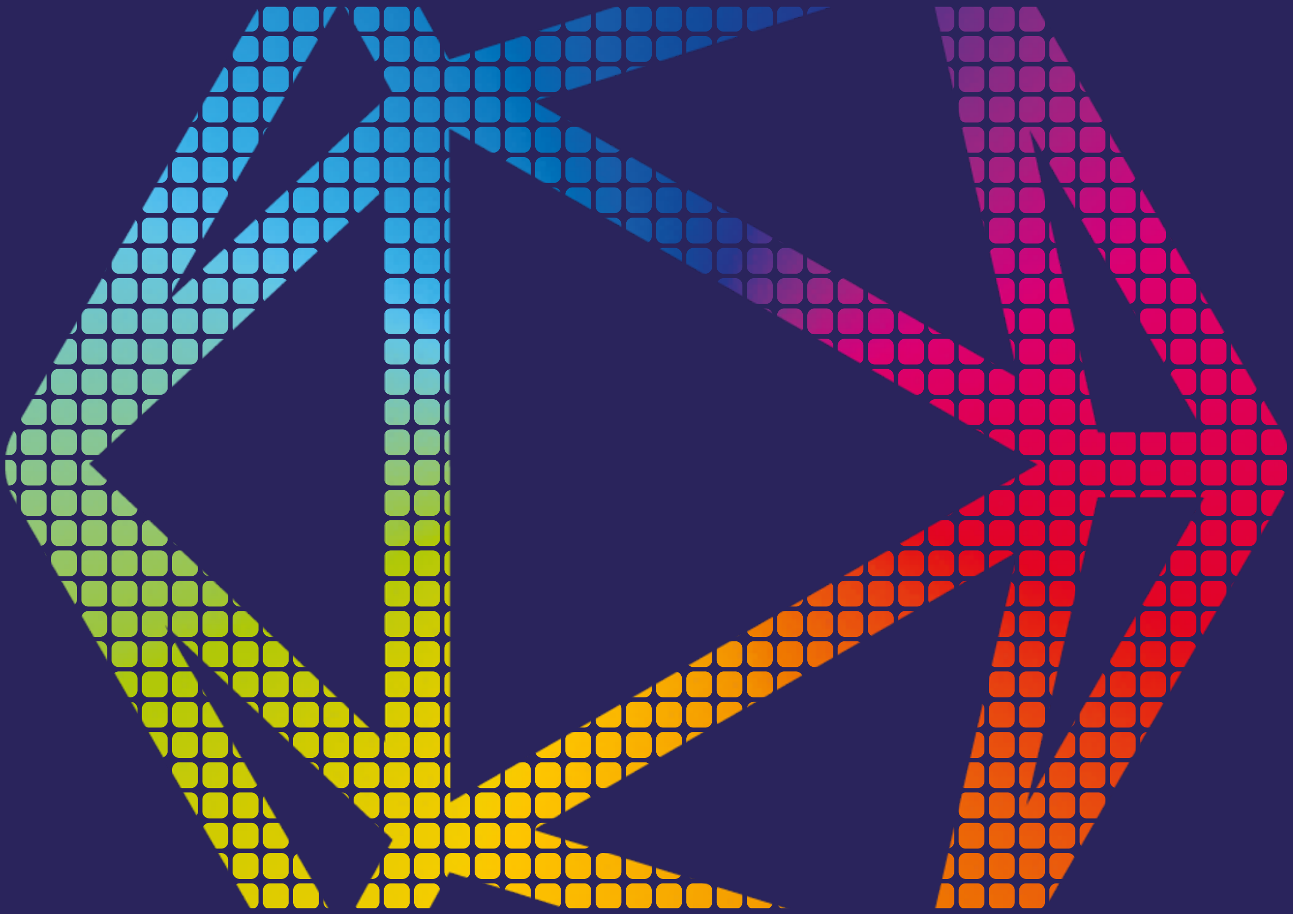


Dimensions Report

A Guide to the Dimensions Data Approach

A collaborative approach to creating a modern infrastructure for data describing research: where we are and where we want to take it

Christian Bode, Christian Herzog, Daniel Hook, Robert McGrath, & Alex Wade

JULY 2023

DIGITAL science

Dimensions

About Digital Science

Digital Science is a technology company working to make research more efficient. We invest in, nurture and support innovative businesses and technologies that make all parts of the research process more open and effective. Our portfolio includes the admired brands Altmetric, Dimensions, Figshare, GRID, IFI CLAIMS, MetaPhacts, Overleaf, ReadCube Papers, Ripeta, Scismic, Symplectic and Writefull. We believe that together, we can help research make a difference. Visit www.digital-science.com

About Dimensions

Dimensions is a modern, innovative, linked research knowledge system that re-imagines discovery and access to research. Developed by Digital Science in collaboration with over 100 leading research organizations around the world, Dimensions brings together grants, publications, citations, alternative metrics, clinical trials, patents and datasets to deliver a platform that enables users to find and access the most relevant information faster, analyze the academic and broader outcomes of research, and gather insights to inform future strategy. Visit Dimensions' website at https://dimensions.ai and find us on Twitter @DSDimensions.

Acknowledgments

We are grateful to all contributors and would like to thank our development team for their time and effort in extracting the data to support this report.

This report has been published by Digital Science which is part of Holtzbrinck, a global media company dedicated to science and education.

Digital Science, 6 Briset Street, London EC1M 5NR, UK.
info@digital-science.com

DOI: 10.6084/m9.figshare.5783094
This report was created in Overleaf

Contents

1 A modern linked research landscape

100+ development partners and Digital Science

Dimensions was created in response to two significant constraints for Digital Science and its development partners. The first was that existing solutions sought to understand the research landscape solely through the lens of publication and citation data. The second was the way that existing solutions exposed their data. Much of the publications research graph had been locked away in proprietary applications, which constrained how the information could be used, including a lack of workable APIs. Where proprietary data existed, there were significant data holes, making the data less useful for core use cases.

Making publication and citation data freely available

To address these constraints and to stimulate innovation to support research, we worked with more than 100 development partners (research organisations and funders) to realise an integrated database covering the entire research process from funding to research, from publishing of results through attention, both scholarly and beyond, to commercial application and policy making - consistently linked in multiple dimensions.

Empowering the research community

We wanted Dimensions to be transformative. A key part of this vision is that Digital Science makes available, without charge, publication citation data via the Dimensions web application (https://app.dimensions.ai) and via an API - the metrics are available via the open Dimensions Metrics API and the Dimensions Badges (https://badge.dimensions.ai) - in both cases for non-commercial purposes.

Does it support your use case?

The current vogue in research evaluation promotes the use of metrics to cope with the vast quantities of material being evaluated. It is clear that a more open data source compatible with more open publications, more open evaluation frameworks and more open metrics are needed. Dimensions aims to be a system that helps the academic community own the formulation and development of metrics that tell the best stories and give the best context to a piece, or a body, of research.

Quick facts on Dimensions - total record count and more

Content type	Items indexed
Publications	137 million
Grants	6.8 million
Patents	155 million
Datasets	12 million
Clinical Trials	790,000
Policy Documents	884,000
GRAND TOTAL	312 million

We will improve it together!

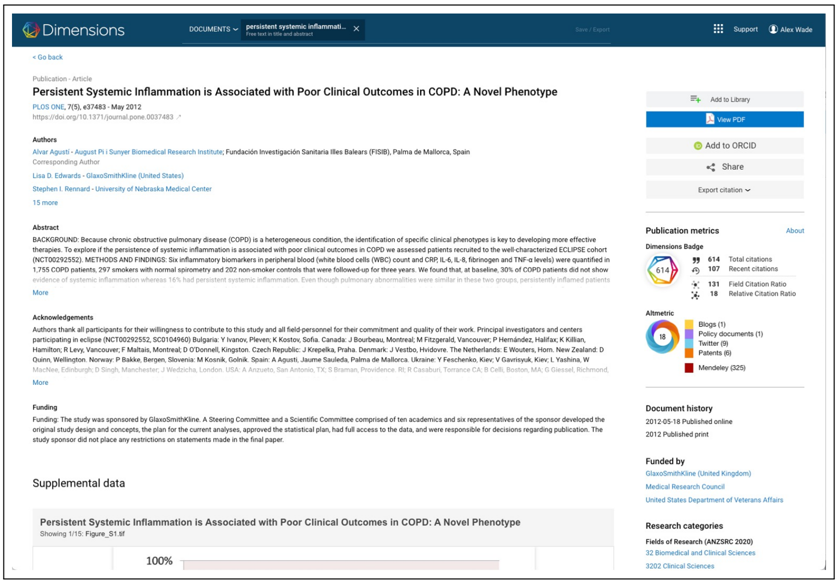
This document provides an overview of the Dimensions data. Please reach out to our team if you want to discuss whether the content scope and coverage can help in your specific situation and use case. We will continue to develop Dimensions with input from research community - any feedback is welcome. Please contact us at info@dimensions.ai.

