

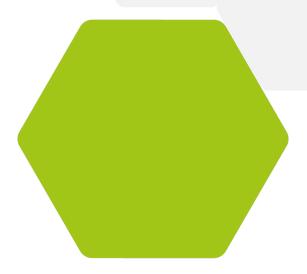
Just Add Context: Levelling up analysis with Dimensions on Google BigQuery

11 February 2021 Danu Poyner - Dimensions Product Specialist (APAC) d.poyner@digital-science.com

Part of DIGITAL SCIENCE

Today's Agenda

- 1) 2-minute Intro: What is Dimensions?
- 2) 2 Lightning Examples
 - a) Covid-19 Research Dashboard
 - b) NZ research collaboration with low-income countries
- 3) How can you get started with Dimensions on Google BigQuery?
- 4) Q&A / Discussion

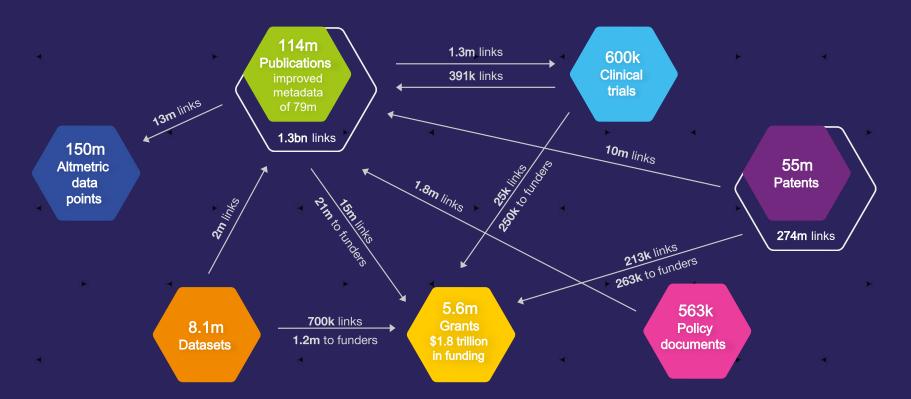




2-minute Intro:
What is Dimensions?

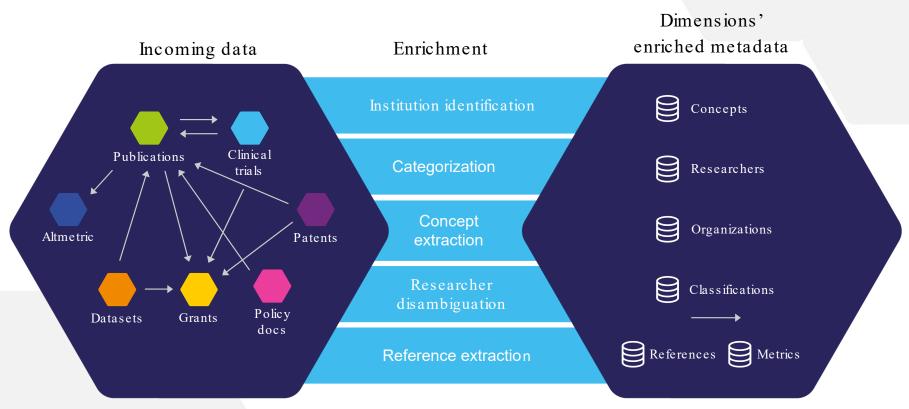


Dimensions: A joined-up perspective on research





A modern approach to data curation





How our academic clients are using Dimensions...

Chancellery, Deans & Planning

Strategic planning & competitive intelligence

- Peer group analysis & comparisons
- Internal benchmarking
- Cost-effectiveness of research
- Horizon scanning

Advocacy

• Impact, value of institutions & research

Talent planning & retention

Networks & context

• Industry collaboration & revenue generation

Research Office

Rankings & assessment

- National assessment
- International ranking
- Custom benchmarking

Impact assessment and narrative

Funding & grants

- Strategy
- Sources (incl. industry)
- Trends
- ROI assessment

Talent

- Recruitment & capability building
- Career tracking

Collaboration networks

Library

Collection Development

- Holdings & ROI assessment
- Publisher negotiations & transformative agreements

Open access

- Trends
- Citation and impact analysis
- Compliance tracking

Publication strategy

• Support researchers on where to publish

Information hub

• High value data asset for training

Faculty & Research

Strategy

- Funding trends & opportunities
- Faculty output tracking
- Find partners and recruit talent
- Emerging research trends
- Patent analysis

