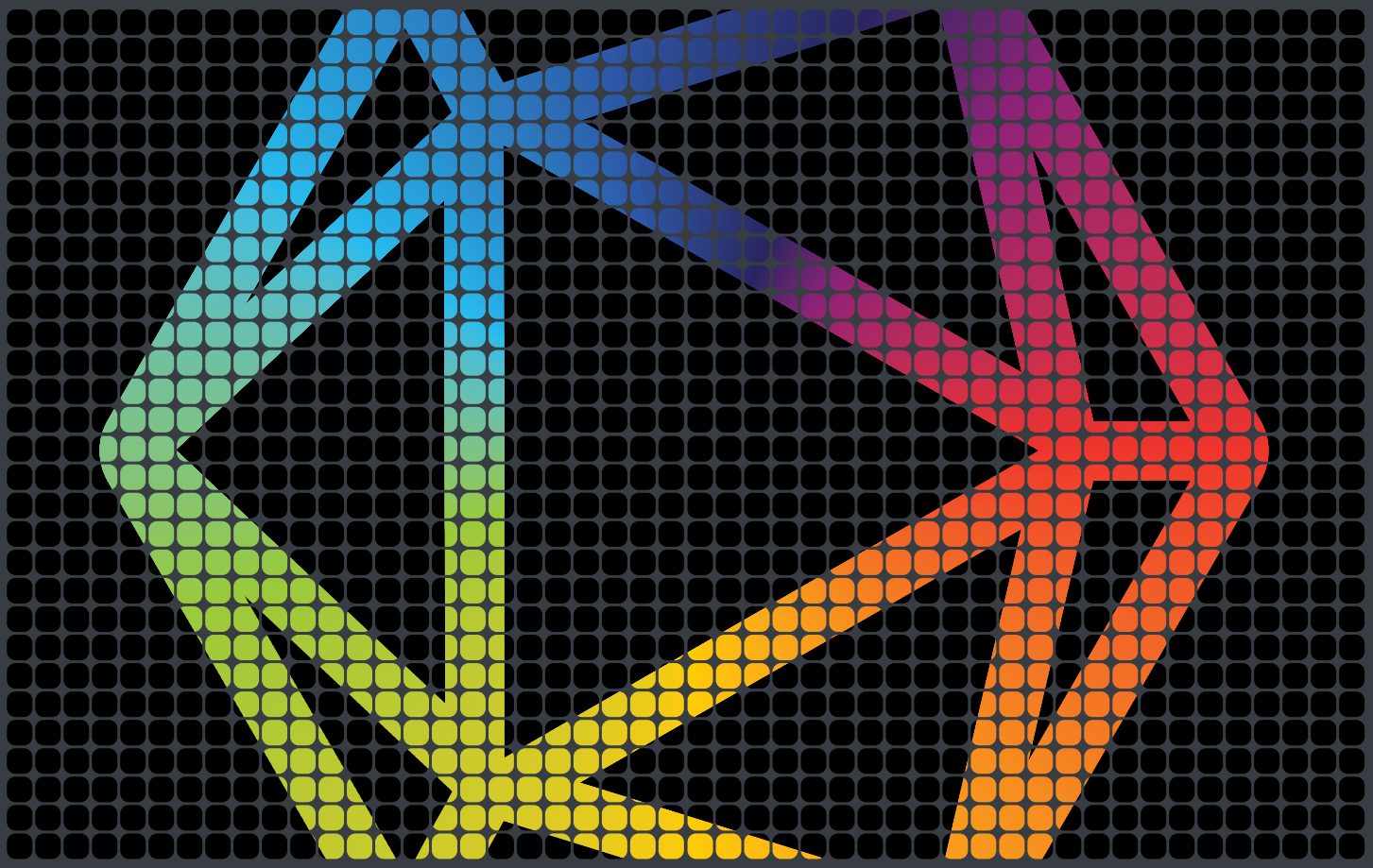


Dimensions Report

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

APRIL 2019

DIGITAL science

Dimensions

About Dimensions

Dimensions® is a modern and innovative, linked research data infrastructure and tool, re-imagining discovery and access to research: grants, publications, citations, clinical trials, patents and policy documents in one place. The development of Dimensions has been triggered by the feedback from clients and partners of the Digital Science portfolio companies. As a result, Dimensions has been developed through a dynamic collaboration across Digital Science and six of its portfolio businesses (ReadCube, Altmetric, Figshare, Symplectic, DS Consultancy and ÜberResearch). With each company focused on a different pain point within the research cycle and serving various stakeholders in the research ecosystem, these teams shared their true passion for innovation, and contribute their unique experiences, opinions, and values into Dimensions. Visit www.dimensions.ai

About Digital Science

Digital Science® is a technology company serving the needs of scientific and research communities at key points along the full cycle of research. We invest in, nurture and support innovative businesses and technologies that make all parts of the research process more open, efficient and effective. We believe that together, we can change research for good. Visit www.digital-science.com

Acknowledgements

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 owned by the Holtzbrinck Publishing Group..

For inquries in respect of Dimensions, please contact info@dimensions.ai, otherwise please write to Digital Science at info@digital-science.com or 625 Massachusetts Avenue, Cambridge, MA, 02139 USA.

Copyright © 2018 Digital Science & Research Solutions Inc.

DOI: 10.6084/m9.figshare.5783094

Contents

A modern linked research data landscape

The broader Dimensions team: 100+ development partners and Digital Science

Dimensions was created in response to two significant constraints for Digital Science and its development partners. The first constraint was that existing solutions sought to understand the research landscape solely through the lens of publication and citation data. The second constraint was the way that existing solutions exposed what data they did have. Much of the publications research graph had been locked away in proprietary applications, which constrained how the information could be used, including through 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 try to stimulate innovation to support research, we worked closely 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

At the heart of Dimensions, we wanted to do something transformative for research and that was always going to have multiple components. A key part of that vision was that Dimensions makes available, without charge, publication citation data via the Dimensions application (visit https://app.dimensions.ai) and via APIs - the metrics in Dimensions are available via the open Dimensions Metrics API and the Dimensions Badges (visit https://badge.dimensions.ai) - in both cases for non-commercial purposes.

Does it support your use case?

Another aspect of supporting the academic community was empowering the community. 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 to own the formulation and development of metrics that tell the best stories and give the best context to a piece of research.

This document provides an overview of the Dimensions content. Feel free to reach out to the Dimensions team if you want to discuss further whether the content scope and coverage of Dimensions can help in your specific situation and use case.

We will improve it together!

One of the most important aspects of Dimensions is that we are going to develop it further with the research community - any feedback is welcome. Please contact us at info@dimensions.ai.