2 Linking and enriching the data

Linked and integrated data from multiple sources is core to Dimensions, providing an integrated view that enables novel insights. The following sections provide a quick overview of the approaches we have taken.

The links between grants, publications, clinical trials, patents and policy documents are key

Dimensions creates these linkages with a data-driven, machine learning and AI-based approach, automatically extracting information to create connections. The content and enrichment pipeline is automated, allowing us to provide publication and citation data to researchers for free, and to research institutions at realistic cost levels.



Publication - Article

Persistent Systemic Inflammation is Associated with Poor Clinical Outcomes in COPD: A Novel Phenotype

PLOS ONE, 7(5), e37483 - May 2012
https://doi.org/10.1371/journal.pone.0037483

Authors

Alvar Agustí - August Pi i Sunyer Biomedical Research Institute; Fundación Investigación Sanitaria Illes Balears (FISIB), Palma de Mallorca, Spain
Corresponding Author

Lisa D. Edwards - GlaxoSmithKline (United States)

Stephen I. Rennard - University of Nebraska Medical Center

15 more

Abstract

BACKGROUND: Because chronic obstructive pulmonary disease (COPD) is a heterogeneous condition, the identification of specific clinical phenotypes is key to developing more effective therapies. To explore if the persistence of systemic inflammation is associated with poor clinical outcomes in COPD we assessed patients recruited to the well-characterized ECLIPSE cohort (NCT00292552). METHODS AND FINDINGS: Six inflammatory biomarkers in peripheral blood (white blood cells (WBC) count and CRP, IL-6, IL-8, fibrinogen and TNF-α levels) were quantified in 1,755 COPD patients, 297 smokers with normal spirometry and 202 non-smoker controls that were followed-up for three years. We found that, at baseline, 30% of COPD patients did not show evidence of systemic inflammation whereas 16% had persistent systemic inflammation. Even though pulmonary abnormalities were similar in these two groups, persistently inflamed patients

Acknowledgements

Authors thank all participants for their willingness to contribute to this study and all field-personnel for their commitment and quality of their work. Principal investigators and centers participating in eclipse (NCT00292552, SC0104960) Bulgaria: Y Ivanov, Pleven; K Kostov, Sofia. Canada: J Bourbeau, Montreal; M Fitzgerald, Vancouver; P Hernández, Halifax; K Killian, Hamilton; R Levy, Vancouver; F Maltais, Montreal; D O'Donnell, Kingston. Czech Republic: J Krepelka, Praha. Denmark: J Vestbo, Hvidovre. The Netherlands: E Wouters, Horn. New Zealand: D Quinn, Wellington. Norway: P Bakke, Bergen, Slovenia: M Kosnik, Golnik. Spain: A Agusti, Jaume Sauleda, Palma de Mallorca. Ukraine: Y Feschenko, Kiev; V Gavrisyuk, Kiev; L Yashina, W MacNee, Edinburgh; D Singh, Manchester; J Wedzicha, London. USA: A Anzueto, San Antonio, TX; S Braman, Providence. RI; R Casaburi, Torrance CA; B Celli, Boston, MA; G Giessel, Richmond,

Funding

Funding: The study was sponsored by GlaxoSmithKline. A Steering Committee and a Scientific Committee comprised of ten academics and six representatives of the sponsor developed the original study design and concepts, the plan for the current analyses, approved the statistical plan, had full access to the data, and were responsible for decisions regarding publication. The study sponsor did not place any restrictions on statements made in the final paper.

Supplemental data

Persistent Systemic Inflammation is Associated with Poor Clinical Outcomes in COPD: A Novel Phenotype
Showing 1/15: Figure_S1.tif

100%

Publication metrics

Dimensions Badge

614

614	Total citations
107	Recent citations
131	Field Citation Ratio
18	Relative Citation Ratio

Altmetric

18

Blogs (1)
Policy documents (1)
Twitter (9)
Patents (6)
Mendeley (325)

Document history

2012-05-18 Published online
2012 Published print

Funded by

GlaxoSmithKline (United Kingdom)
Medical Research Council
United States Department of Veterans Affairs

Research categories

Fields of Research (ANZSRC 2020)
32 Biomedical and Clinical Sciences
3202 Clinical Sciences

Figure 1: An example of a publication record in Dimensions with links to related content types

Example "Persistent Systemic Inflammation is Associated with Poor Clinical Outcomes in COPD: A Novel Phenotype" (DOI 10.1371/journal.pone.0037483	
Supplemental data	15
Associated datasets	4
Publication references	44
Supporting grants	2
Clinical trial references	1
Publication citations	620
Patent citations	6
Policy document citations	1
Altmetric Attention Score	18

2.1 Full-text index — enabling deep discovery

Full-text indexing - real discovery instead of missing relevant information

Our approach to indexing the full text makes publications and patents much more discoverable. Full-text search is currently available for over 99 million publications and 110 million patents in Dimensions, and covers the entirety of each publication, from title to references. For example, a search for the term 'CRISPR' in just titles and abstracts brings back about

60,000 publications and 13,800 patents, while a search using the full-text index returns more than 287,000 publications and 118,000 patents. The full-text index makes Dimensions a very powerful discovery tool - especially in combination with the filtering options, which help researchers to further refine their results.

2.2 Machine learning based topic classification

Article-level indicators need to be paired with article-level classifications

Traditionally, academic databases have categorised publications using the journal's subject areas as a proxy, with a few research categories being assigned at the journal level. This approach has created unintended consequences, from content coverage in academic databases to citation benchmarking practices.

NLP and machine learning enable categorisation approaches which take the substance into account

Advances in the fields of natural language processing, machine learning, and artificial intelligence have enabled Dimensions to solve a very practical problem: to consistently categorise grants, patents, clinical trials and policy documents, a journal proxy is no longer viable. Instead, Dimensions has implemented machine learning based models for a growing number of classification systems. The models are then applied to assign the relevant research categories at the *publication* level.

The leading categorisation system with broad coverage of subject areas is the Australian and New Zealand Fields of Research system. This classification "lens" is available as part of the free Dimensions version.

2.3 Research categorisation — Australian and New Zealand Standard Research Classification

FoR - part of the Australian and New Zealand Standard Research Classification system (ANZSRC)

The Fields of Research (FoR) classification, part of the Australian and New Zealand Standard Research Classification system (ANZSRC), was developed in 2008 and updated in 2020. The ANZSRC FoR is used in all areas of research and education in Australia and New Zealand, and more broadly allows any R&D activity to be categorised using a single classification scheme. The FoR classification has three hierarchical levels: Divisions, Groups and Fields. Each Division represents a broad subject area or research discipline, while Groups and Fields represent increasingly detailed subsets of these categories.

Dimensions has implemented a machine learning approach to assign FoR categories to publications, including only the first (Division) and second (Group) levels, and excluding the 'Indigenous Studies' division as well as the generic 'other' fields from the group level. This reduces the ANZSRC 2020 Division-level categories from 23 to 22 and the Group-level categories from 213 to 171.

FoR classification covers all areas of academic research at a high level, so it works well for non-granular comparative analyses across all academia. The FoR classification in Dimensions also supports other subsequent data enrichment procedures: namely, concept relevance scoring and the field citation ratio. A full list of the ANZSRC FoR categories used in Dimensions can be found on the Dimensions website.

2.4 Other classification systems

NIH's RCDC and UK HRCS are implemented as well

Other classification systems have been implemented in addition to the FoR codes. The choice of these additional classification lenses has mainly been driven by the needs of research funders, the majority of whom are focused on the biomedical sciences. A similar machine-learning approach has been used to implement these schemes.

Examples include:

- **Broad Research Areas (BRA)** - a classification used by the Australian National Health and Medical Research Council (NHMRC) consisting of four categories - Basic Science, Clinical Medicine & Science, Health Services Research, and Public Health.

