Dimensions

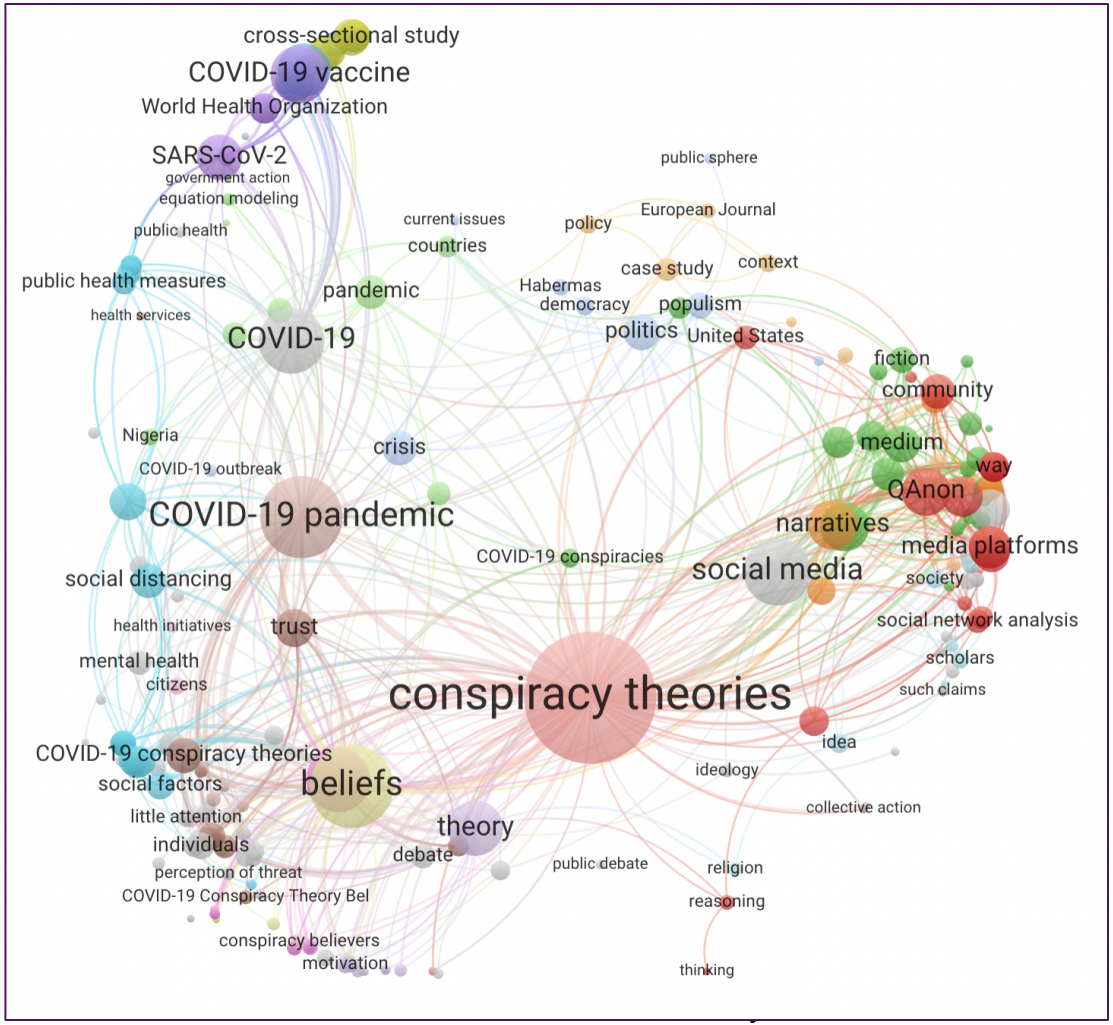
Generating large-scale network analyses of scientific landscapes in seconds using Dimensions on Google BigQuery

Michele Pasin
Head of Dimensions Data Solutions
m.pasin@digital-science.com

Richard J. Abdill
Data Scientist
richard.abdill@pennmedicine.upenn.edu

2022-09 STI / Granada

Part of DIGITAL science


cross-sectional study
COVID-19 vaccine
World Health Organization
SARS-CoV-2
government action
equation modeling
public health
public health measures
health services
pandemic
COVID-19
current issues
countries
public sphere
policy
European Journal
case study
context
Habermas
democracy
populism
politics
United States
fiction
community
medium
way
QAnon
narratives
media platforms
society
social media
COVID-19 conspiracies
social network analysis
scholars
such claims
idea
ideology
collective action
Nigeria
COVID-19 outbreak
crisis
COVID-19 pandemic
social distancing
health initiatives
trust
mental health
citizens
conspiracy theories
COVID-19 conspiracy theories
social factors
beliefs
little attention
individuals
perception of threat
COVID-19 Conspiracy Theory Bel
theory
debate
public debate
religion
reasoning
conspiracy believers
motivation
thinking