Quick facts on Dimensions - the total record count and more	
Content type	**Number of items indexed**
Publications	100 million
Grants	4.6 million
Patents	38 million
Clinical Trials	455,000
Policy Documents	422,000
Records with Altmetric attention	10 million
Grand total	**153 million**

Linking it all together and enriching it for the user

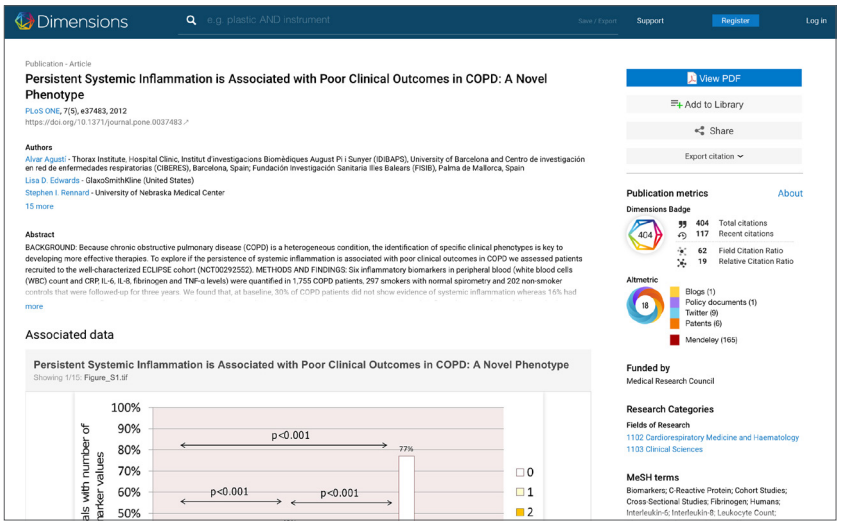
Linked and integrated data from multiple sources are core to Dimensions. This has been a key feature in discussing the product scope and direction with development partners, who agree that the integrated view enables novel insights. The following sections provide a quick overview of the key approaches which are visible to the user.

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

We are realising these linkages with a data driven, machine learning and AI-based approach, automatically extracting the information to create the connections. The content and enrichment pipeline is as automated as possible, allowing us to provide Dimensions with publication / citation data to researchers for free, and to research institutions at realistic cost levels. While an automated approach allows us to offer a more open, free approach it also results in some data issues, which we will continue to have to work on and improve. If you see anything that doesn't seem correct in our data case please reach out to us. We are always looking to improve the processing pipeline and subsequently the data and services that Dimensions provides - please email us at support@dimensions.ai.

Automated process, efficient and effective, but we need your help to constantly improve the quality

An example of a publication record in Dimensions with links to all other content sources - allowing the user already in the freely available version to explore these relations:



Example "Persistent Systemic Inflammation is Associated with Poor Clinical Outcomes in COPD: A Novel Phenotype" (DOI 10.1371/journal.pone.0037483)	
Publication references	44
Associated data sets	15
Supporting Grants	2
Publication citations	401
Patent citations	5
Linked Clinical trials	2
Policy document citations	1
Altmetric Attention Score	18

Full text indexing - real discovery instead of missing relevant information

Full text index - enabling deep discovery

Dimensions provides researchers with a free discovery service. Our approach to indexing the full text makes publications and books much more discoverable. Full text search is already available for over 69 million publication records in Dimensions. For example, a search for CRISPR in just title and abstracts brings back about 15,000 results, while the Dimensions search using the full-text index results in more than 77,000 results. The full text index makes Dimensions a very powerful discovery tool - especially with the filtering options, which helps researchers to further refine their results.

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

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

Machine learning based research topic classification - Fields of Research and other classification systems

In existing databases such as Web of Science and Scopus, the documents are typically categorized using a journal as a proxy, with a few research categories being assigned at the journal level. This approach has created unintended consequences across research, from content coverage in databases to citation benchmarking practices.

Technology has developed further. The fields of natural language processing, machine learning and artificial intelligence have all made huge advances in recent years. Dimensions has been able to leverage these technologies to solve a very practical problem requiring a different approach: If you want to consistently categorize grants, patents, clinical trials and policy documents, a journal proxy is no longer available. The path we have chosen for Dimensions is to use existing classification systems and a machine learning based approach to automatically assign a consistent set of categories to all documents - regardless of the source.

We implemented established research classification systems that have existing associated datasets that we are able to use to train our classification algorithms. The leading categorization system with broad coverage of subject areas and a large general corpus of training material is the Australia/New Zealand Fields of Research system. This classification "lens" has been made available as part of the free Dimensions version.

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

Research categories in Dimensions - Australian and New Zealand Standard Research Classification (ANZSRC)

The Fields of Research (FOR) classification is a component of the Australian and New Zealand Standard Research Classification (ANZSRC) system, developed in 2008. It allows all R&D activity to be categorized using a single system. The ANZSRC is used in all areas of research and education in Australia and New Zealand. The FOR classification has three hierarchical levels: Divisions, Groups and Fields. Division represents a broad subject area or research discipline, while Groups and Fields represent increasingly detailed subsets of these categories. There are 22 Divisions, 157 Groups and 1238 Fields. We have emulated the second level of the system only (Groups) in Dimensions. We have used a reverse-engineering technique, based on machine learning, where a corpus of manually-coded grants are examined and the manually-applied codes FOR - part of the Australian and New Zealand Standard Research Classification (ANZSRC) system are

reproduced by the algorithm. This is then checked against actual codes, and the algorithm is iterated.

FOR classification covers all areas of academic research at a high level, so it works well for non-granular investigations by broad subject areas. Therefore, FOR is good for comparative analyses across all academia.

Other classification systems

Other classification systems have been implemented in addition to the FOR codes. The choice of these different classification lenses is mainly driven by the needs of research funders, the majority of whom are focused on the biomedical sciences. An analogous machine-learning approach has been used to implement these schemes. Examples include:

NIH's RCDC and UK HRCS implemented as well

- The Research, Condition, and Disease Categorization (RCDC) is a classification scheme used by the US National Institutes of Health (NIH) for the public reporting required by the US Congress. The ÜberResearch team has implemented the technology for RCDC at the NIH and is still supporting it.
- The Health Research Classification System (HRCS) is a classification system used by nearly all UK biomedical funders to classify their portfolio of health and biomedical projects. There are two strands to HRCS – Research Activity Codes (RAC) and Health Categories (HC).

Any other classification system can be generated in a similar way with very little effort. Several additional schemes have been implemented for clients with specific topic classification needs. Examples could be classification systems on a national level or very specific topic focused systems. If required, it is also possible to categorize documents that are not part of Dimensions. Please reach out to the Dimensions team if you would like to learn more.