Bibliometrics

• Co-citation analysis

Use for research

- e.g. Computer Science, Business, Economics, Social Science
- Dataset for analysis and visualisation

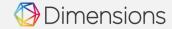
And Beyond

Commercialization of research

- Science-->Innovation
- Tech transfer
- Support university spinouts & deep tech
- Patent analysis
- Find Key Opinion Leaders
- Industry partnerships

Marketing & Communications

- Market segmentation & targeting
- Monitor & grow wider impact & attention, publicity for research



One amazing database – multiple ways to access it

Web App



Search & discovery; top analytical use cases

Dedicated UI, inbuilt visualizations

In the browser, no specialized knowledge required

For everyone

API



Ad hoc analyses & topic modelling

Full-text search & special functions e.g. affiliation extraction

Product integrations e.g. CRIS

For API users + data & analytics teams

Google BigQuery



Fast, large scale analyses; dynamic dashboards

Join private & public data, access previously unsurfaced links

Direct integration with BI tools e.g. Tableau, Qlik, PowerBI

For data & analytics teams + dashboards for everyone



So what?





Dimensions on BigQuery: So what?

Ultimate flexibility

Analyze however you want: ask and answer questions that you've always wanted to ask.

Speed and agility

Aggregate and manipulate data at a speed that was simply impossible before, to create bespoke real-time insights.

Interoperability with other platforms and data

Integrate with your existing infrastructure, join with your own data, pull into your visualisation tool of choice (e.g. Tableau, Qlik, PowerBI, Plot.ly etc).



Example 1:

A dashboard for exploring COVID-19 research in near real-time



Dimensions COVID-19 Dashboard

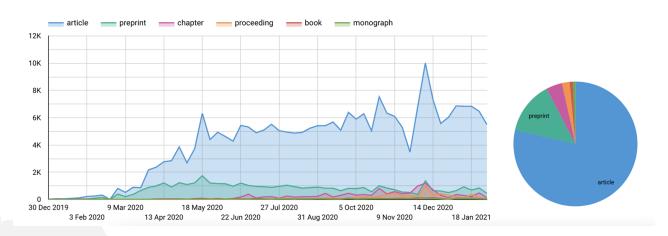
Publications

Report pages: Publications | Clinical Trials | Funding

Total publications	Total organizations	Total countries		
308,782	21,460	196		

Weekly new publications, by type of publication

Records displayed are sorted based on the week they have been added to the Dimensions database. Note: all document types are included.



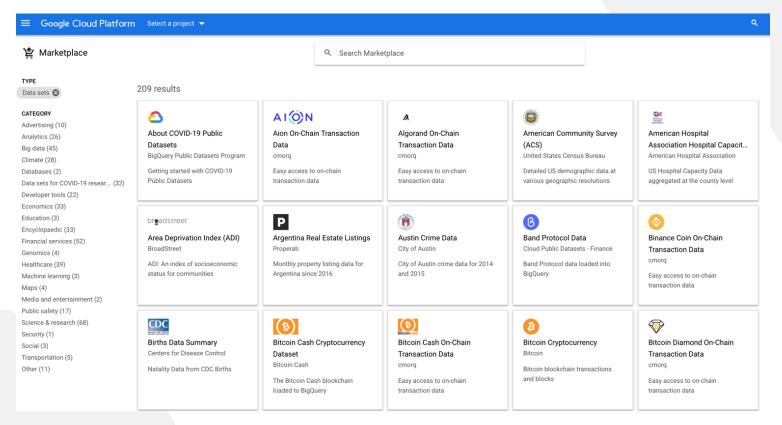


Example 2:

"How often does NZ collaborate on research with low-income countries?"



Hundreds of public datasets on GCP Marketplace...





World Development Indicators from The World Bank



World Development Indicators (WDI)

The World Bank

The primary World Bank collection of development indicators



OVERVIEW

SAMPLES

Overview

This dataset contains the most current and accurate global development data available including national, regional and global estimates. Data has been collected from the early 1960's to present and is updated regularly depending on new data available on the indicators. This time series data offers indicators such as agriculture and food security, climate change, population growth, economic growth, education, energy, natural Resources and many more.