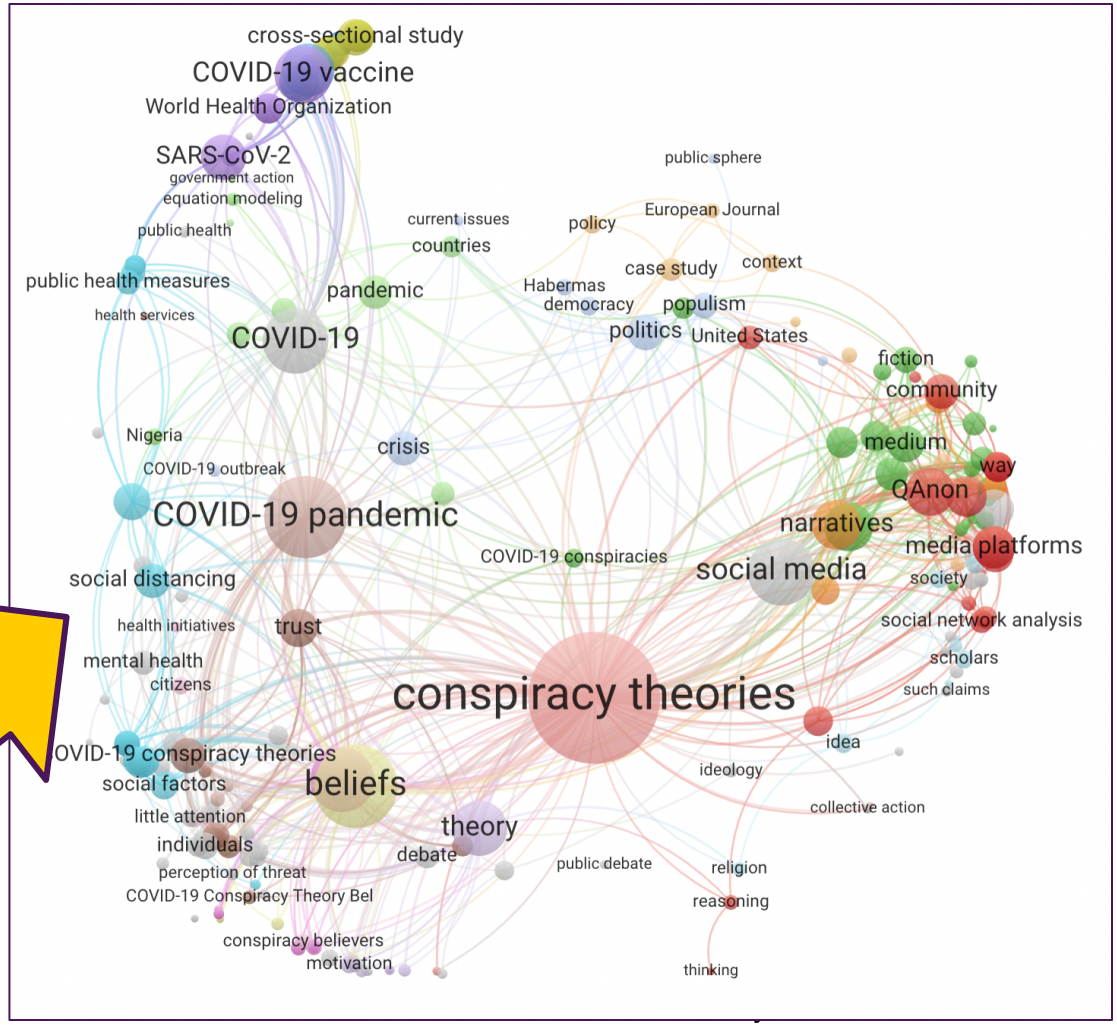
Dimensions



Google BigQuery

python™



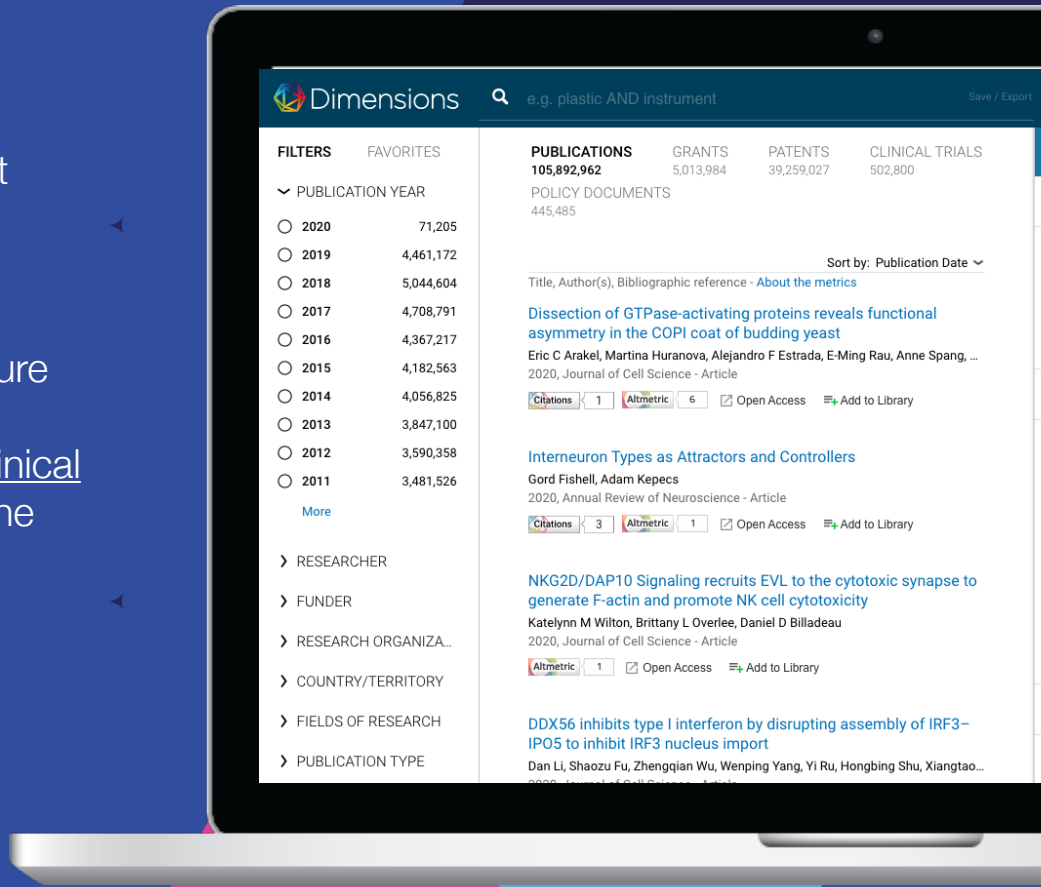


What is Dimensions?