Other classification systems can be implemented

The challenge of affiliation names

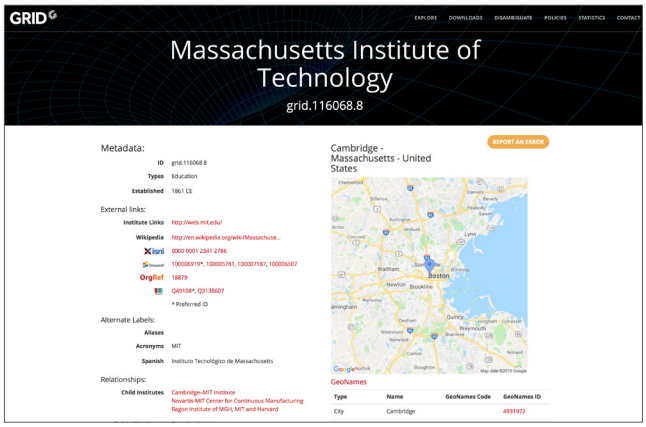
Disambiguating institution names - based on GRID

Authors of publications (as well as the other research objects such as grants and patents) express their institutional affiliations in non-standard ways. Indeed, most institutions have a few name variants but for some organizations we 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. Furthermore, each institution in that unique identifier list must be well defined according to a policy that helps to quantify what we classify to be an institution, why it has been included and what type of institution we believe it to be. On top of this, there must be useful metadata, such a geolocation information, date of foundation and, most importantly, a persistent identifier.

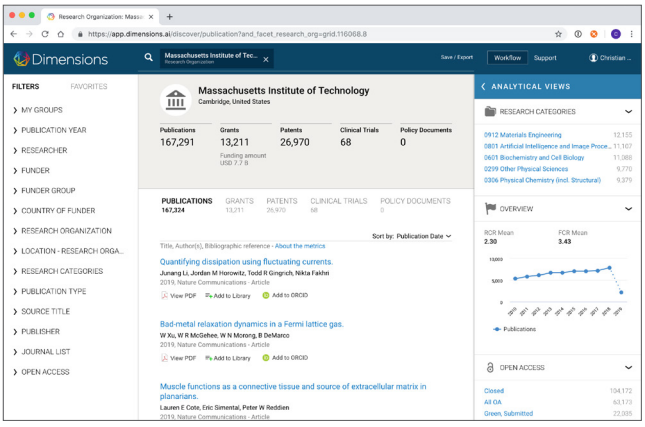
GRID - an open resource provided by Digital Science

Digital Science has already started to tackle that challenge - resulting in the release of the open GRID database, which has grown to cover more than 90,000 institutions, where the data has been curated and each institution assigned a persistent identifier. GRID is continuously improved and used in many other systems, for example ORCID (see ORCID blog post).

In Dimensions, the GRID system is used to allow us to create a consistent view of an organization within one content source, but also across the different types of content.



MIT GRID record



MIT Dimensions profile

Feedback to improve GRID is appreciated!

GRID is continuously improved as we encounter more data and feed that back into the GRID database. Digital Science is committed to providing GRID on an ongoing basis as an open dataset under a CC0 license to support the research community. GRID is not yet perfect and never will be. Research organizations change: some merge, some rename themselves, new institutions appear. Change here is more fluid than you think! For more information on GRID please visit www.grid.ac or submit a support request via Dimensions for improvement suggestions.

Person disambiguation across publications, grants, patents and clinical trials - a challenging task

Researcher disambiguation across multiple sources

Automatically assigning the correct publications to a researcher has always been a challenging task. Even with the growing adoption of ORCID identifiers by an increasing proportion of the research community there still exist software solutions such as Symplectic Elements to help researchers, institutions and funders manage the link between publications, researchers and grants. However, automated assignment is four times as challenging for the team behind Dimensions. The aim of Dimensions is to connect a researcher to all their research objects across at least four content sources: grants, publications, clinical trials to patents. Consequently, we have invested a lot of resources to developing an automated researcher disambiguation process that takes into account not only the metadata in each of the content sources but also the publicly available ORCID data to provide the best outcome that can be achieved at this point.

In technical terms, Dimensions has taken an approach that focuses more on precision and less on recall of the disambiguation routine. This is because we believe that assigning the wrong publications and other documents to a researcher is worse than suggesting an incomplete record since data errors undermine the trust in the results and can be highly 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.

Citations, acknowledgements and adding context

Extracting references - creating a network across sources

The extraction of the references and links between the different content sources is key to Dimensions. Our aim is to allow a user to gain a far superior understanding of the context of a piece of research by eliminating the walls and 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. The user is then able to draw conclusions and gain new insights, which previously would have taken an enormous amount of effort.

References between the different records are either harvested from existing databases (such as Crossref, PubMed Central, Open Citation Data) or extracted directly from the full text record provided by the content publisher. This is not only limited to journal publication references, but also includes acknowledgement and citation from and to books, conference proceedings, patents, grants and clinical trials.

PUBLICATIONS
Associated data
Publication references
Publication citations
Supporting grants
Patent citations
Linked clinical trials
Policy document citations

PATENTS
Publication references
Supporting grants
Patent citations
Patent references

POLICY DOCUMENTS
Publication references

CLINICAL TRIALS
Linked publications
Supporting grants

GRANTS
Resulting publications
Resulting patents
Resulting clinical trials

More than 1.3 billion references between documents

In total, we have extracted more than 1.3 billion direct connections between the document records, with more than 1 billion between publication records alone. This number is continually growing as we integrate more content, as we improve the representation of the content from more and more publishers, and as we work on perfecting our extraction routines.

Broadening the view beyond publications - bringing content together from as many places as possible

How does Dimensions compare to other databases like Google Scholar, Pubmed, Scopus or Web of Science?

Not comparable - a new and innovative approach, linking grants, publications, 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 (Dimensions has about 25% more publication records). But Dimensions transcends these existing tools and databases: The bringing together of grants, publications, clinical trials, patents, and policy documents, consistently linked and contextualised, opens up a world of proper discovery, research planning and impact communication possibilities. In addition, the Dimensions user interface presents search results in context allowing a user to understand the setting of a search result at a glance, while at the same time, facilitating greater exploration of potentially relevant works, funding or routes to impact.

Dimensions provides:

- A solid citation graph of the kind 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 with a powerful API, allowing a 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.

Citation counts in different systems and databases - there is no single truth!

Citations counts - why do they differ?

One question we are asked when talking about Dimensions is, 'how does our citation count compare to Google Scholar, Scopus or Web of Science'? As much as we would like to be able to give a simple answer, it is not possible. First of all, Dimensions and the reference that it contains is 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 databases and content to build its citation graph
- each database may include content from different date ranges (e.g. 1996 to present)
- each database may include 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 pre-prints or e-prints (where versioning and disambiguation of pre-print and post-print versions of the same paper adds yet more complexity)
- the frequency at which the content is updated differs by database, 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, and in other cases false positives may occur.
- as algorithms for matching improve and new data sources become available, reference graphs may be updated, resulting in changes to citation counts.

While spot checking Dimensions records against source data we found that for some articles we were under reading citation counts while for other publications our counts were notably higher than publicly available higher citation sources. We know that there are some fields where we need to engage with more publishers or more funders for greater coverage. Likewise, we know that there are some geographies where more work is needed to achieve greater patent coverage. As ever, we look for feedback from the community to prioritise our development focus for content integration.

As an illustratration, from the available data, an example from PLOS One:

"FastTree 2 – Approximately Maximum - Likelihood Trees for Large Alignments"

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

Dimensions:	3,041
Scopus:	2,691
WoS:	2,713
CrossRef:	2,658
Google Scholar:	3,777

Given the many variables described above, it is not possible for multiple parties to arrive at a single absolute count. As a result, in practice many researchers consider citations counts to be a useful relative metric when comparing other content within a single system.