This public dataset is hosted in Google BigQuery and is included in BigQuery's 1TB/mo of free tier processing. This means that each user receives 1TB of free BigQuery processing every month, which can be used to run queries on this public dataset. Watch this short video to learn how to get started quickly using BigQuery to access public datasets. What is BigQuery \(\mathbb{C} \)

Additional details

Type: Datasets

Last updated: 31/01/2020

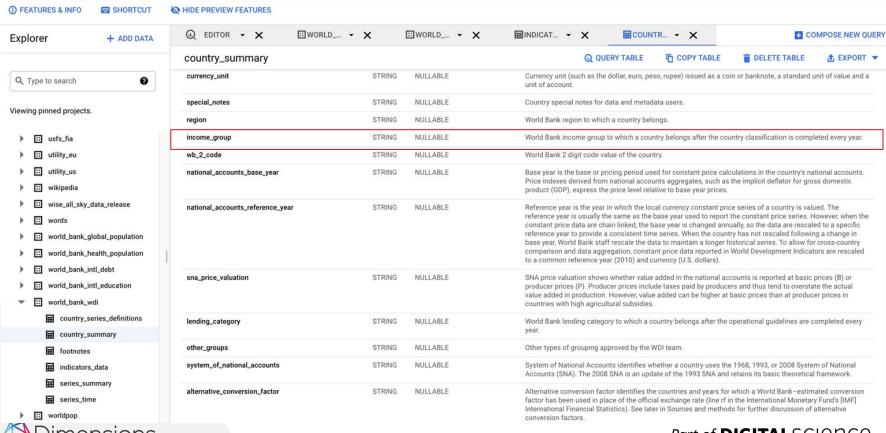
Category: Healthcare, Financial services, Economics

Dataset source: The World Bank ☑
Cloud service: BigQuery

Update frequency: Annually



World Bank data has an 'income_group' indicator



Part of **DIGITAL**SCIENCE

How often did University of Auckland collaborate on research with low-income countries from 2010 onwards?"

```
SELECT
        p.year ,
    count(distinct CASE WHEN income group = 'Low income' THEN p.id ELSE null END) low income,
    count(distinct CASE WHEN income group = 'Lower middle income' THEN p.id ELSE null END) lower middle,
    count(distinct CASE WHEN income group = 'Upper middle income' THEN p.id ELSE null END) upper middle income,
    count(distinct CASE WHEN income group = 'High income' THEN p.id ELSE null END) high income,
    count(distinct p.id) all publications
FROM
    `dimensions-ai.data analytics.publications` p,
       unnest(authors) al,
       unnest(al.affiliations address) aff1,
       unnest(authors) a2,
        unnest(a2.affiliations address) aff2
   inner join 'dimensions-ai.data analytics.grid' gl
       on aff1.grid id = g1.id
    inner join 'bigquery-public-data.world bank wdi.country summary' wb
        on gl.address.country code = wb.two alpha code
    WHERE
       aff2.grid id = 'grid.9654.e'
       and aff1.grid id != aff2.grid id
       and p.year >= 2010
    group by p.year
    order by year desc
```



How often did University of Auckland collaborate on research with low-income countries from 2010 onwards?"

		1			
year	low_income	lower_middle	upper_middle_income	high_income	all_publications
2021	5	42	158	458	535
2020	19	260	1056	3574	4049
2019	7	250	935	3311	3734
2018	8	243	773	3046	3369
2017	22	234	654	2810	3097
2016	9	185	591	2482	2759
2015	7	148	413	2165	2362
2014	5	143	415	2075	2280
2013	2	150	369	2020	2159
2012	5	144	324	1766	1885
2011	3	125	289	1504	1625
2010	1	30	156	1326	1399



How often did NZ institutions collaborate on research with low-income countries from 2010 onwards?"

name	low_income	lower_middle	upper_middle_income	high_income	all_publications	low_as_percentage
Canterbury Health Laboratories	22	50	47	368	369	5.962060
Middlemore Hospital	12	26	54	1048	1054	1.138520
University of Otago	153	949	2570	18274	19452	0.786552
Landcare Research	14	76	425	1741	1850	0.756757
Auckland University of Technology	35	327	1007	5241	5804	0.603032
University of Auckland	93	1954	6133	26537	29253	0.317916
Victoria University of Wellington	23	322	1230	7161	7775	0.295820
Massey University	23	553	1824	8135	9206	0.249837
University of Waikato	11	204	928	4300	4790	0.229645
University of Canterbury	15	1341	2477	8907	9625	0.155844