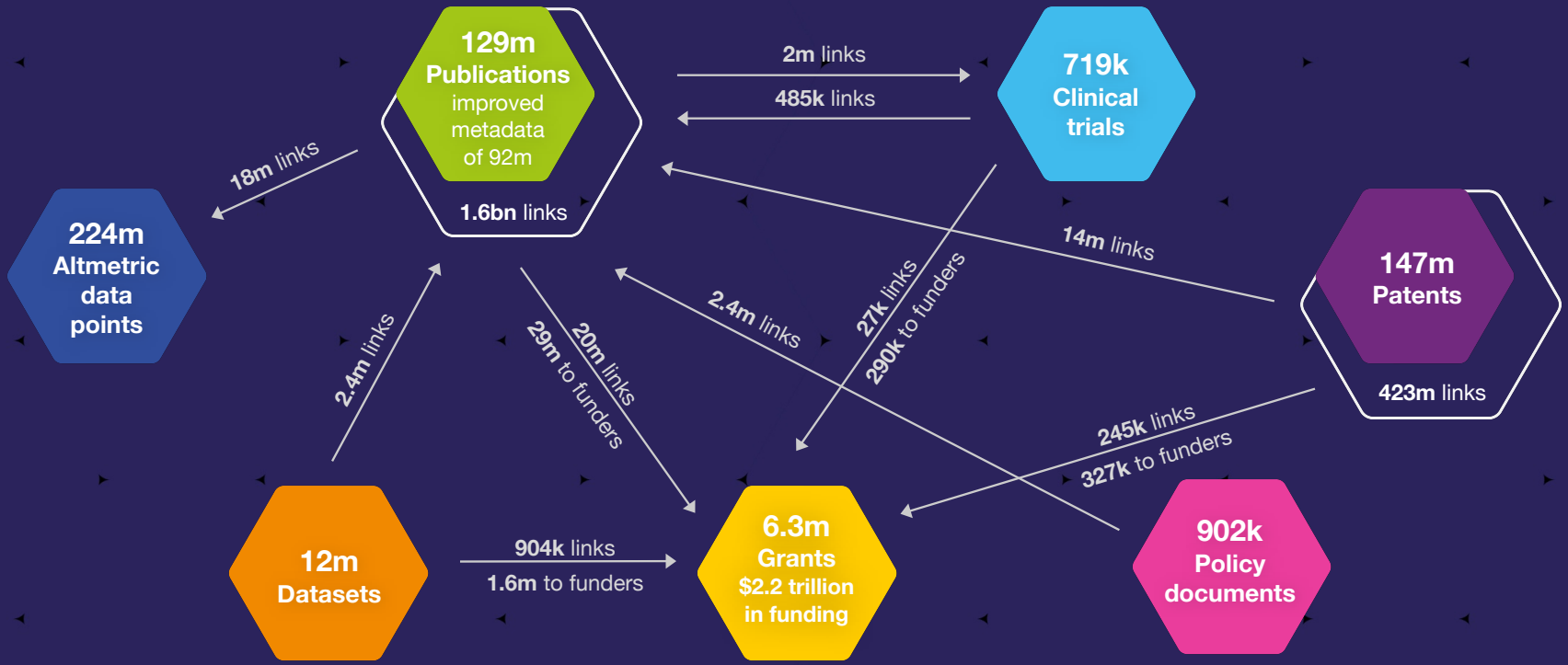
Dimensions delivers an array of search and discovery, analytical, and research management tools, all in a single platform.

- **Moving beyond citations** to better capture the research lifecycle: the **links among** publications, grants, policy documents, clinical trials, datasets, patents, researchers are the key to understanding this landscape.

https://www.dimensions.ai

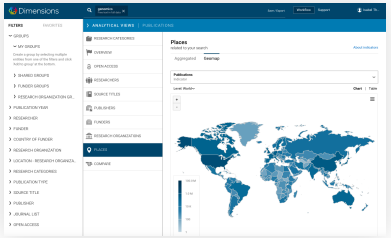


More than a citations database



Dimensions ‘flavors’

Web App



- Search & discovery; top analytical use cases
- Dedicated UI, inbuilt visualizations
- In the browser, no specialized knowledge required

For everyone

API

- Ad hoc analysis & single data-type analyses
- 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
- Direct integration with BI tools e.g. Tableau, Qlik, PowerBI
- Join with private & public data

For data & analytics teams + dashboards for everyone

The main argument

THE CHALLENGE
What is our motivation?

- Cloud databases are powerful, but can be hard to operate
- Performing large scientometric analysis at scale requires a lot of technical knowledge - eg. building a network-analysis

THE OPPORTUNITY
What action are we taking?

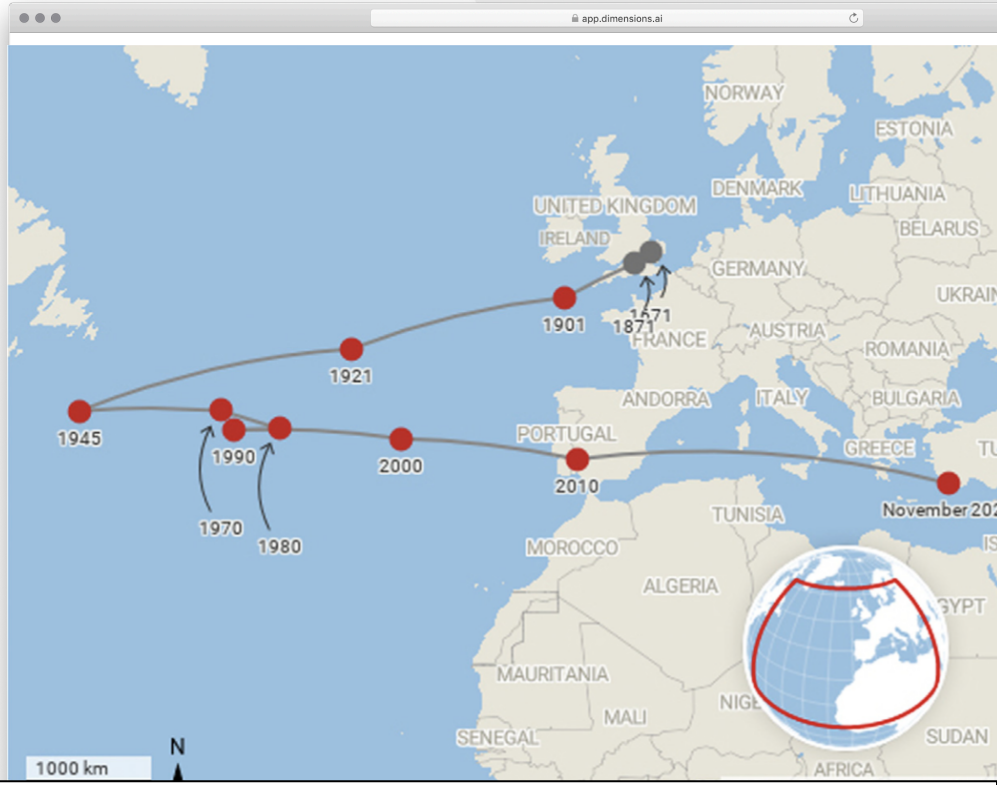
- Can we automate parts of these tasks?
- Can we empower more researchers who have domain knowledge (but lack time / technical skills) ?