Dimensions and the underlying data is an ongoing effort

The current content scope and quality is just the starting point

It took a large amount of effort and resource to bring all the current sources and content together - we consider this only to be a starting point:

- Grants are added continuously - every few months new funders and their portfolios become part of the Dimensions data universe
- We are going to add more publication data, new patent offices, new clinical trial registries and publishing organizations of policy documents during the course of 2019
- Pre-prints will be consistently integrated
- We continue to support publishers who wish to work with us to make their content more discoverable in Dimensions.

A joint effort to improve the data - please be as critical as possible!

Most important to us is your input and feedback. We are looking forward to being challenged and to receiving many suggestions from you as to where we can improve the data. We already have a long list of tasks from our development partners and friends, but we can always be better! This is clearly a team effort and we need you as the users, the research community and the broader Dimensions team!

Grants - a real glimpse into the future

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. The aim of their "pitch" 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. That position in the research cycle makes it a very special 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.

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

ÜberResearch (one of the six businesses in the Digital Science portfolio creating Dimensions) was founded in 2013 to work with research funders on aggregating a large grant database. Its 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. ÜberResearch's early effort has now become part of the new and broader version of Dimensions, which covers the entire flow from input to academic attention, commercialization, policy formulation and routes to impact.

ÜberResearch aggregated a grant database with $1.5 trillion in funding

Grants are a difficult content source for several reasons: They do not follow a common metadata schema in the way that publications do, nor do they yet have a persistent identifier such as the DOI; 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 a skew toward block funding, which will never show up in a funded grants database. Of course, 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 grant data should not be taken as a complete view on all research related funding, as we pointed out in a recent report. 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.

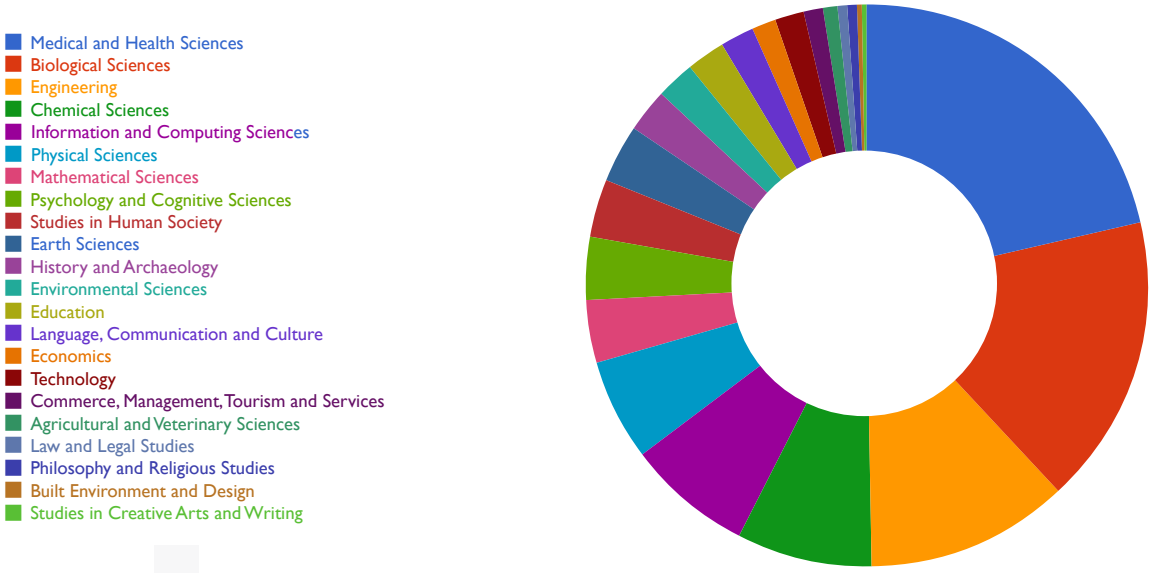
Grant data provides particular insights, not a complete research funding view

Key statistics on the Dimensions grant data

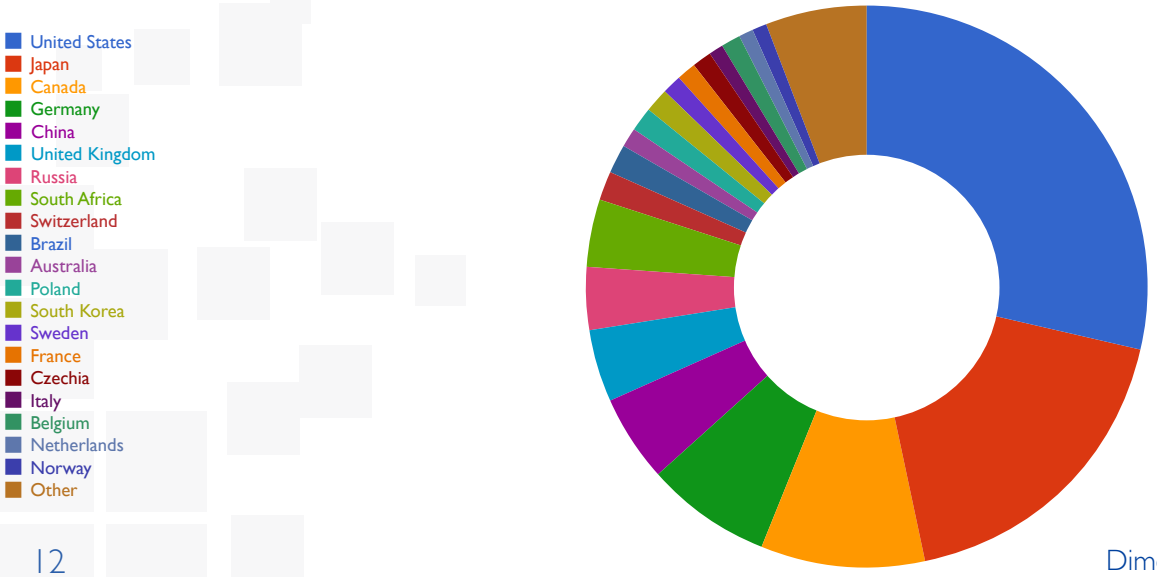
The following key statistics have been captured on April 1, 2019 and are changing on a monthly basis - this means that the values in this document can vary from the actual results in the Dimensions application or API.

