research projects, you can simply fill out the form at https://dimensions.ai/data_access.

[9] https://www.vosviewer.com/

[10] http://cluster.cis.drexel.edu/~cchen/citespace/

In return, we ask that researchers commit to

- Acknowledging their use of Dimensions data in all related publications, presentations, posters, and software;
- Sharing their Dimensions-related research with us, so we can share it with others in the community;
- Reaching out to collaborate, allowing us to add context to Dimensions-related research findings; and
- Helping us improve Dimensions by suggesting enhancements and pointing out data issues and inconsistencies—this feedback improves Dimensions data for all who use it.

When you fill out the no-cost data access application form, we will then follow up with more information about data access terms and conditions (basically, how you can and cannot use and share Dimensions data; we need to explain this to keep our legal department happy), which you will need to accept to receive a Dimensions account. Once we have your agreement on file, our team will review your application and, if you qualify for access, get you set up with the Dimensions credentials you need to begin your research.

We invite members of the community to join the Dimensions Scientometrics User Group[11]. Members can receive support for their Dimensions-related research projects provided via email, office hours with Dimensions data scientists, and in-person meetings at conferences. We host regular community calls and events that offer networking and collaboration opportunities with Digital Science's experts and the discipline's top researchers.

8. SUMMARY

If research evaluation is to move forward, then the scientometric research that underpins research evaluation needs also to move forward. We believe that the task of data providers such as Dimensions should be to lower the barriers that make it difficult for scientometricians to find and work with data at scale. Since our launch, the Dimensions team has focused on making such data easier to access and analyze, through providing integrated, easy-to-use tools; offering no-cost access for researchers who need it; and making billions of linked data points available for analysis in a single database. Going forward, we look forward to working with the community to find ways to lower barriers to scientometric reproducibility, as well.

COMPETING INTERESTS

The authors of this paper are Digital Science employees. Digital Science runs Dimensions, the database discussed in this article.

REFERENCES

Bode, C., Herzog, C., Hook, D., & McGrath, R. (2018). A guide to the Dimensions data approach. https://doi.org/10.6084/m9.figshare.5783094.v7

Bornmann, L. (2018). Field classification of publications in Dimensions: A first case study testing its reliability and validity. *Scientometrics*, *117*(637). https://doi.org/10.1007/s11192-018-2855-y

Brown, R. J., & Head, M. G. (2018). Sizing up pneumonia research. https://doi.org/10.6084/m9.figshare.6143060.v1

García-Sánchez P., & Cobo, M. J. (2018). Measuring the impact of the international relationships of the Andalusian universities using Dimensions database. In H. Yin, D. Camacho, P. Novais, & A. Tallón-Ballesteros (Eds.), *Intelligent Data Engineering and Automated Learning—IDEAL 2018* (pp. 138–144). Switzerland: Springer Nature. https://doi.org/10.1007/978-3-030-03496-2_16

Harzing, A. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft

[11] To join the Dimensions Scientometrics User Group, fill out the form at https://ds.digital-science.com/DimensionsSUG.

Academic, Scopus and the Web of Science? *Scientometrics*, *120*(1). https://doi.org/10.1007/s11192-019-03114-y

Herzog, C., Hook, D. W., & Adie, E. (2018). Reproducibility or producibility? Metrics and their masters. Paper presented at STI 2018, Leiden, the Netherlands. Retrieved from https://openaccess.leidenuniv.nl/handle/1887/65257

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, *3*(23). https://doi.org/10.3389/frma.2018.00023

Hutchins, B. I., Yuan, X., Anderson, J. M., Santangelo, G. M. (2016). Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLOS Biology*, *14*(9), e1002541. https://doi.org/10.1371/journal.pbio.1002541

Orduña-Malea, E., & Delgado-López-Cózar, E. (2018). Dimensions: Re-discovering the ecosystem of scientific information. *El Profesional de la Información (The Information Professional)*. *27*(2), 420–431. https://doi.org/10.3145/epi.2018.mar.21

Seppänen, J. T. (2018). Conservation Biology is not a single field of science: How to judge citation impact properly. Paper presented at 5th European Congress of Conservation Biology, Jyväskylä, Finland. https://doi.org/10.17011/conference/eccb2018/108222

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics*. *12*(2), 430–435. https://doi.org/10.1016/j.joi.2018.03.006

Visser, M., van Eck, N. J., & Waltman, L. (2019). Large-Scale comparison of bibliographic data sources: Web of Science, Scopus, Dimensions, and Crossref. Paper presented at 17th International Conference on Scientometrics and Informetrics (ISSI 2019), Rome, Italy.