THE RESULT
What changes can we expect to see?

- Approach for generating scientific network visualizations from simpler SQL queries
- Uses GBQ + Python + VosViewer, but core ideas applicable elsewhere

Cloud-based scientometrics databases 1/2 performance

- **How they work**
  - Cost-efficient storage
  - Distribute access control
  - Perform complex calculations with an on-demand infrastructure
- **Case study: “centre of mass of research”**
  - authors affiliations locations weighted by citations to outputs



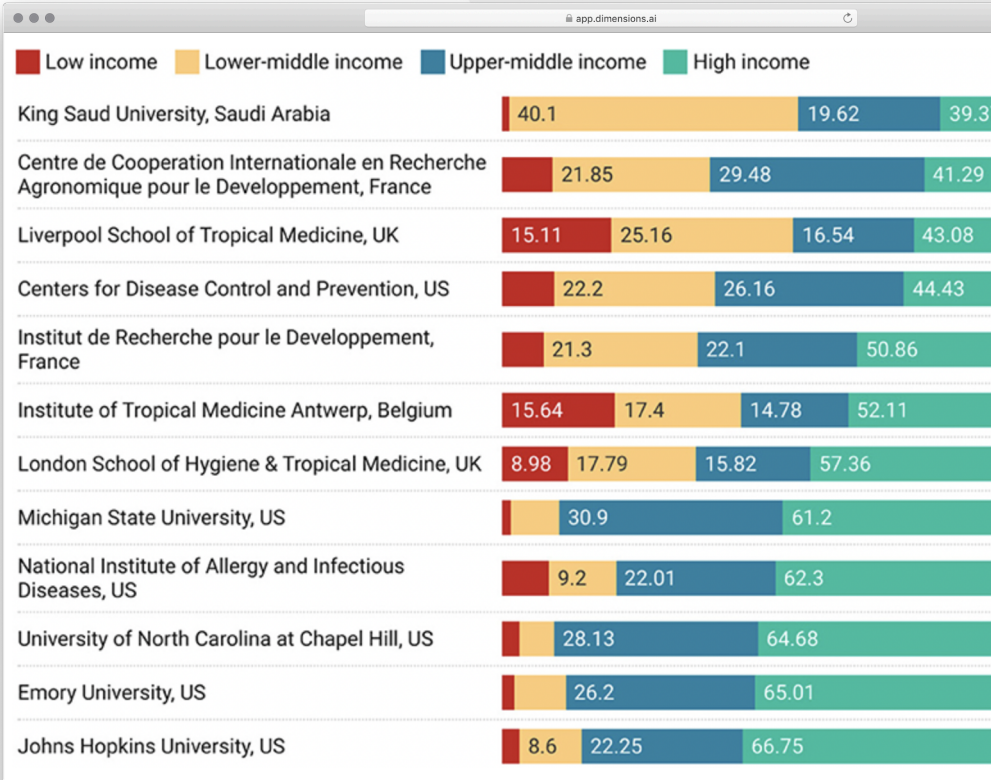
Hook, D. W., Porter, S. J., (2021). Scaling Scientometrics: Dimensions on Google BigQuery as an Infrastructure for Large-Scale Analysis. *Front. Res. Metr. Anal.* 6.

Dimensions

Google BigQuery

Cloud-based scientometrics databases 2/2 linked data

- How they work
  - Cost-efficient storage
  - Distribute access control
  - Perform complex calculations with an on-demand infrastructure
- Case study: “centre of mass of research”
  - authors affiliations locations weighted by citations to outputs
- **Case study: “connecting scientometrics”**
  - GRID + World Bank Data
  - High/Low-income-country research organizations collaborations