Grants	4.6 million
Research funders covered	>340
Total funding amount	USD 1.5 trillion
Average funding amount	USD 403,000
Total amount of funding of projects active in 2019 and beyond	USD 341 billion
Number of links to research organizations (GRID IDs)	4.6 million
Number of links to researchers (Researcher IDs)	6.3 million

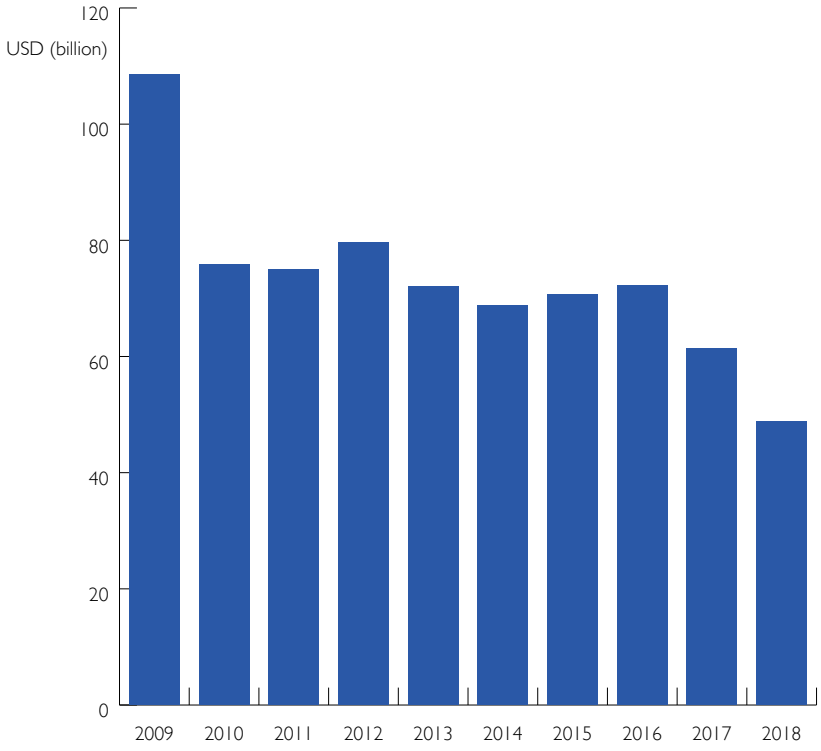
Distribution of funded projects across disciplines



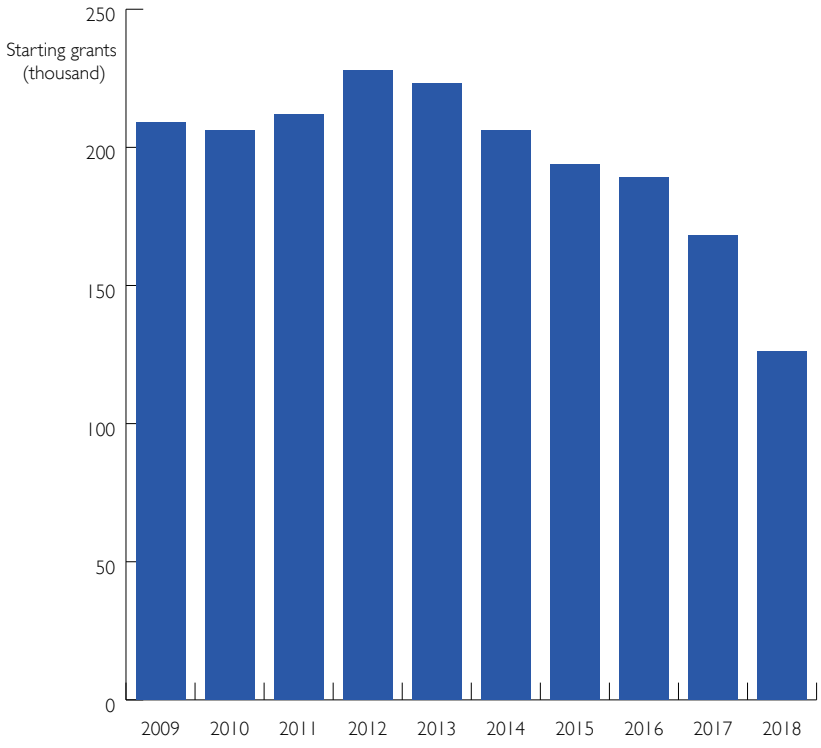
Geographical distribution of grants (total number of countries = 210)



Aggregated funding amount of starting grants over time



Number of starting grants over time



Publications, books and citations

Lack of innovation due to data being 'locked up'

Dimensions and publications / citations - a database, not a judgement call

With Dimensions, a powerful publication and citation database has been made available to increase access and usage of metadata for researchers and institutions, which has been, for a long time, an aspiration for Digital Science. An uncompetitive landscape has led to a slower-than-desirable pace of innovation to support researchers in many use cases. Rather than a lively research-led discussion about the needs of researchers, administrators and evaluators, there has been a narrower approach born of historical legacies both technological and practical as well as specific drivers from the research policy arena.

Dimensions - not a replication of the usual approach - a different approach

But it was clear that simply replicating existing approaches to create a third (or fourth, or fifth, depending on how you classify and count) abstracting and indexing database would not be in the sector's interest, so we decided to do two things in a fundamentally different way:

- Dimensions should be open to integrate all relevant research objects - in essence, less editorial choices over the content to be included (within reason, predatory journals, for example, clearly need to be treated differently)
- Consistent integration and linking of other sources (grants, patents and more.) treated in on the same basis as publications.

As comprehensive as possible - the decision power belongs in the hands of the user, not a vendor

Making the Dimensions database as comprehensive as possible is a central driver. We firmly believe that technological advances have led to different expectations from users. People no longer expect or desire that a search engine should filter content based on the preferences of a vendor. Indeed, as we write this report, net neutrality is becoming a big issue and in a very real sense we are consciously choosing to be neutral with respect to the content that we index and display to users. This means that we should not make the decision as to what is a 'worthy' research output (e.g. journal) to be included 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. Rather, it is our job to give users the best tools to navigate content and arrive at the most relevant results in the most efficient way.

Quality related filters: whitelists and blacklists as tools for the user

To ensure that users have the tools that they need to make content the right content filtering decisions for their use case we have implemented features that allow the user to limit the results that they obtain to certain subsets. The standard filters are specified by pre-defined, curated lists, which can be white

or black lists. We started our list definition with accepted openly available listed defined by others in the community, but are looking forward to receive new suggestions, again from the research community.

At launch, the following journal lists have been implemented in Dimensions:

- **DOAJ list**: Directory of Open Access Journals (DOAJ) is a community-curated online directory that indexes high quality, open access, peer-reviewed journals. The DOAJ journal list includes over 10,000 journal titles covering all areas of science, technology, medicine, social science and humanities.
- **ERA list**: The ERA 2015 journal list was designed by the Australian Research Council (ARC) in cooperation with the National Health and Medical Research Council (NHMRC) and the broader research community, with the purpose of supporting Australia's national research evaluation framework, Excellence in Research for Australia (ERA). Included are journals that were eligible for institutions' ERA 2015 submissions. We will include the ERA 2018 list as a filter once it is released.
- **Norwegian Register**: The Norwegian register, officially the 'Norwegian Register for Scientific Journals, Series and Publishers' is operated jointly by the Norwegian Centre for Research Data (NSD) and the National Board of Scholarly Publishing (NPU). The list shows which scientific publications are recognized in the weighted funding model and includes around 30,000 source titles.
- **PubMed list**: PubMed is a search engine of the abstracts and references of life science and biomedical publications mainly sourced from MEDLINE, and maintained by the United States National Library of Medicine (NLM) at the National Institutes of Health (NIH). The PubMed filter in Dimensions filters to only publications which have a PubMed identifier (PMID), as used in PubMed (website).