- **Cancer Types** - the ICRP's scheme complements the CSO and is linked to the International Classification of Diseases.

- **Common Scientific Outline (CSO)** - a scheme organized into six broad areas of scientific interest in cancer research.

- **Health Research Areas (HRA)** - to classify research on the basic to applied spectrum, and to identify research that is of public and global health concern, we have created the HRA categories - Biomedical, Clinical, Health services & systems and Population & Society.

- **Health Research Classification System (HRCS)** - a system used by UK biomedical funders to classify health and biomedical projects. HRCS contains two strands – Research Activity Codes (RAC) and Health Categories (HC).

- **Research, Condition, and Disease Categorization (RCDC)** - a scheme used by the US National Institutes of Health (NIH) for the public reporting required by US Congress. Dimensions implemented the technology for RCDC at the NIH and is still supporting it.

- **Sustainable Development Goals (SDG)** - areas for global development defined by the United Nations. They are a call to action to end poverty, protect the planet, and improve the lives of everyone everywhere to achieve a more sustainable future for all.

- **Units of Assessment (UoA)** - a scheme used by the Research Excellence Framework (REF) for assessing the quality of research in UK Higher Education Institutions. The Dimensions-assigned UoA framework is that used in REF 2021.

It is also possible to categorise documents that are not part of Dimensions. Please reach out to the Dimensions team if you would like to learn more.

2.5 Researcher name disambiguation

Disambiguation of people across publications, grants, patents and clinical trials

Automatically assigning publications to a researcher profile has always been a challenging task. Even with the adoption of ORCID identifiers by an increasing proportion of the research community, there still exist software solutions, such as Symplectic Elements, which help researchers, institutions and funders manage the link between publications, researchers and grants.

Dimensions aims to connect a researcher automatically to all their research objects across at least five content sources: grants, publications, datasets, clinical trials and patents. Consequently, we have developed a researcher disambiguation process that takes into account not only the metadata in each of the content sources but also publicly available ORCID data.

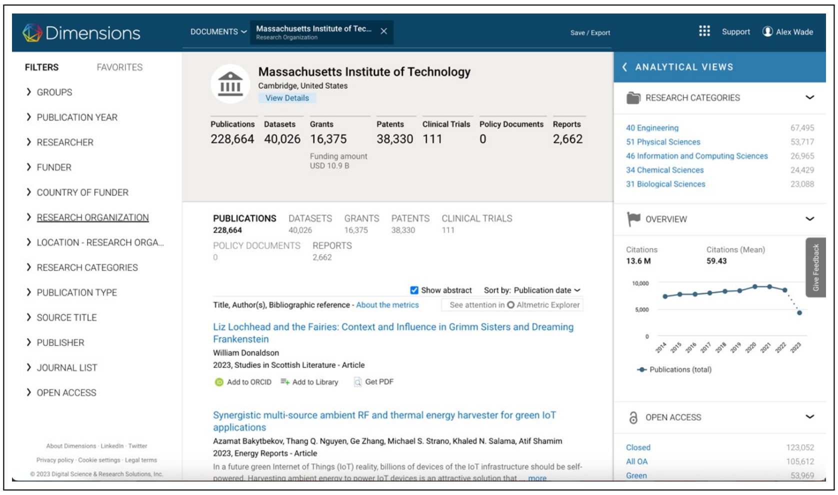
Dimensions focuses more on precision and less on recall, because we believe that assigning the wrong publications to a researcher is worse than suggesting an incomplete or fragmented record, since data errors undermine trust in the results and can be confusing. Completeness, on the other hand, can be easily fixed with the help of the user and is not as detrimental to the user experience as a basic lack of trust in the results.

2.6 Institution name disambiguation

The challenge of affiliation names

Authors of publications express their institutional affiliations in non- standard ways. Indeed, most institutions have a few name variants, but for some organizations we have found hundreds of name variants. For a data infrastructure like Dimensions it is important to be able to assign documents automatically to a unique identifier that corresponds to a single institution. On top of this, there must be useful metadata, such as geolocation information and date of foundation.

Digital Science has tackled this challenge - resulting in a curated organization database covering more than 109,000 institutions. This system allows us to create a consistent view of an organization within one content source, but also across the different types of content.



Massachusetts Institute of Technology
Cambridge, United States

Publications	Datasets	Grants	Patents	Clinical Trials	Policy Documents	Reports
228,664	40,026	16,375	38,330	111	0	2,662
		Funding amount USD 10.9 B				

PUBLICATIONS 228,664 · DATASETS 40,026 · GRANTS 16,375 · PATENTS 38,330 · CLINICAL TRIALS 111 · POLICY DOCUMENTS 0 · REPORTS 2,662

Title, Author(s), Bibliographic reference

Liz Lochhead and the Fairies: Context and Influence in Grimm Sisters and Dreaming Frankenstein
William Donaldson
2023, Studies in Scottish Literature - Article

Synergistic multi-source ambient RF and thermal energy harvester for green IoT applications
Azamat Bakytbekov, Thang Q. Nguyen, Ge Zhang, Michael S. Strano, Khaled N. Salama, Atif Shamim
2023, Energy Reports - Article
In a future green Internet of Things (IoT) reality, billions of devices of the IoT infrastructure should be self-powered. Harvesting ambient energy to power IoT devices is an attractive solution that ...

RESEARCH CATEGORIES	
40 Engineering	67,495
51 Physical Sciences	53,717
46 Information and Computing Sciences	26,965
34 Chemical Sciences	24,429
31 Biological Sciences	23,088

Citations 13.6 M · Citations (Mean) 59.43

OPEN ACCESS	
Closed	123,052
All OA	105,612
Green	53,969

Figure 2: An example Dimensions organizational profile: MIT

More recently, the Digital Science organization data has been used to seed the Research Organization Registry (ROR). In order to support interoperability with external systems, Dimensions now also includes ROR identifiers for organizations that are also in ROR.

2.7 Citations, acknowledgements and adding context

Extracting references — creating a network across sources

The extraction of the references and links between the different content types (grants, publications, clinical trials, etc.) is key to the completeness and comprehensiveness of the Dimensions citation graph. Our aim is to allow a user to gain a superior understanding of the context of a piece of research by eliminating the separations between isolated data silos. Bringing data together in this way allows a much improved view on the nature of research in a particular field as well as the associated research process. Users are then able to draw conclusions and gain new insights that previously would have taken an enormous amount of time and effort.

PUBLICATIONS

- Associated data
- Publication references
- Supporting grants
- Publication citations
- Patent citations
- Clinical trial citations
- Policy document citations

DATASETS

- Associated publications
- Supporting Grants

POLICY DOCUMENTS

- Publication references

PATENTS

- Patent references
- Publication citations
- Supporting grants
- Patent citations

CLINICAL TRIALS

- Publication references
- Publication citations
- Supporting grants

GRANTS

- Resulting publications
- Resulting patents
- Resulting clinical trials

References between records are incorporated from existing databases (such as Crossref, PubMed Central, Open Citation Data) and also are extracted directly from the full-text records provided by the content publishers. This includes not only journal publication references, but also acknowledgements and citations from and to books, conference proceedings, patents, grants and clinical trials.

In total, we have extracted more than 5.6 billion direct connections between the document records, with more than 1.7 billion between publication records alone. This number is continually growing as we integrate more content, improve the representation of the content from more publishers, and continue to improve our extraction routines.

More than 5.6 billion connections

2.8 Concepts

Concepts are key terms that are found in the titles and abstracts of publications through machine learning. They are then scored by determining the likelihood of those terms occurring in the first-level FoR category to which the publication was assigned. The relevance score of a concept is higher where the term occurs more frequently and more unexpectedly in that field of research.

3 Publications, books and citations

3.1 Dimensions publications and citations

Dimensions — a different approach

Dimensions is a powerful publication and citation database for both researchers and institutions. But simply replicating prior approaches to create yet another abstracting and indexing database is not as useful as a taking a more expansive and interconnected approach.

With Dimensions, we have set out to:

- include all relevant research objects — in essence, less editorial discrimination as to the content included;
- integrate and link other content types — grants, patents, datasets, clinical trials and more.