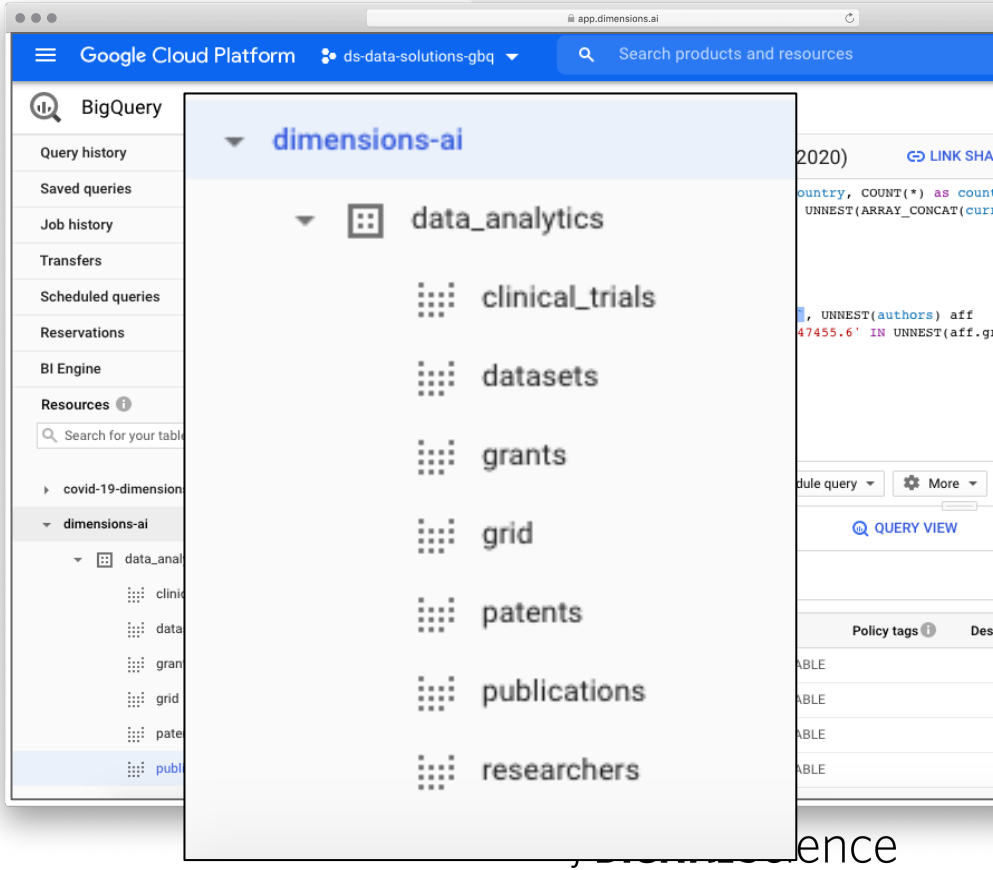
Porter, Hook, D. W., Porter, S. J., (2022). Connecting Scientometrics: Dimensions as a Route to Broadening Context for Analyses. *Front. Res. Metr. Anal.*

Dimensions

Google BigQuery

Technical challenges

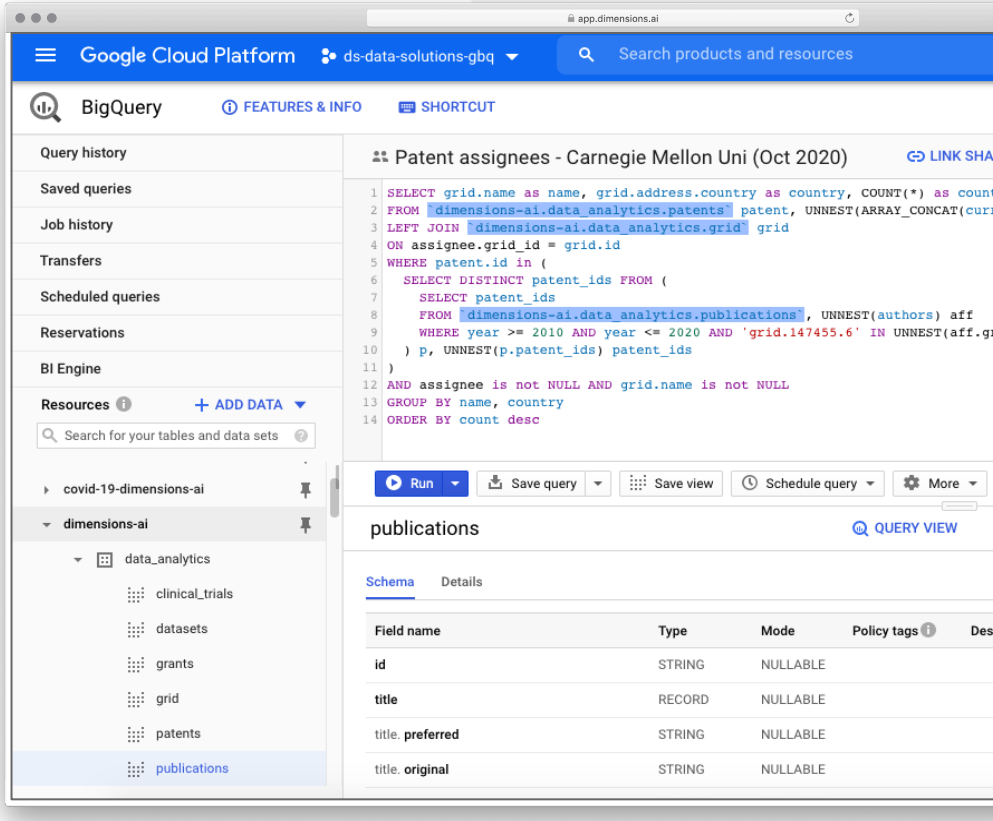
1. **DB Structure can be complex**
2. SQL Query can be complex
3. Cloud systems have their own peculiarities
   EG nested fields