DOAJ, ERA list, Norwegian Register and PubMed

These filters are just a starting point and only address specific use cases. We are keen to learn about other general, national or institutional filters that should be considered, as well as different use cases where other lists may be helpful and welcome feedback so that we can develop this concept further.

Any idea for an additional 'quality' list? Please get in touch!

Aggregating the Dimensions publication and citation data

The publication and citation content in Dimensions is aggregated in a complex process. Below we sketch the key points in a two-step process for those who are keen to understand "why the data looks that way".

Step 1: Creating a backbone

An extensive metadata backbone was assembled and is continuously updated. This data spine integrates data from many sources, including openly-available databases together with those with permissive content licenses, such as PubMed, PubMed Central and Crossref. This initial step resulted in a large

100 million publication metadata records assembled

index of uniquely-identified publications containing about 100 million records. The Crossref records associated with a DOI sourced from the publishers among the 12,442 Crossref members form a significant core of this spine. This provides the Dimensions database with a very robust metadata backbone, but even with this great resource there are some limitations on metadata completeness, most notable is affiliation data for authors.

Step 2: Enhancing the data

The metadata records resulting from Step 1 are enhanced by processing full-text records, where those have been made available to us, significantly improving discoverability of content.

69 million records enriched - from more than 100 publishers already

This step includes deriving reference/citation data from the full-text and mining acknowledgements sections to identify links to funded projects, research funders and clinical trials. This step has been completed for more than 69 million full-text records, some open access but many made available to Digital Science for such purpose. These records are sourced from more than 100 publishers including some of the largest STM publishers in the world. Searching Dimensions will quickly indicate where we have coverage.

A key part of this data enhancement Step is that we are able to index full-text records. This means that a user can search for any term in a paper - it doesn't have to be in the title or the abstract. In concert with the filtering mechanisms that we've put in place for users, this means that you are increasingly likely to locate the research work that you're looking for.

New content added continuously

New publication data is added as more and more publishers join the effort and make their content more discoverable. Over the last 12 months, we have focused primarily on the large- and medium-sized publishers to be ready for the launch of Dimensions. If you are a publisher and want to see your content representation improved in Dimensions - just reach out to us via this form and we will be in touch.

Beyond academic attention - Altmetrics data 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, score, unique sources like policy documents and university syllabi, and the always popular Altmetric Top 100.

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, in a number of countries, 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.

Open Access, Open Citation Data and Dimensions

Open Access in Dimensions

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

Dimensions is an example of the power of making metadata including citations publicly available, in order to stimulate innovation and novel solutions / tools. Dimensions has been developed with the same goal in mind: Making good quality, consistent and linked metadata available to the community not just to ensure access for all but 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 aligned with the very important Initiative for Open Citations. Indeed, Dimensions is an example of what can be done if citation data is more openly available. In building Dimensions, Digital Science had to invest significant effort to make a good enough citation graph so that a good quality discovery experience could be delivered to users. We hope that the I4OC and similar initiatives continue to lower that barrier going forward. This will allow 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 the Dimensions team, both Altmetric and Figshare are members of the initiative.

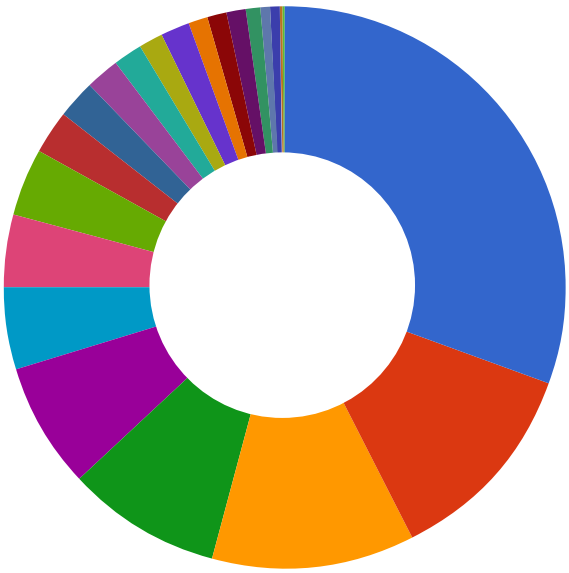
Key statistics on the Dimensions publication and citation data

The following key statistics were captured on April 1, 2019 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	100 million
Source titles covered (Journals, Book series, Preprint server, Conference proceedings)	More than 50,000
Number of links to research organizations (GRID IDs)	158 million
Number of links to researchers (Researcher IDs)	209 million
Number of cited references	1.1 billion
Number of links to grants	11 million
Number of links to funders	17 million
Number of links to clinical trials	891,000

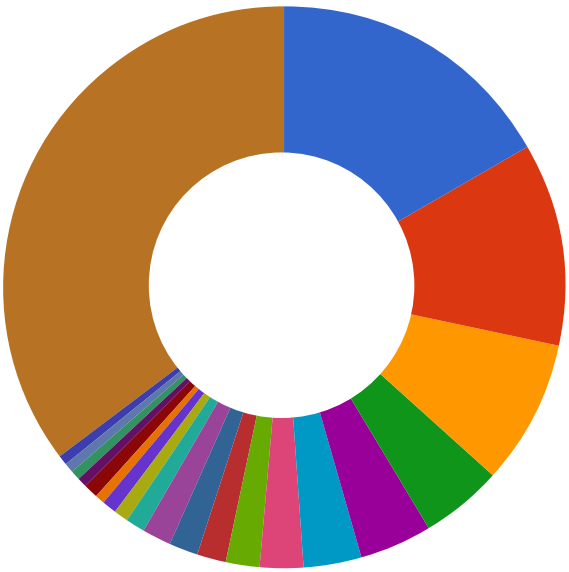
Distribution of publications across disciplines

- Medical and Health Sciences
- Biological Sciences
- Engineering
- Chemical Sciences
- Information and Computing Sciences
- Physical Sciences
- Mathematical Sciences
- Psychology and Cognitive Sciences
- Studies in Human Society
- Earth Sciences
- Technology
- Economics
- Environmental Sciences
- History and Archaeology
- Language, Communication and Culture
- Education
- Commerce, Management, Tourism and Services
- Agricultural and Veterinary Sciences
- Law and Legal Studies
- Philosophy and Religious Studies
- Studies in Creative Arts and Writing
- Built Environment and Designng



Distribution across publisher

- Elsevier
- Springer Nature
- Wiley
- Taylor & Francis
- IEEE
- Oxford University Press
- SAGE Publications
- Wolters Kluwer
- JSTOR
- Cambridge University Press
- American Chemical Society
- De Gruyter
- BMJ
- IOP Publishing
- AIP Publishing
- Thieme
- Royal Society of Chemistry
- American Medical Association
- American Physical Society
- SPIE
- Other