How many NZ-affiliated researchers have dual affiliation with low-income countries from 2010 onwards?"

name	country	low_income	lower_middle	upper_middle_income	high_income	all_researchers
University of Otago	New Zealand	17	41	209	2224	2379
University of Auckland	New Zealand	6	86	446	2989	3324
Auckland University of Technology	New Zealand	2	34	86	467	543
AgResearch	New Zealand	2	2	26	284	309
Massey University	New Zealand	2	56	145	869	1018
Lincoln University	New Zealand	1	13	42	205	253
University of Canterbury	New Zealand	1	36	106	760	864



How can you get started with Dimensions on Google BigQuery?



Why Google BigQuery?

- A cloud-based data warehouse where users can store and analyse their large datasets.
- We chose Google BigQuery as it is designed to be a very cost effective data warehouse: no ongoing maintenance costs, no set-up costs, and a low overall cost of ownership. Designed for fast analytics on massive datasets.
- Eliminates the need for expensive local server infrastructure.
- Work with the data immediately: Just log in to start querying.
 - Allows you to save valuable developer time: no need for an ingest pipeline or on-going maintenance.

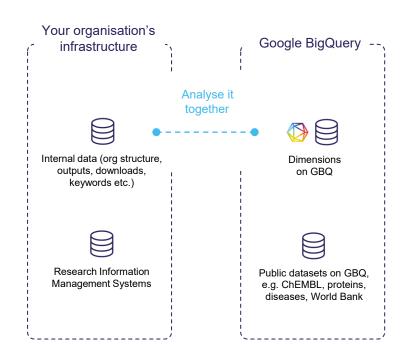
Infrastructure
as it should be barely visible,
hugely supportive
and affordable



Create broader context:

Add your own data

- Join with your own internal datasets securely.
- Expand to other public & private datasets.

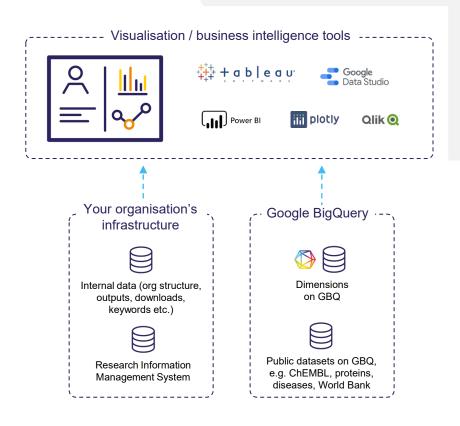




Augment your infrastructure:

Do all this with the tools you already use

- Direct connectors to BI & data viz tools.
- Programmatic access via client libraries.

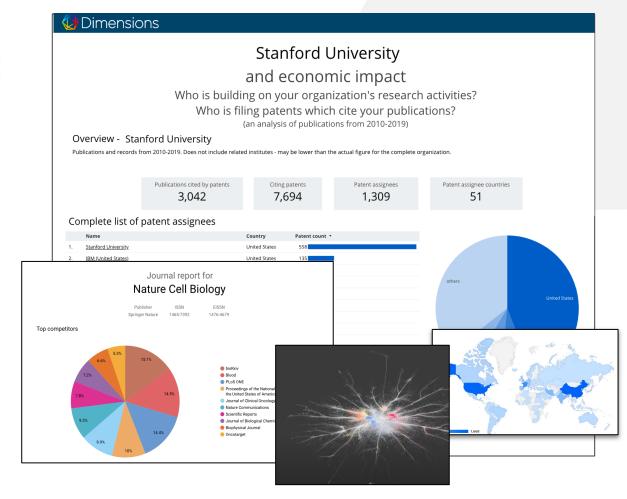




From idea to impact:

View and analyse Dimensions data however you want

- Same data, new ways to look at it.
- Build custom dashboards that provide answers on a specific topic.