Dimensions

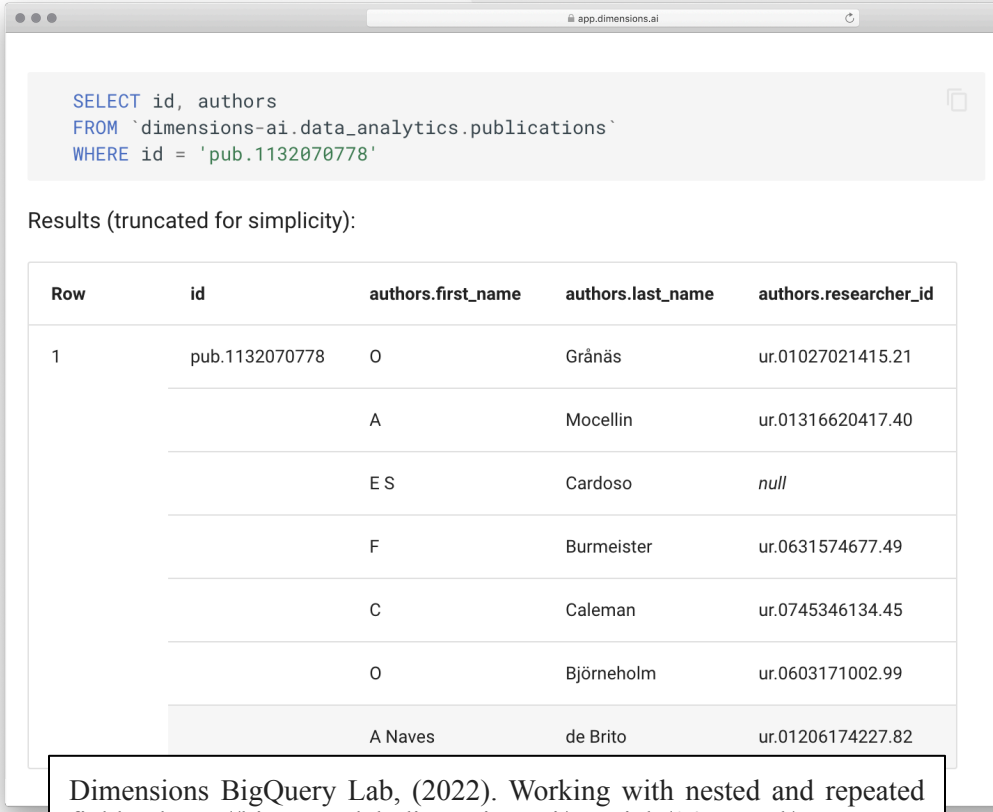
Technical challenges

1. DB Structure can be complex
2. **SQL Query can be complex**
3. Cloud systems have their own peculiarities
   EG nested fields



Technical challenges

1. SQL Query can be complex
2. DB Structure can be complex
3. **Cloud systems have their own peculiarities EG nested fields**

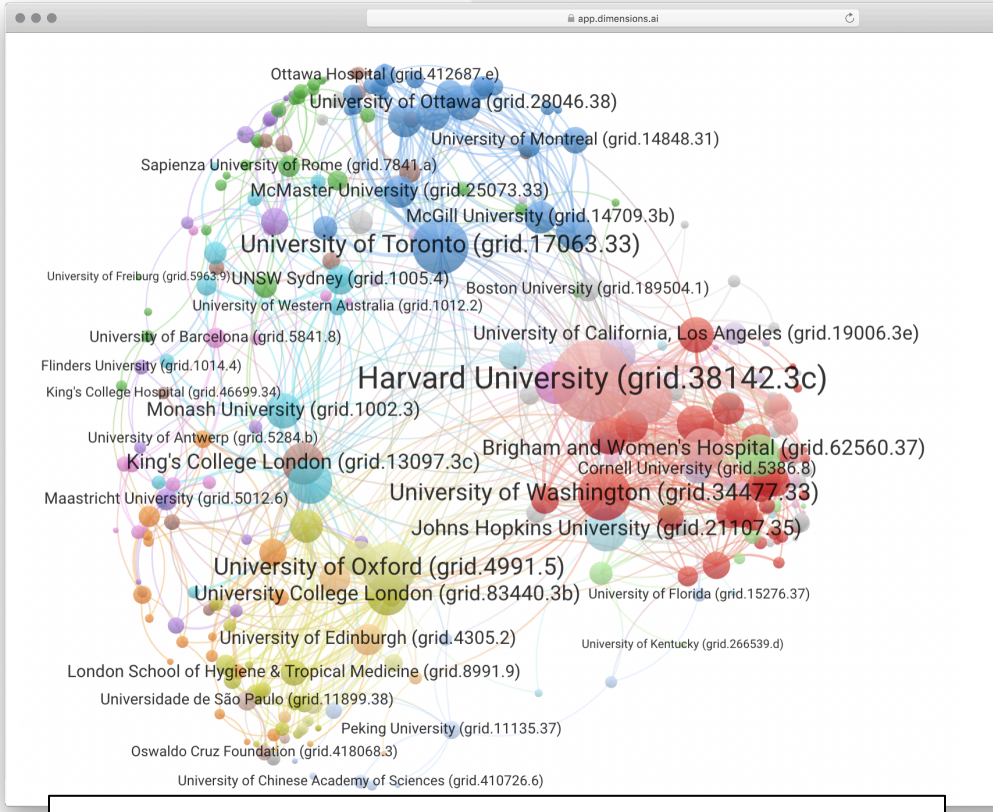


Dimensions BigQuery Lab, (2022). Working with nested and repeated fields - https://bigquery-lab.dimensions.ai/tutorials/04-nested/

Dimensions

Our Prototype: main ideas

- **Goal:** simplified workflow for GBQ-based analytics
- **App** that allows users to generate network visualizations of a scientific *topic*
  - Organization collaborations
  - Concepts co-occurrence