The decision-making power belongs in the hands of the user

Making the Dimensions database as comprehensive as possible is a central driver. Technological advances have led to higher expectations from users. People no longer expect or desire that a search engine should filter content based on the preferences of a vendor. Indeed, we have chosen to be neutral with respect to the content that we index and display to users. This means that we are not making the decision as to what is a 'worthy' research output (e.g., from which journal) to include in our database - these decisions belong in the hands of the research community or, depending on the use case, in the hands of the individual user. The goal of Dimensions is to give users the best tools to navigate content and arrive at the most relevant results in the most efficient way.

3.2 Aggregating Dimensions publication and citation data

Publication and citation data is aggregated via a two-step process:

137 million publication metadata records

Step 1: Create a metadata backbone
An extensive metadata backbone has been assembled and is continuously updated, integrating data from many sources, such as PubMed, PubMed Central, and Crossref. This step has resulted in a large index of 137 million research publications. Crossref records associated with a DOI are sourced from Crossref's 17,000 publisher members, which forms the core of the spine. But limitations remain on metadata completeness, most notably affiliation data for authors.

Step 2: Enhance the data
The metadata records are then enhanced by processing full-text records, where those have been made available to us, significantly improving discoverability of content.

99 million records enriched — from more than 160 publishers

This step includes deriving reference/citation data from the full-text and identifying mentions of funded projects, research funders and clinical trials. These full-text publications are sourced from more than 160 publishers including some of the largest STM publishers in the world.

Indexing the full-text of publications also means that a user can search for any term in a paper — not just in the title or the abstract. This index, in concert with the filtering mechanisms that we've put in place, means that your searches are more comprehensive and you are more likely to locate the research that you are looking for.

New content added continuously

New publications are added as more publishers join the effort and make their content more discoverable. If you are a publisher and want to see your content coverage improved in Dimensions, see our Help page here.

3.3 Field Citation Ratio (FCR)

After the citations to each publication have been aggregated, the field citation ratio is calculated. This is done by dividing the number of citations a paper has received by the average number received by documents published in the same year within the same Field of Research (FoR) category. In order to calculate an FCR score for an article, a minimum number of 500 articles need to be categorised in the applicable 4-digit FoR category for the year in which the article was published. If an article is categorised in more than one FoR code, then only those codes which meet the threshold of 500 labelled articles in that year will be used to calculate the FCR score. An FCR of 1.0 represents an average number of citations for a publication in the same field.

3.4 Quality related filters

To ensure that users have the right tools to make the right content filtering decisions, we enable users to limit their results to certain subsets. The standard filters are specified by pre-defined, curated lists, which can be included as search filters in the Dimensions application, as well as via Google BigQuery. We started with accepted openly available lists defined by the community, and we encourage further suggestions.

Currently, the following journal lists have been implemented in Dimensions:

Limit search results to DOAJ, ERA list, PubMed, and other lists

- **DOAJ list:** The Directory of Open Access Journals is a community-curated online directory that indexes high quality, open access, peer-reviewed journals. The DOAJ list includes over 19,000 journals covering all areas of research.
- **DOAJ non-APC journals:** A subset of the DOAJ list, limited to journals that do not charge article processing charges (APCs).
- **ERA list:** The ERA 2023 journal list was designed by the Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC), to support the national research evaluation framework, Excellence in Research for Australia (ERA). Also included are previous ERA lists from 2018 and 2015.
- **ERIH PLUS:** Operated by the Norwegian Centre for Research Data (NSD), the aim of ERIH (European Reference Index for the Humanities) is to enhance global visibility of high quality research in the humanities published in academic journals in European languages all over Europe. All journals included in ERIH have to meet threshold standards for scholarly journals.
- **J-Stage:** J-STAGE is an electronic journal platform for science and technology information in Japan, developed and managed by the Japan Science and Technology Agency (JST). J-STAGE aims to ensure the internationalisation of the science and technology information published in Japan.

- **Nature Index journals:** Compiled by Nature Research, the Nature Index Journals are science journals included in the ”Nature Index,” a database of author affiliation information collated from research articles published in an independently selected group of 82 high-quality science journals.

- **Norwegian Register:** The Norwegian Register is operated by the Norwegian Centre for Research Data (NSD) and the National Board of Scholarly Publishing (NPU). It includes journals recognized in the weighted funding model and includes $\sim$ 30,000 source titles. Journals are divided by quality into Levels 0, 1, and 2, where Level 2 is superior than Level 1. Source titles which do not meet these criteria are identified as Level 0.

- **PubMed**: PubMed is an index of life science and biomedical publications mainly sourced from MEDLINE, and maintained by the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH). The PubMed list in Dimensions filters to publications which have a PubMed identifier (PMID).

- **SciELO:** Scientific Electronic Library Online (SciELO) is a platform for scientific journals, with national focal points within Latin America and the Caribbean. Countries with collections in SciELO include Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Ecuador, Mexico, Paraguay, Peru, Uruguay and Venezuela.

- **UGC journals:** Maintained by the Indian University Grants Commission - Consortium for Academic and Research Ethics (UGC-CARE), the Reference List of Quality journals is divided into two groups:
  - **UGC-CARE List Group I:** Journals found qualified through UGC-CARE protocols
  - **UGC-CARE List Group II:** Journals indexed in globally recognised databases

- **VABB-SHW:** Operated by the Centre for Research & Development Monitoring, the Flemish Academic Bibliography for Social Sciences and Humanities (VABB-SHW) is a database of publications from the social sciences and humanities from researchers affiliated with Flemish universities.

We are keen to learn about general, national or institutional filters that should be considered, as well as new use cases where lists may be helpful.

3.5 Dimensions Research Integrity Trust Markers

As part of our commitment to enhancing the comprehensiveness of the Dimensions database, in 2023 we introduced a new dataset referred to as Dimensions Research Integrity. This dataset comprises what we call trust markers, the hallmarks of responsible science as evidenced within published research papers, related to study transparency and reproducibility.

These trust markers are increasingly prevalent across scientific research, often mandated by publishers and funders, and shed light not only on research integrity but on research practices. Incorporated into research landscape analysis, they can provide a more complete assessment of current practices and identify areas for improvement to support open science. For example, they highlight where researchers are storing their data so that universities can provide better support for their infrastructure needs. They can help funders visualize the impact of policy decisions and monitor compliance. They can help publishers to monitor the adoption of article templates.

The following trust markers are available as of July 2023:

Trust Marker	Description
Funding	States if the author(s) of the paper were granted funding to conduct their research.
Ethical approval	Statement affirming that ethical approval from a human subjects or animal welfare committee was either obtained or was not required for the study.
Competing interests	Declares possible sources of bias, based on personal interests of the author(s) in the research findings. For example, funding sources, past or present employers, or the author(s) financial interests.
Author contributions	Details of each author's role in the development and publication of the manuscript.
Data availability	A dedicated section of a scientific work indicating whether data from the research is available and where it can be found.
Data location	Locations where research data (raw or processed) can be accessed.
Repositories	The names of any research data repositories used by the author(s) to preserve, organize and facilitate access to study data.
Code availability	States how one could gain access to the code used to conduct the study/research.

We analyze the full-text of published research articles using natural language processing to identify and categorise trust markers. We have scanned over 33 million research articles from 2010 onward, providing a rich data source on the quality of scientific literature.

On average, the model for each trust marker incorporated in Dimensions Research Integrity has been trained, evaluated, and validated across 10 different fields of study to ensure they are not concentrated or biased toward a single field. Additional trust markers are in development to align with various study reporting guidelines and will be added to the Dimensions Research Integrity dataset over time.

Dimensions Research Integrity utilizes methodologies developed by Ripeta, one of the Digital Science portfolio companies

The Dimensions Research Integrity dataset is available as a module within the Dimensions Google BigQuery offerings. Dimensions Research Integrity consultancy reports for individual funders, publishers, and institutions are also available and the Dimensions Research Integrity dashboard is available via Dimensions Apps&Modules.

3.6 Beyond academic attention - Altmetrics in Dimensions

Altmetrics - an immediate and different type of impact

Digital Science was an early supporter of the alternative metrics movement and Altmetric has played a key part in defining the agenda around altmetrics. Indeed, Altmetric has lead the field with a number of innovations including the colorful Altmetric badges, Altmetric Attention Score, and the inclusion of unique sources like policy documents and syllabi.

Dimensions includes high-level Altmetric data for each article in the index and displays this on the article details page. In this way, we bring together the academic attention (citations), innovation attention (patents), clinical attention (trials) alongside public and policy engagement attention including social media, traditional media, policy attention and the other forms of attention that Altmetric indexes.