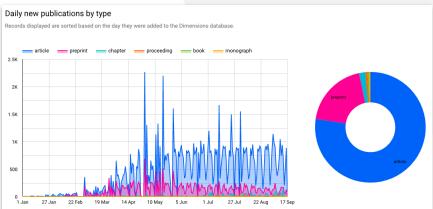


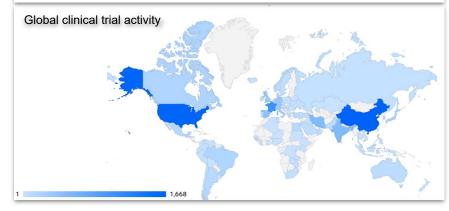
Try it out! All content relevant to COVID-19 has been made available openly on BigQuery

Daily new publications by type

- Explore how Dimensions on BigQuery works.
- Data covers:
 - Publications (180k+)
 - o Grants
 - o Datasets
 - o Clinical trials
- Take a look at our <u>interactive Covid-19</u> dashboard.
- Make your own dynamic dashboard using BigQuery + Google's free data visualization tool Data Studio (or connect to Qlik etc).

Find the data here, and explore our documentation here to get started.





Q&A / Discussion

Danu Poyner - Dimensions Product Specialist (APAC) d.poyner@digital-science.com

BOOK A TIME TO CHAT: https://calendly.com/d-poyner



Thank you!

Danu Poyner - Dimensions Product Specialist (APAC) d.poyner@digital-science.com

BOOK A TIME TO CHAT: https://calendly.com/d-poyner





ADDITIONAL SLIDES- NOT PART OF PRESENTATION



How does Dimensions on Google BigQuery work?



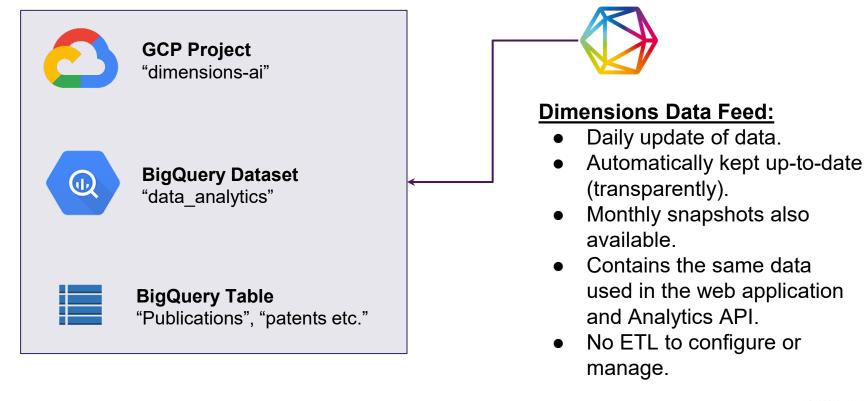
Google BigQuery

- Cloud Data Warehouse the ease/simplicity of Google Docs applied to data warehouses and analytics.
- Uses ANSI SQL for performing queries minimal learning curve from other OLTP and OLAP database systems.
- Highly scalable, easily capable of analysing petabytes of data.
- Zero infrastructure serverless approach completely managed by Google.
- Easy Sharing of datasets between colleagues, teams and external parties.



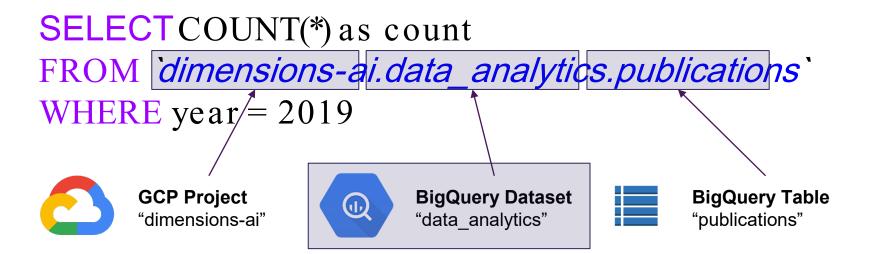


Dimensions on BigQuery: How it works





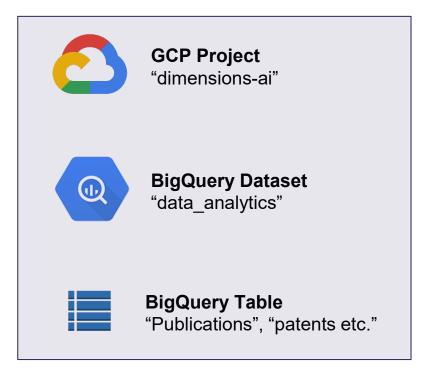
A Simple Query



Read access onto this dataset and tables contained within is provided. We keep it maintained with Dimensions data.