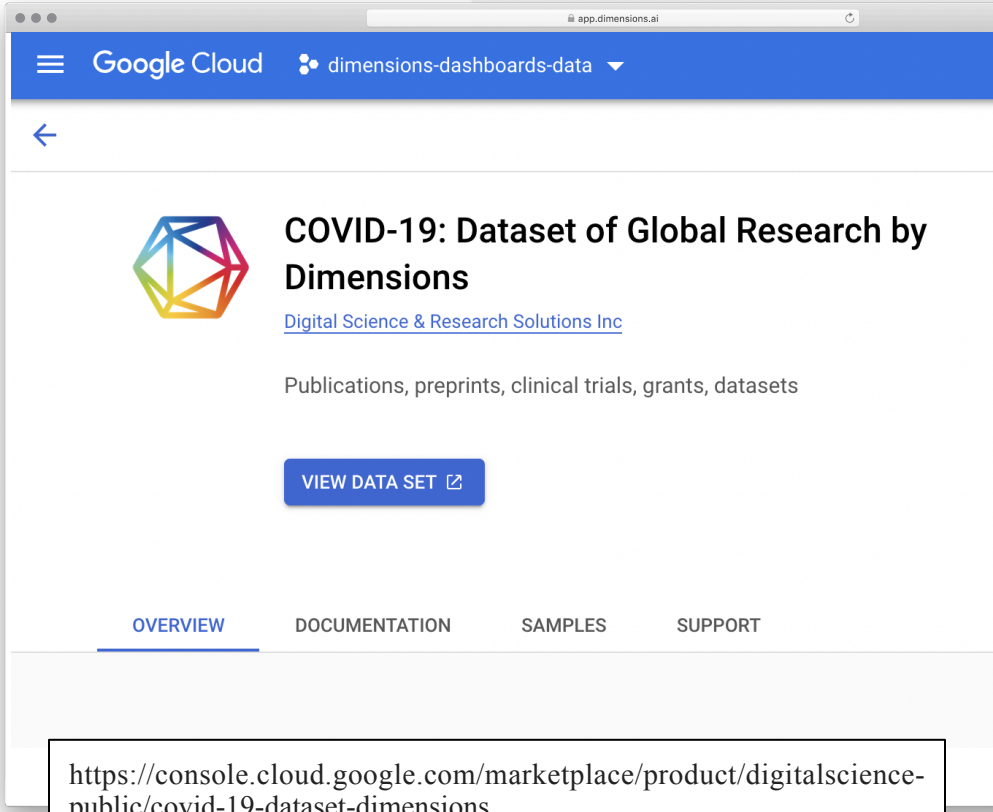


https://digital-science.github.io/dimensions-network-gen

Dimensions

Our Prototype: main ideas

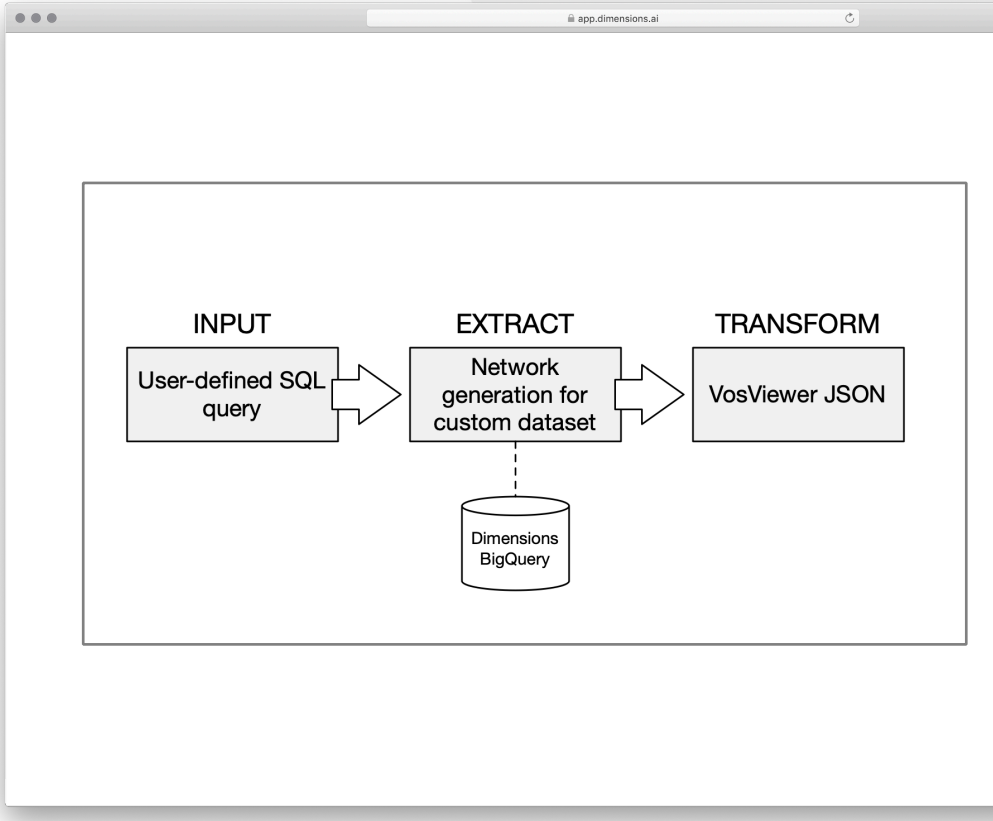
- Goal: simplified workflow for GBQ-based analytics
- App that allows users to generate network visualizations of a scientific *topic*
  - Organization collaborations
  - Concepts co-occurrence
- **Data**: Dimensions COVID-19 on GBQ (public domain)
  - Publications: 1,031,972
  - Clinical Trials: 14,723
  - Grants: 16,703
  - Patents: 41,473
  - Datasets: 32,784
  - Organizations: 36,670



https://console.cloud.google.com/marketplace/product/digitalscience-public/covid-19-dataset-dimensions