The need to demonstrate the impact of research has sought to bring together data to tell stories to describe the route to impact. The inclusion of Altmetric data natively in Dimensions moves the community a step closer to understanding the impact of research in more quantifiable terms.

3.7 Open Access, Open Citation Data and Dimensions

Open Access in Dimensions

Digital Science supports Open Access and Dimensions can be a helpful tool for the community in supporting these efforts. We have integrated OA data from Unpaywall and maintain a list of full OA journals based on DOAJ to create a comprehensive view of Open Access publications. Dimensions allows users to access OA articles with a single click, opening article directly in a ReadCube overlay window on top of the Dimensions interface to get the user to the content as quickly as possible.

Dimensions - making citation and metadata available

Dimensions is an example of the power of making publication metadata and citations data, publicly available in order to stimulate innovation and novel solutions. Dimensions has been developed with this goal in mind: making good quality, consistent and linked metadata available to the community to improve access and to stimulate creativity. So much can be done with these data and to create innovation that supports research.

Dimensions and open citation data

Dimensions is also aligned with Initiative for Open Citations (I4OC). Indeed, Dimensions is an example of what can be done if citation data is more openly available. In building Dimensions, Digital Science invests significant effort to make a rich citation graph so that a good quality discovery experience can be delivered to our users. We hope that the I4OC and similar initiatives continue to lower that barrier going forward, allowing the community to focus on more valuable functionality for users who want to push their research forward faster.

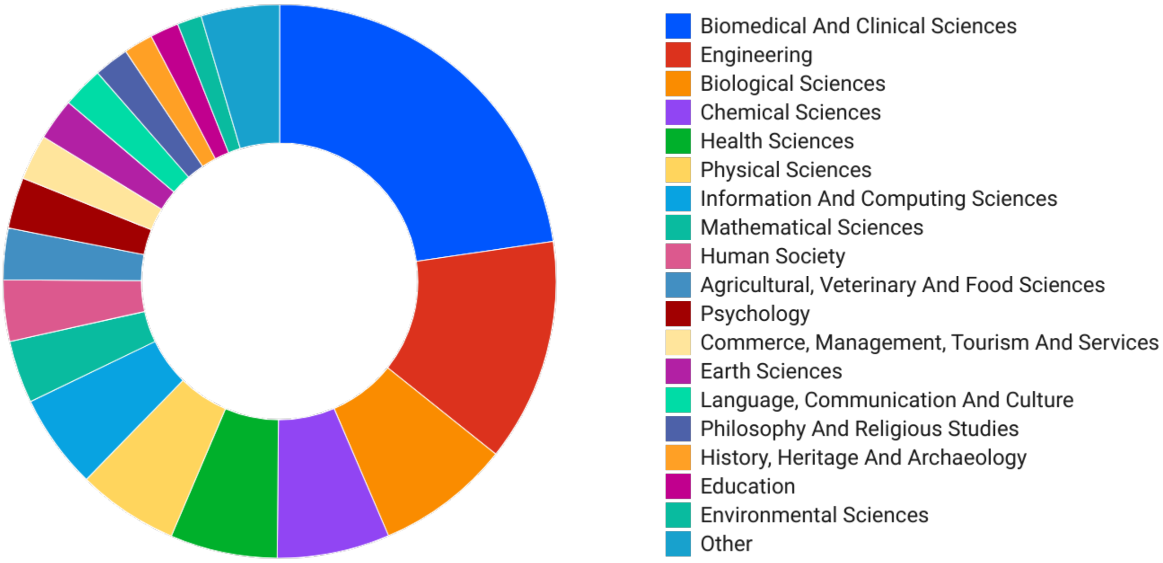
Since we have been asked this question often: Digital Science is not a publisher and is not in the best position to contribute citation data to I4OC — we believe this should come from publishers themselves. From Digital Science, both Altmetric and Figshare are members of the initiative.

3.8 Key statistics on publication and citation data

The following statistics were captured in July 2023 and are changing on a daily basis - this means that the values in this document can vary from the actual results in the Dimensions application or API.

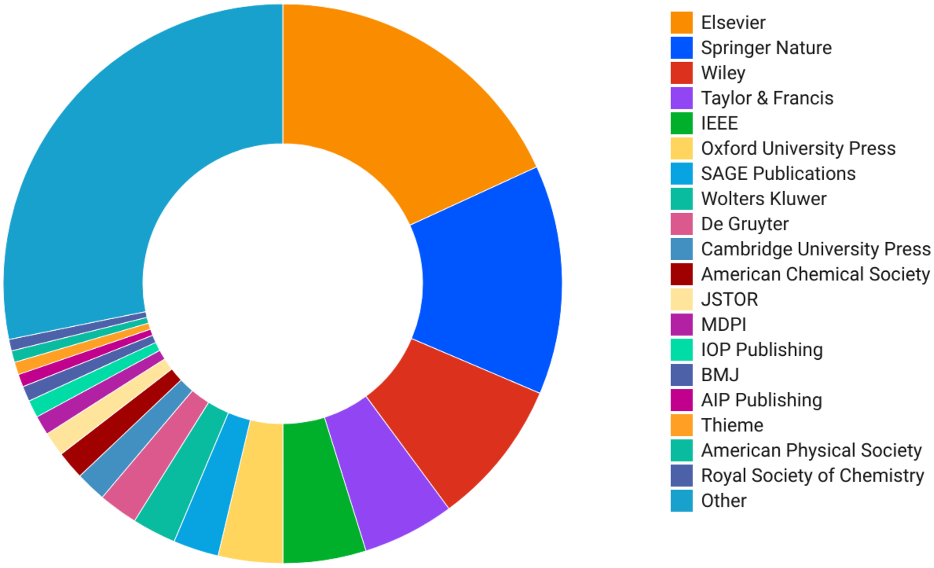
Publications	137 million
Source titles (journals, book series, preprint servers, conference proceedings)	209,000
Number of links to research organizations	117 million
Number of links to researchers	318 million
Number of cited references	1.7 billion
Number of links to grants	22 million
Number of links to funders	34 million
Number of links to clinical trials	2.3 million

3.9 Distribution of publications across disciplines



Created with Datawrapper

3.10 Distribution of publications by publisher



Created with Datawrapper

4 Beyond publications — a broader view

4.1 How does Dimensions compare to other services

Not comparable — a new and innovative approach, linking grants, publications, datasets, patents, clinical trials and policy documents

Dimensions is not directly comparable to PubMed, Google Scholar, Scopus or Web of Science since it is much broader. It covers the 'basics' in terms of a robust and even more comprehensive publication and citation database. But Dimensions transcends these existing tools and databases. The bringing together of grants, publications, datasets, clinical trials, patents, and policy documents, linked and contextualised, opens up a world of proper discovery, research planning and impact communication possibilities. In addition, the Dimensions web application presents search results in context allowing a user to understand the setting of a search result at a glance, and facilitating greater exploration of potentially relevant works, funding or routes to impact.

Dimensions provides:

- A more expansive citation graph than offered by Scopus or Web of Science;
- Wide coverage and an enhanced experience around discovering the right (or most relevant) research based on indexing the full-text, in a similar approach to Google Scholar;
- Grants as an early trend discovery method showing the intended rather than published research;
- A broad linked and rich view on content relevant for the research process — to avoid the narrow focus on publications and citations, allowing a deeper understanding of the inputs, outputs and impact and how they are related.

The data is provided via a powerful API and easy integration with other data sources through Google BigQuery, allowing machine-to-machine interaction. This is available in the institutional subscription but can also be made available to individual researchers for research purposes upon request.

4.2 Citation counts in different systems and databases

Citations counts - why do they differ?

One question we are often asked is, 'how do Dimensions citation counts compare to Google Scholar, Scopus or Web of Science?' As much as we would like to give a simple answer, it is not possible. First of all, Dimensions and the references that it contains are not directly comparable with other databases since Dimensions also captures references and links to sources beyond classic publication-based citations. Even if we only examine the publication-based citation count, it is not possible to establish a simple ranking. (This type of work was already found by the bibliometrics community in the comparison of the Scopus and Web of Science databases following the launch of Scopus in 2006.)