Clinical trials - research results en route to clinical application

Clinical trials, aggregated from different registries

To be clear about definitions: A clinical trial is any research study that prospectively assigns human participants or groups of humans to one or more 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 April 2019 we have integrated 10 registries:

Registry name	Country / Territory
ClinicalTrials.gov	United States
UMIN-CTR	Japan
EU-CTR	European Union
ISRCTN	International
CTRI	India
ANZCTR	Australia / New Zealand
CHICTR	China
GCTR	Germany
NTR	Netherlands
CRIS	South Korea

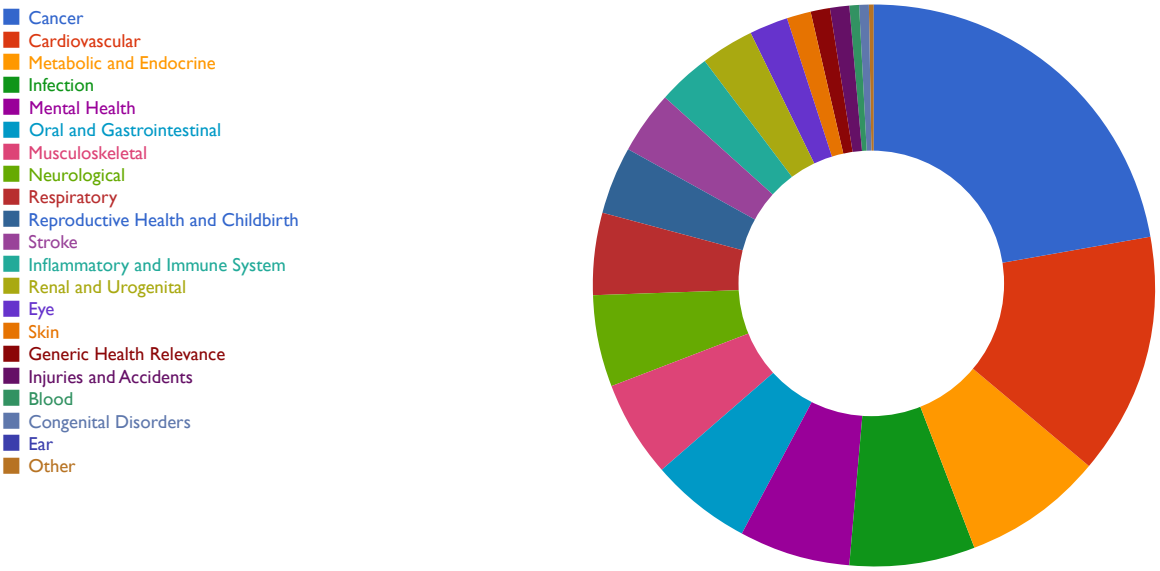
More will follow in the future. We integrate and map all relevant source data into Dimensions' coherent data model with filters, for e.g. research categories, research organizations or years, applicable across content types.

Key statistics on the Dimensions clinical trials data

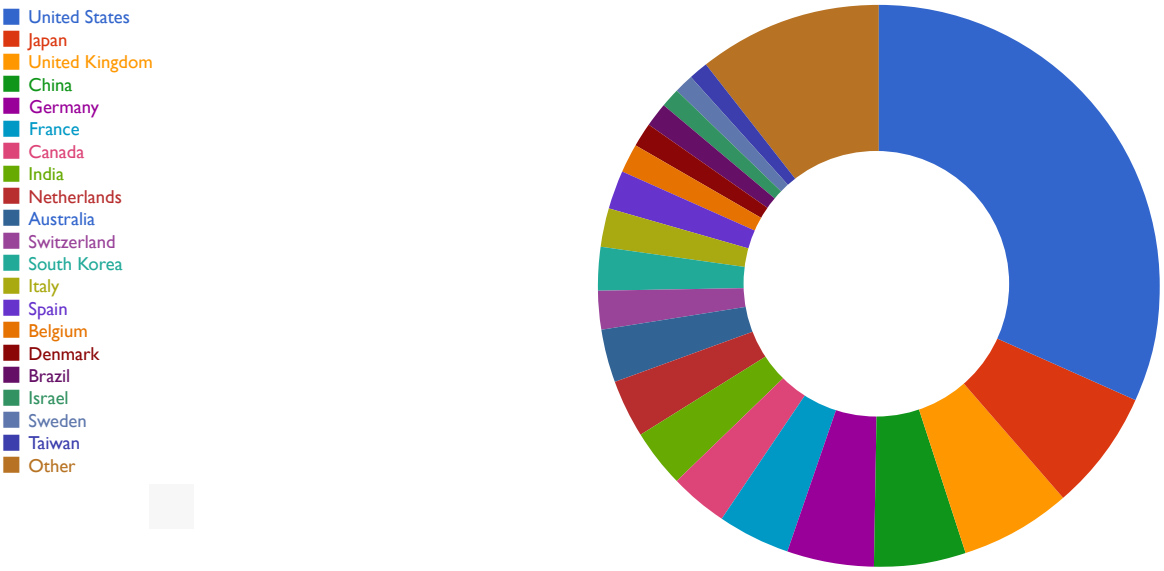
The following key statistics were captured on April 1, 2019 and are changing on a daily basis - this means that the values in this document can vary from the actual results in the Dimensions application or API.

Clinical trials	455,000
Clinical trial registries covered	10
Number of links to sponsors / collaborators (GRID IDs)	1.3 million
Number of links to publications	441,000
Number of links to grants	22,000
Number of links to funders	571,000

Distribution of clinical trials across disciplines (based on the Health Research Classification System (HRCS) from the UK)



Geographical distribution of clinical trials



Patents - research resulting in practical and commercial applications

Patent data - to show the translation of research activities into the commercial space

We started with an initial tranche of patent offices for the launch of Dimensions. We are now in the process of adding more, which will appear in Dimensions during the course of 2019. 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 and potential use of research results.