Dimensions on BigQuery: How it works



Data we share to you.



Your data – you control all access rights.



Another Simple Query

SELECT

d.dept_code, year, metrics.field_citation_rat category_for.first_level.codes as for_codes

Pulling data from Dimensions dataset based on a list of your own departments and DOI pairs (your data).

FROM

`lilliput-uni-analysis.2020_analysis.departments_doi`d

LEFT JOIN

`dimensions-ai.data_analytics.publications `p

ON p.doi = d.doi

Shared read access onto Dimensions datasets.



GBQ model = less need for JOINs

```
"first_level": {
    "codes":
            "id": "2211".
            "code": "11"
            "name": "Medical and Health Sciences"
            "id": "2213",
            "code": "13",
            "name": "Education"
            "id": "2217".
            "name": "Psychology and Cognitive Sciences"
```

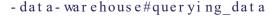
```
SELECT
COUNT(c.id) as tot, category
FROM
'covid-19-dimensions-
ai.data.clinical_trials' c,

UNNEST(category_for.first_level.full) AS
category

GROUP BY category
ORDER BY tot DESC

LIMT 10
```

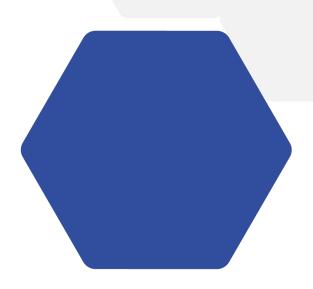






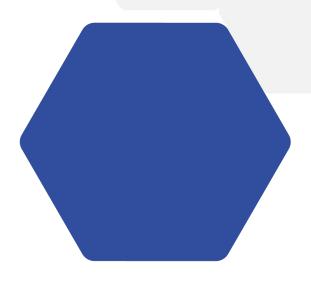
Moving data in and out

- Web UI import/export via CSV, JSON files, Google Sheets.
 - Allows for exploring the data, answering simple oneoff questions, preparing queries for other avenues below.
- Programmatic access Google provided REST API endpoints with client libraries allowing full control over BigQuery.
 - Execute queries, import/export datasets etc. (we run our Dimensions daily ingest ETL using the Google provided client libraries).
- BI Tool integrations most popular business intelligence tools allowing integrations, allowing data to be sourced from BigQuery.
 - o i.e. Data Studio, Qlik, PowerBI, Tableau



Programmatic Data Access

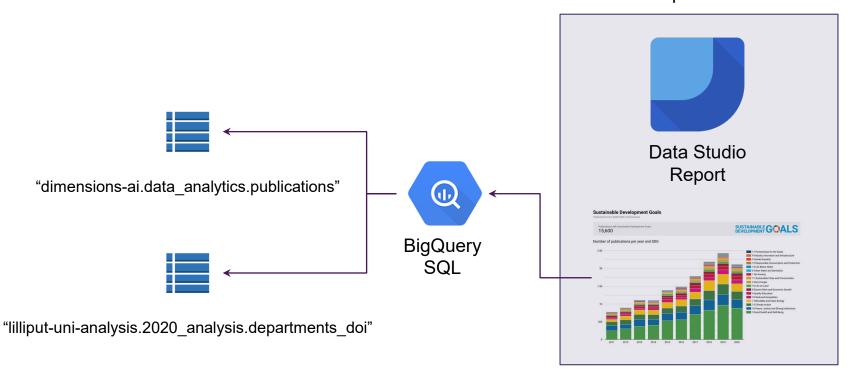
- Simple REST API to control all aspects of BigQuery (creating datasets, tables, performing queries etc)
- Client libraries provided by Google (Python, NodeJS etc).
- Advanced data access methods for bulk data access BigQuery Storage API
- BigQuery Storage API provides a binary serialised (Avro), transaction safe, parallelizable method for pulling down very large queries.
- Export Jobs export large tables/queries off onto Google Cloud Storage (JSON, Avro, CSV etc).







Shareable "Google Doc" like report link.