There are several reasons why Google Scholar, Scopus, Web of Science and other services may show different citation counts for the same content. Some of the reasons for these disparities include:

- each database covers different sets of content to build its citation graph;
- each database includes content from different date ranges (e.g. 1996 to present);
- each database includes different types of content. For example, some sources may only include references from peer-reviewed journals, while others may include references from non-published or not-yet-published works, such as student theses published on a website, citations from preprints (where versioning and disambiguation of pre- and post-print versions of the same paper adds yet more complexity);
- the frequency at which the content is updated differs, from daily to weekly and beyond;
- extracting references from a paper and uniquely matching them to the reference graph is a challenge which each database solves in different ways (there is no standard, industry-defined approach and, as a result, in some cases references may not properly match);
- as algorithms for matching improve and new data sources become available, reference graphs may be updated, resulting in changes to citation counts.

As ever, we look for feedback from the community to prioritise our development focus for content integration.

As an illustration, an example from a paper in PLOS One:

FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Price MN, Dehal PS, Arkin AP (2010) *PLOS ONE* 5(3): e9490. doi:10.1371/journal.pone.0009490

Database	Citations
Crossref	8,875
Dimensions	9,525
Google Scholar	10,636
Scopus	7,552
Web of Science	7,477

Given the variables described above, it is unlikely that multiple parties will arrive at a single absolute citation count. As a result, in practice many researchers consider citation counts to be a useful *relative* metric when comparing other content within a single system. It has required a large amount of effort and resources to bring the current sources and content together - we consider this only to be a starting point. This is clearly a team effort and we need you as the users, the research community and the broader Dimensions team!

5 Grants - a glimpse into the future

Grants, a forward looking data source - neglected for too long

Funded grants are the result of an extensive process in which a researcher or team of researchers describe the research project that they wish to undertake. Their aim is to convince a research funder, through an anonymous peer review panel, that the research problem is interesting, tractable and worthy, and that the team is qualified and capable of achieving the outcomes suggested. This process is even more important since, in most cases, the money being spent is public money and hence must be accounted for in a responsible manner. Grants are the first manifestation of a research idea in a cogent format that must convince a third-party of their value — a little like a beta software release.

This position in the research cycle makes grants a unique source for discovery since it allows analysis of trends and movements in fields by looking at the research that is intended to be carried out in the coming years — a glimpse into the future. For funders, research policy strategists and planners, analysis of the funding landscape allows early intervention and strategy formulation, not only the retrospective identification of fast facts or wrong decisions.

A grant database with $2.3 trillion in funding

We have been working with research funders since 2013 to aggregate a large grant database. Our aim was to enable, for the first time, a broad view across national and institutional borders on the resource input aspects of the research system and to make this available not just to the largest funders, who have the responsibility to commission custom systems to ensure appropriate reporting to public stakeholders, but also to smaller funders with smaller teams and more limited resources. Our early efforts are now a part of the broader version of Dimensions, which covers the entire flow from input to academic attention, commercialisation, policy formulation and routes to impact.

Grant data provides particular insights, not a complete research funding view

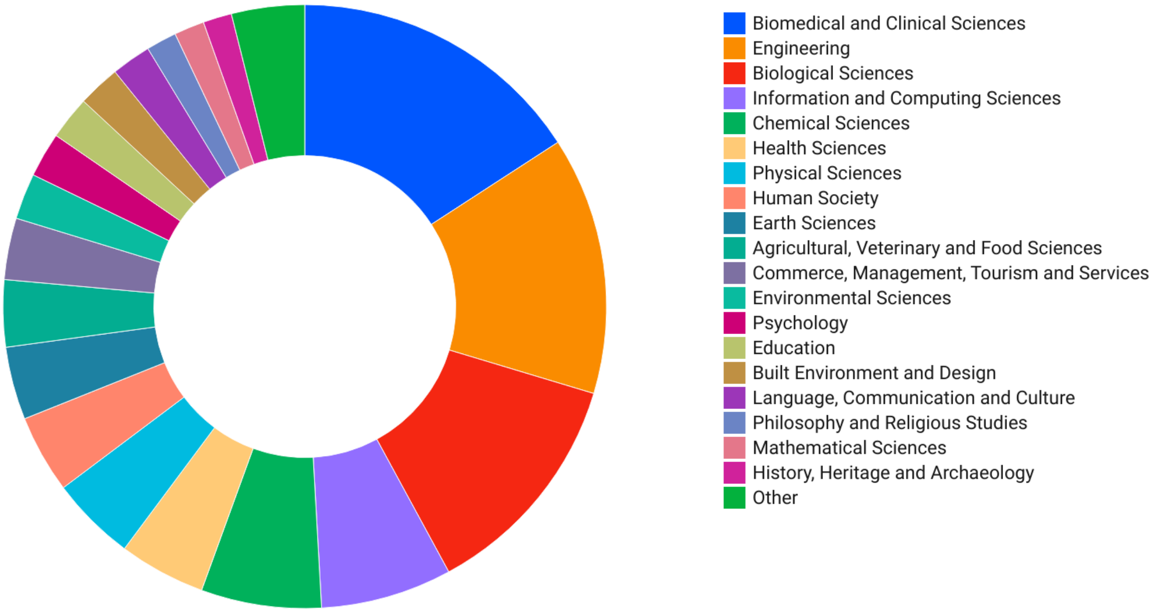
Grants are a difficult content source for several reasons: they do not follow a common metadata schema in the way that publications do; many do not yet have a persistent identifier such as a DOI; and they are highly dependent on individual national frameworks of research funding. Geographic differences are not trivial. In some countries, the majority of the research funding is given out in competitive project grants, while in other countries there is more emphasis on block funding, which never appears in funded grants databases. And there are a lot of countries that fall between the ends of this spectrum with a mix of block funding and project-based funding. For that reason the Dimensions grant data should not be taken as a complete view on all research related funding. It covers project-based funding from different types of funders (government, multinational, charities, etc.). If you have any questions related to your use case do not hesitate to reach out to us.

5.1 Key statistics on the Dimensions grant data

The following key statistics have been captured in July 2023 and are changing on a monthly basis — this means that the values in this document can vary from the actual results in the Dimensions application, the API, or via Google BigQuery.

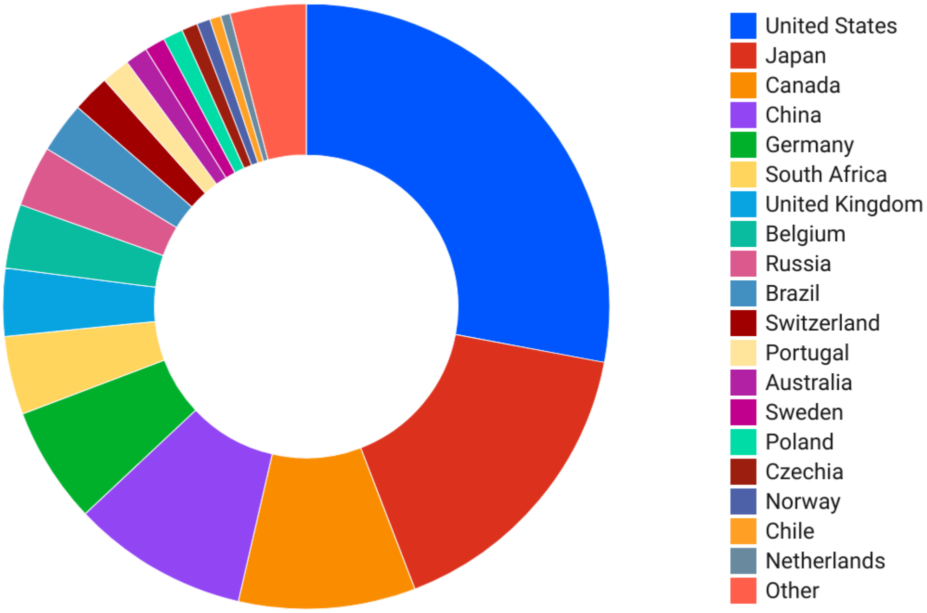
Grants	6.8 million
Research funders covered	670
Total funding amount	USD 2.3 trillion
Average funding amount	USD 455,000
Total funding of projects active in 2023	USD 528 billion
Number of links to research organizations	7.1 million
Number of links to researchers	9.9 million