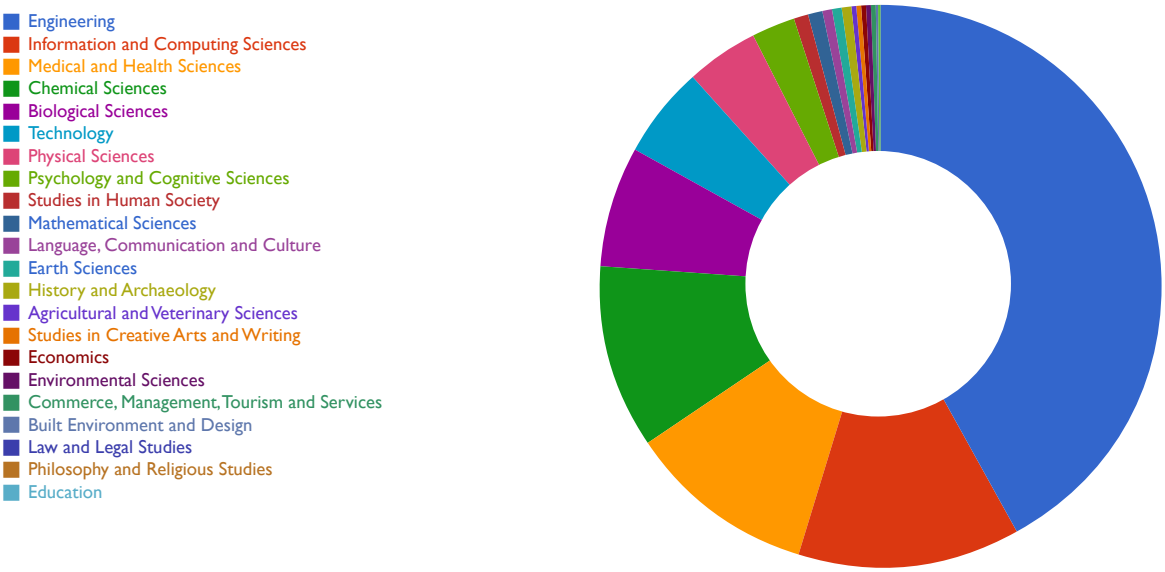
Office name	Country / Territory
United States Patent and Trademark Office (USPTO)	United States
European Patent Office (EPO)	Europe
World Intellectual Property Organisation (WIPO)	International
German Patent and Trademark Office (DPMA)	Germany
Canadian Intellectual Property Office (CIPO)	Canada
Intellectual Property India (IPI)	India
Intellectual Property Office (IPO)	United Kingdom
National Industrial Property Institute (INPI)	France
Intellectual Property Department (IPD)	Hong Kong
Russian Patent Office	Russia

Key statistics on the Dimensions patent data

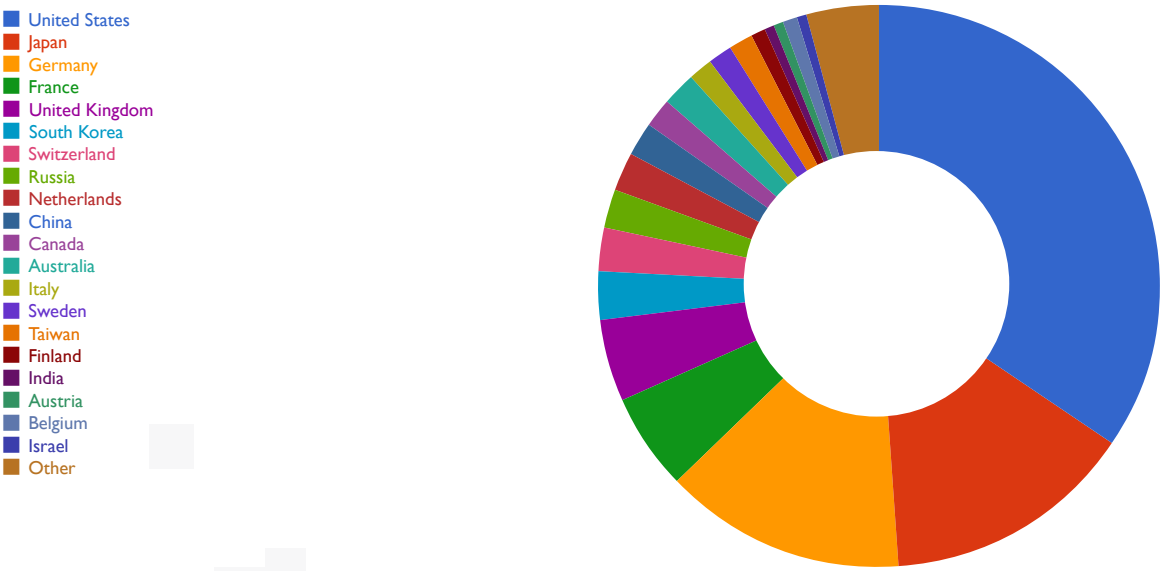
The following key statistics were captured on April 1, 2019 and are changing on a weekly basis - this means that the values in this document can vary from the actual results in the Dimensions application or API.

Patents	38 million
Patent offices covered	10
Number of links to research organizations (GRID IDs)	37 million
Number of cited patent references	227 million
Number of links to publications	10 million
Number of links to grants	165,000
Number of links to funders	221,000

Distribution of patents across disciplines



Geographical distribution of assignees



Policy documents - research resulting in policy and guidance documents

The policy document data in Dimensions is provided by the Digital Science portfolio company Altmetric. It 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. We aim to curate a broad scope of policy sources from organisations around the world and cover topics from climate change to health, transport and economics. Wherever possible we deep-index the full text, allowing us to categorize the record and extract references.

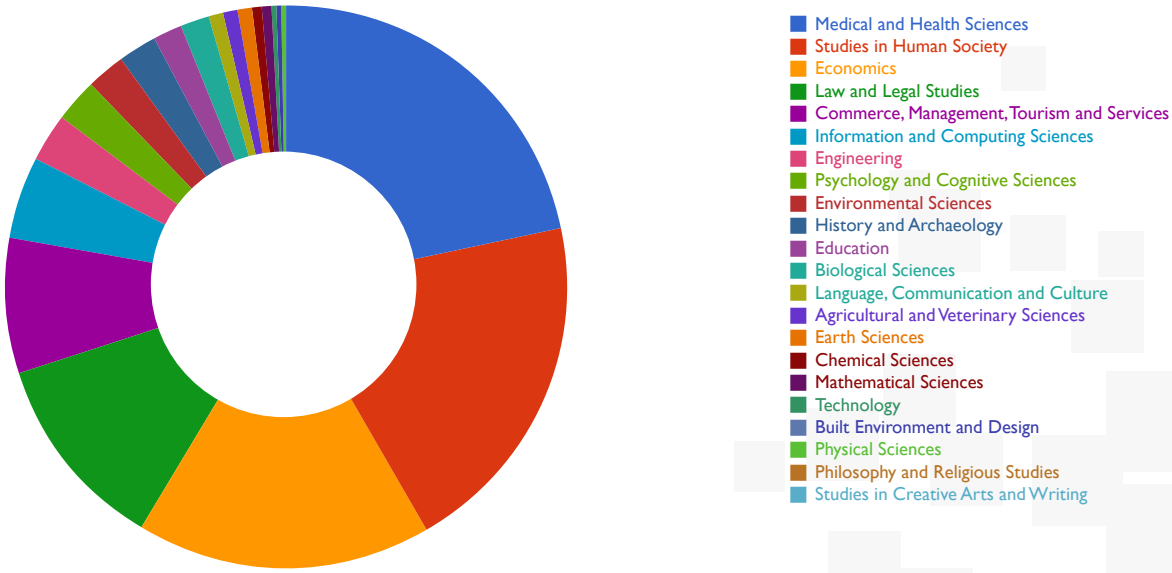
Policy documents from over 70 publishing organization

Key statistics on the Dimensions policy document data

The following key statistics were captured on April 1, 2019 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.

Policy documents	421,000
Publishing organizations covered	72
Number of links to publications	1.5 million

Distribution of policy documents across disciplines



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

www.dimensions.ai digital-science.com