Sustainable Development Goals

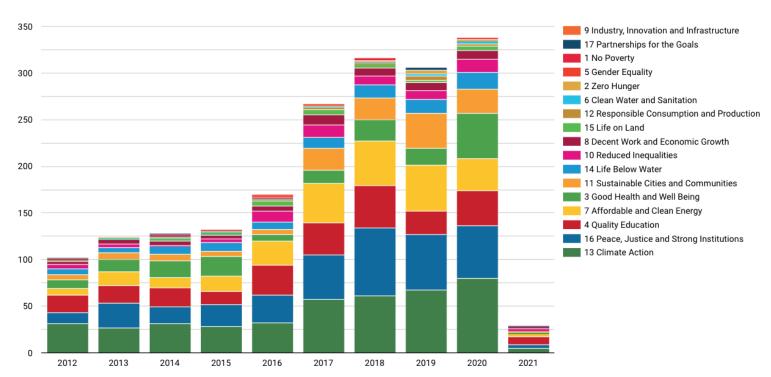
Publications from 2009-2020 in Dimensions.

Publications with Sustainable Development Goals

2,013



Number of publications per year and SDG

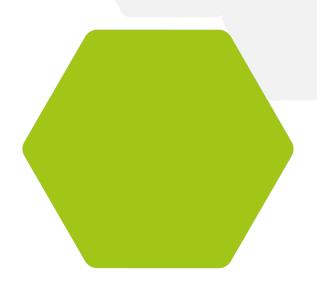


BigQuery Costs

Charges associated with Google BigQuery are based on two factors:

- How much data is involved (i.e. 'touched') in a query.
- How much data is stored on an ongoing basis in tables you own.

The UN Sustainable Development Goals (SDG) Data Studio report for example costs approximately a few cents to generate - even though this touches most of the different data sources over a wide range of years and does an international comparison against the organization in question.



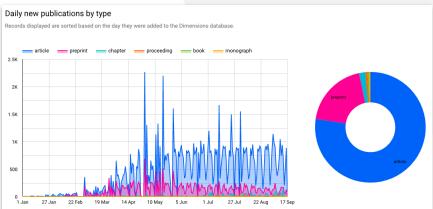


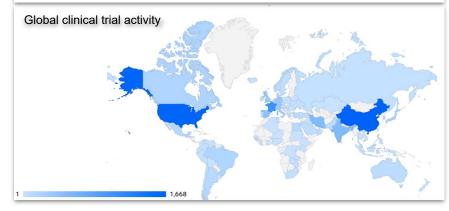
Try it out! All content relevant to COVID-19 has been made available openly on BigQuery

Daily new publications by type

- Explore how Dimensions on BigQuery works.
- Data covers:
 - Publications (180k+)
 - o Grants
 - o Datasets
 - o Clinical trials
- Take a look at our <u>interactive Covid-19</u> dashboard.
- Make your own dynamic dashboard using BigQuery + Google's free data visualization tool Data Studio (or connect to Qlik etc).

Find the data here, and explore our documentation here to get started.





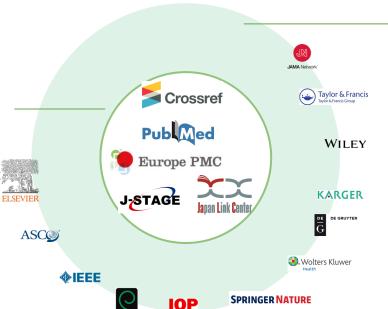
Data and indicators in Dimensions





Publications

- 115M records based on metadata
- 'Backbone' for Dimensions
- OA tagging
- Rule-based document type identification



Institute of

- Full text for 78,620,438 publications (direct relationships with >130 publishers)
- Improved representation compared to the 'backbone' record
 - Additional metadata
 - o Fill gaps
 - o Deep indexing



Open access data

- More than 32M open access publications
- Based on Unpaywall data integration and additional improvements by Dimensions
- All publications categorized as
 - o Closed
 - o Gold (Pure gold, Hybrid, Bronze)
 - Green (published, accepted, submitted)





Datasets



- More than 8 million datasets
- Sourced from DataCite and Figshare
- Linked to publications, supporting grants and funders
- Filters for research organizations, funders, researchers and more

DATACITE

Data Cite FIND, ACCESS, AND REDUSE DATA	TO DRYAD	PANGAEA.	** MENDELEY DATA	zenodo
Dataverse Project	ICPSR Sharing data to advance science	採	UK•DATA ARCHIVE	 OSF
<u>UNAVCO</u> _O	QDR The Qualitative Data Repository	SEAMOE	≰ GBIF	800+ more