5.2 Distribution of funded projects across disciplines



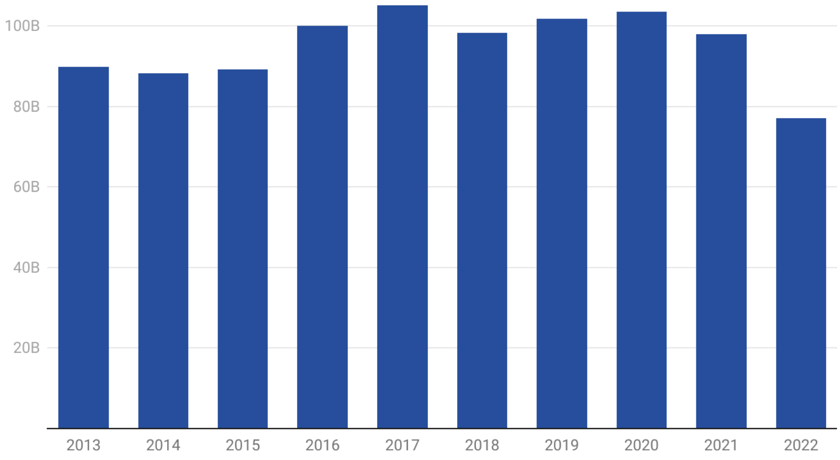
Created with Datawrapper

5.3 Geographical distribution of grants



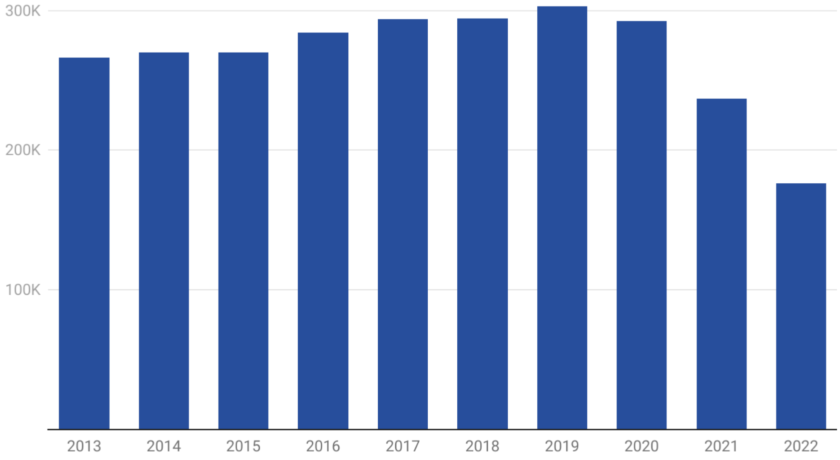
Created with Datawrapper

5.4 Aggregated funding amount of starting grants



Created with Datawrapper

5.5 Number of starting grants over time



Created with Datawrapper

6 Clinical Trials — research results en route to clinical application

Clinical trials, aggregated from different registries

A clinical trial is any research study that prospectively assigns human participants to health-related interventions to evaluate the effects on health outcomes. Interventions include, but are not restricted to drugs, cells and other biological products, surgical procedures, radiological procedures, devices, behavioural treatments, process-of-care changes, preventive care, etc. (source: WHO).

Dimensions provides a single point of access to multiple clinical trial registries. As of July 2023 we have integrated 15 registries:

Registry	Country/Territory
ANZCTR	Australia / New Zealand
CHICTR	China
ClinicalTrials.gov	United States
CRIS	Korea
CTRI	India
ENCePP	Netherlands
EU-CTR	European Union
GCTR	Germany
IRCT	Iran
ISRCTN	International
jRCT	Japan
NTR	Netherlands
PACTR	Africa
ReBEC	Brazil
UMIN-CTR	Japan

We map source data into Dimensions' data model with filters, for e.g. research categories, research organizations or years, across content types.

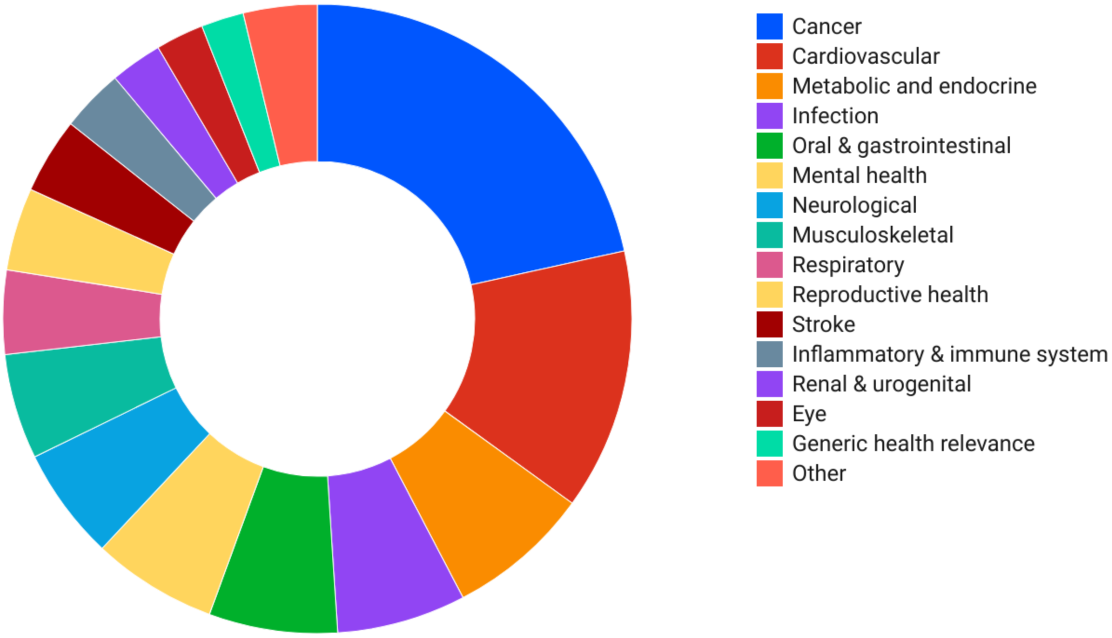
6.1 Key statistics on the Dimensions clinical trials data

The following statistics were captured in July 2023 and are growing daily — this means that the values in this document can vary from the actual results in the Dimensions application, Google BigQuery, or API.

Clinical trials	789,000
Clinical trial registries covered	15
Number of links to sponsors / collaborators	3.6 million
Number of links to publications	534,000
Number of links to grants	34,000
Number of links to funders	634,000

6.2 Distribution of Clinical Trials by discipline

The following chart shows a distribution of clinical trials across disciplines (based on the Health Research Classification System (HRCS) from the UK):



Created with Datawrapper

7 Patents — practical and commercial applications of research

Patent data - translation of research activities into the commercial space

The patent data in Dimensions is provided by Digital Science portfolio company IFI CLAIMS. The focus of the patent data in Dimensions is to provide a downstream view on how research funding is impacting and enabling the commercial protection of intellectual property and the potential commercial use of research results.

As of July 2023, Dimensions includes over 155 million patent records from 50 countries around the world. The top 20 jurisdictions by number of patent records are listed below.

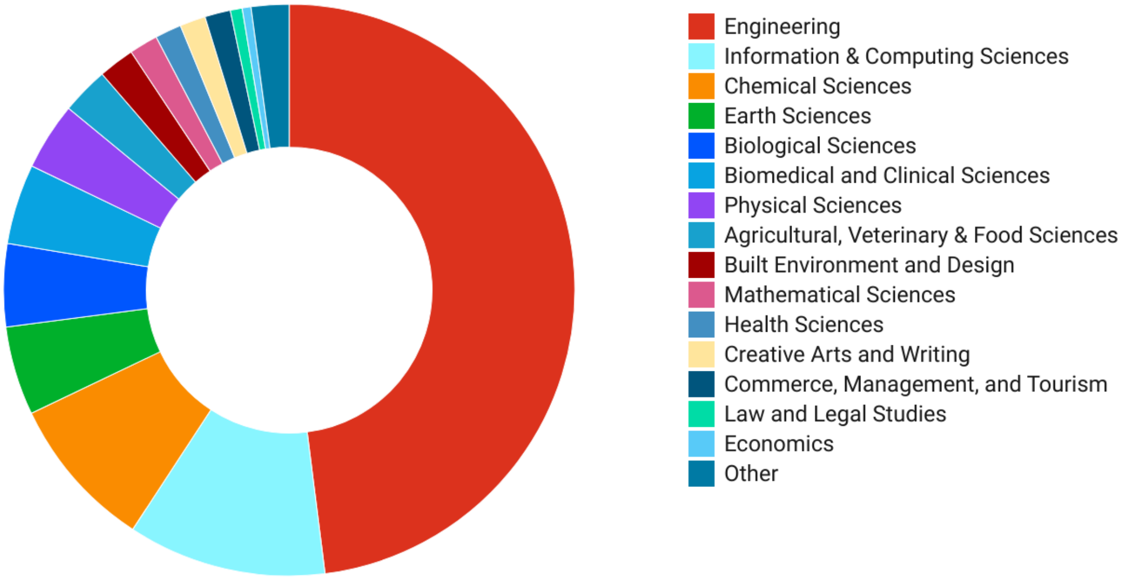
Jurisdiction	Patents
China	42,445,374
Japan	27,231,046
United States	19,647,310
Germany	8,193,086
European Union	7,870,794
Korea	7,071,406
World Intellectual Property Organization	5,307,461
Great Britain	3,901,958
France	3,272,121
Canada	3,124,791
Australia	3,051,474
Taiwan	2,187,231
Spain	2,026,482
Russia	1,733,404
Soviet Union	1,452,748
India	1,304,245
Italy	1,143,613
Austria	1,126,798
Brazil	1,055,389
Sweden	752,416

7.1 Key statistics on the Dimensions patent data

The following key statistics were captured in July 2023 and are changing on a weekly basis — this means that the values in this document can vary from the actual results in the Dimensions application, the API, or in Google BigQuery.

Patents	155 million
Jurisdictions covered	107
Number of links to research organizations	67 million
Number of cited patent references	475 million
Number of links to publications	15 million
Number of links to grants	385,000
Number of links to funders	590,000

7.2 Distribution of patents across disciplines



Created with Datawrapper

8 Policy documents — research resulting in policy and guidance documents

Policy documents from over 70 publishing organizations

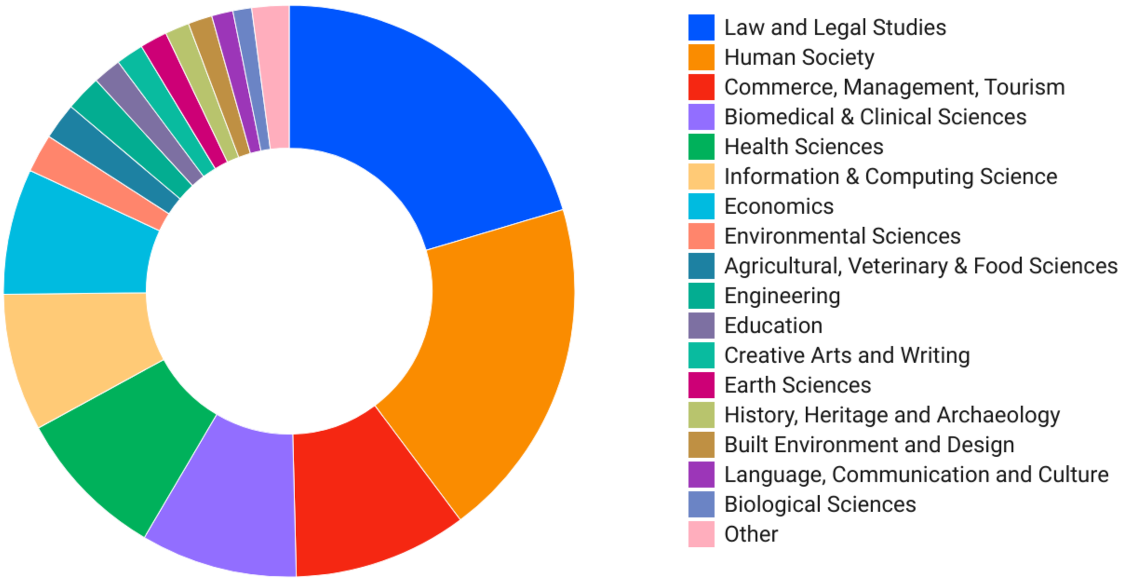
The policy document data in Dimensions includes policy sources that are designed to change or otherwise influence guidelines, policy or practice. Tracked policy sources range from government guidelines, reports or white papers, independent policy institute publications, advisory committees on specific topics, research institutes, and international development organisations. Digital Science curates a broad scope of policy sources from organisations around the world which cover topics from climate change to health, transport and economics. Wherever possible we index the full text, allowing us to categorise the record and extract references.

8.1 Key statistics on the Dimensions policy document data

The following key statistics were captured in July 2023 and are changing on a daily basis — this means that the values in this document can vary from the actual results in the Dimensions application, the API, or via Google BigQuery.

Policy documents	871,000
Publishing organizations covered	396
Number of links to publications	2.4 million

8.2 Distribution of Policy documents across disciplines



Created with Datawrapper

9 Datasets

9.1 Key statistics on the Dimensions Datasets

Full coverage for open research datasets

The addition of datasets within Dimensions allows users to discover and analyse trends in publicly available data at an institutional level and makes even more linked data available in one platform, rather than via disconnected databases.

Dimensions indexes more than 12 million datasets from more than 1,500 repositories via the following sources:

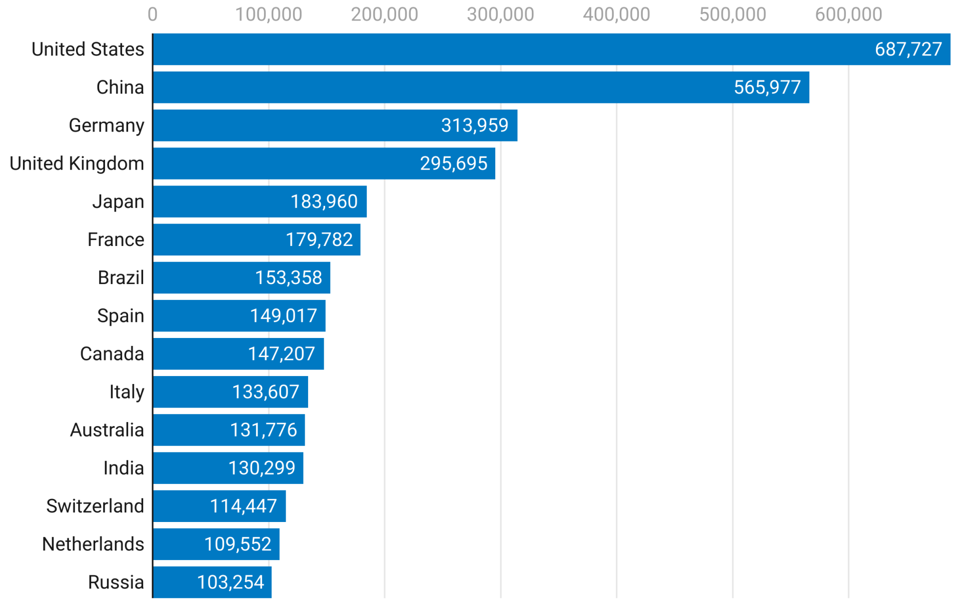
- **DataCite**: all works of Resource Type = "Dataset";
- **Figshare and more than 100 repositories hosted by Figshare**: includes all articles of Item Type = ”Dataset”.

Dimensions does not index as a dataset any journal articles, media or other content types from any of these sources / repositories.

Due to increasing funder mandates around data sharing, the number of datasets published has increased by an order of magnitude in the last decade. This is also reflected by the steady increase in the number of datasets indexed each year by Dimensions, which has grown to approximately 2 million new datasets per year.

9.2 Geographical distribution of datasets

The global nature of the Dimensions dataset indexing is a reflection of the global funder push for open research data. There are now 54 funders globally that require data archiving and 38 that encourage data archiving, as per Sherpa Juliet from JISC. Within the corpus we can already see 15 countries with more than 100,000 datasets indexed.



Thank you

Thank you for your interest in Dimensions. We look forward to improving both the tool and the data in cooperation with you and the research community.

Legal note: while we have tried to ensure the accuracy of this report, it is subject to change and provided for information only on an ”as is” basis, and is not intended to form part of any legal contract. Any reference to a third party in this report should not be considered as an endorsement by or of, or indication of any association with, Dimensions or Digital Science.

Part of **DIGITAL** science

Altmetric

CC Grant Tracker

DIGITALscience Consultancy

Dimensions

figshare

gigantum

GRID

ifi CLAIMS

Overleaf

readcube

ripeta

scismic

SYMPLECTIC

writefull

digital-science.com