FIGSHARE & FIGSHARE HOSTED REPOSITORIES

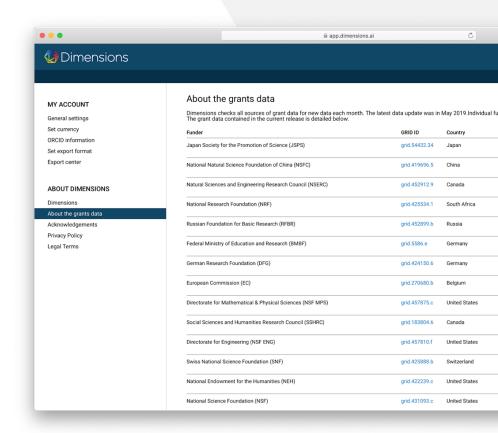
figshare	frontiers	SPRINGER NATURE	SeiFLO	IOP Publishing
© PLOS	WILEY	FICCOResearch	Karger	\$SAGE
Taylor & Francis Taylor & Francis Croup	ROYAL SOCIETY	CELL IMAGE LIBRARY	NIH	70+ more

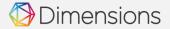


Grants data



- Project funding
- Over 5M grants from 600+ funders globally
- \$1.8 trillion of funding
- Sourcing
 - Direct relationships with funders
 - Data available via APIs
 - Data available via websites which we crawl

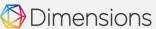




Patents data



- United States
- European Patent Office
- Germany
- World Intellectual Property Organisation (WIPO)
- Australia
- United Kingdom
- Canada
- Russian Federation
- France
- India
- Nicaragua
- Hong Kong
- New Zealand
- Switzerland
- Ireland





...most patent registries will be included in Dimensions by early 2021.

Clinical trials data



- ClinicalTrials.gov
- EU-CTR
- UMIN-CTR
- ISRCTN
- ANZCTR
- CHICTR
- NTR
- GCTR
- CTRI
- CRIS
- IRCT







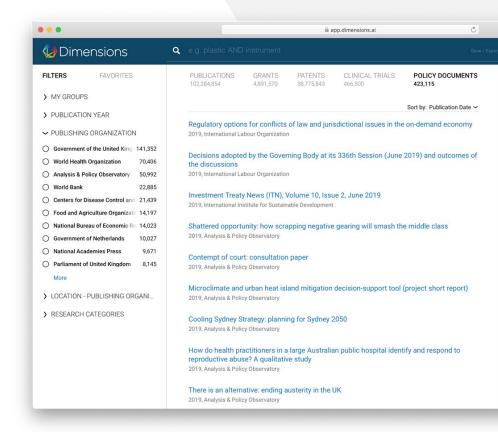
Policy documents data

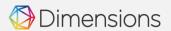


Over 500,000 policy document records, linked to publications

Including but not limited to:

- World Health Organization
- World Bank
- Centers for Disease Control & Prevention
- Government of the United Kingdom
- Analysis and Policy Observatory
- NZ Treasury





Publication indicators in Dimensions

Publication citations

The publication citations value is the number of times that a publication has been cited by other publications in the database. Citing publications can be of any publication type.

Recent citations

The recent citations value is the number of citations that were received in the last two years. It is currently reset at the beginning of each calendar year.

Field Citation Ratio (FCR)

The Field Citation Ratio (FCR) is an article-level metric that indicates the relative citation performance of an article, when compared to similarly-aged articles in its subject area. A value of more than 1.0 indicates higher than average citation, when defined by Field of Research Subject Code, publishing year and age. The FCR is calculated for all publications in Dimensions which are at least 2 years old and were published in 2000 or later.

Altmetric Attention Score

The Altmetric Attention Score is a weighted count of all of the online attention Altmetric have found for an individual research output. This includes mentions in the mainstream news, social networks, Wikipedia, blogs and more.

Relative Citation Ratio (RCR)

The Relative Citation Ratio (RCR), developed by the National Institutes of Health (US), indicates the relative citation performance of an article when comparing its citation rate to that of other articles in its area of research. A value of more than 1.0 shows a citation rate above average. The article's area of research is defined by the articles that have been cited alongside it.

The RCR is calculated for all PubMed publications which are at least 2 years old.