Implementation: three steps

1. SQL input
2. Network generation
3. JSON / dataviz output

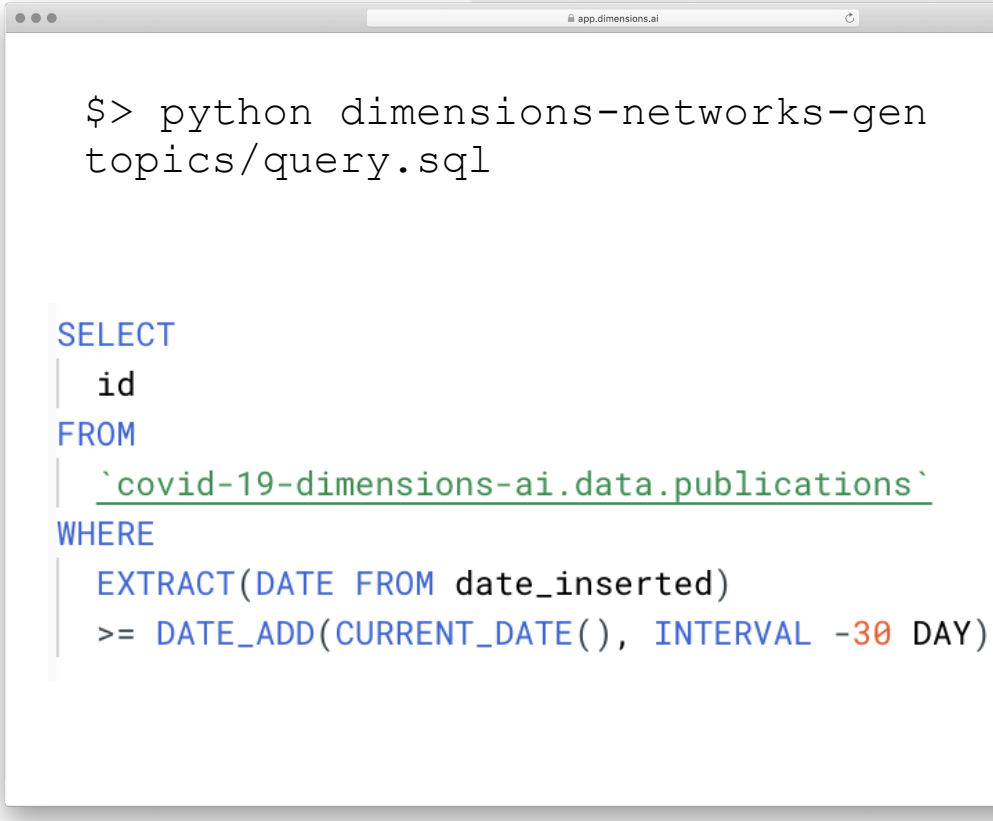


Implementation **1/3**: SQL input

1. **SQL input**
2. Network generation
3. Dataviz output (from JSON)

User define a ‘document set’ using a simple SQL query returning IDs

Document IDs can be used later as part of a more complex query.



Implementation **2/3**

1. SQL input
2. **Network generation**
3. Dataviz output (from JSON)

Automatically, the user query gets embedded in more complex ‘network’ query

- **Concept co-occurrence** network
  - how many publications are shared between these concepts
- **Organisation network**
  - how many publications are shared between these organisations.



```
WITH subset AS (
    {user-topic-query}
),
top_nodes AS (
    SELECT orgid, COUNT(p.id) AS pubs
    FROM `{DIMENSIONS_DATASET}.publications` p
    CROSS JOIN UNNEST(p.research_orgs) orgid
    WHERE id IN (SELECT id FROM subset)
    GROUP BY 1 ORDER BY 2 DESC
    LIMIT @max_nodes
),
links AS (
    SELECT
        CONCAT(g1.name, ' (', org1_id, ')') AS org1
        ,CONCAT(g2.name, ' (', org2_id, ')') AS org2
        ,COUNT(DISTINCT p.id) AS collabs
    FROM `{DIMENSIONS_DATASET}.publications` p
    CROSS JOIN UNNEST(p.research_orgs) org1_id
    CROSS JOIN UNNEST(p.research_orgs) org2_id
    INNER JOIN `{DIMENSIONS_DATASET}.grid` g1
    ON org1_id=g1.id
    INNER JOIN `{DIMENSIONS_DATASET}.grid` g2
    ON org2_id=g2.id
    WHERE
        p.id IN (SELECT id FROM subset)
        AND org1_id > org2_id -- to prevent dupes
        AND org1_id IN (SELECT orgid FROM top_nodes)
        AND org2_id IN (SELECT orgid FROM top_nodes)
    GROUP BY 1,2
)
SELECT *
FROM links
WHERE collabs >= @min_edge_weight
```

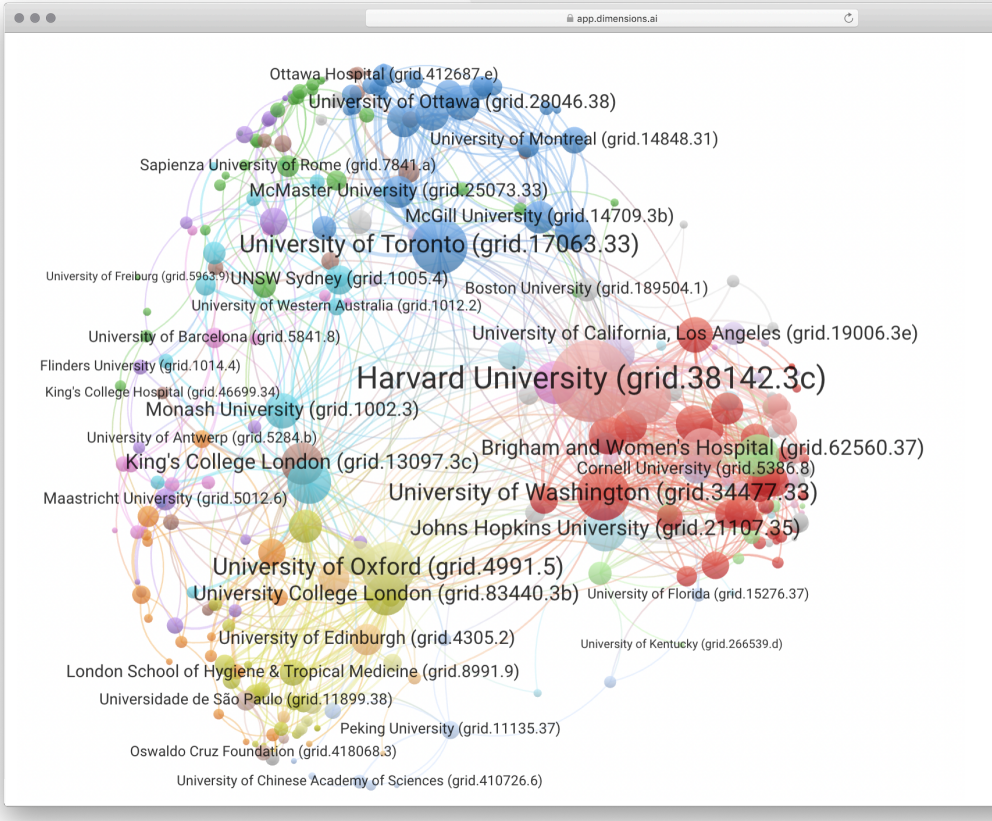
Dimensions

Implementation **3/3**: Data Visualization

1. SQL input
2. Network generation
3. **Dataviz output (from JSON)**

Network data processed so to be compatible with **VosViewer** JSON format.

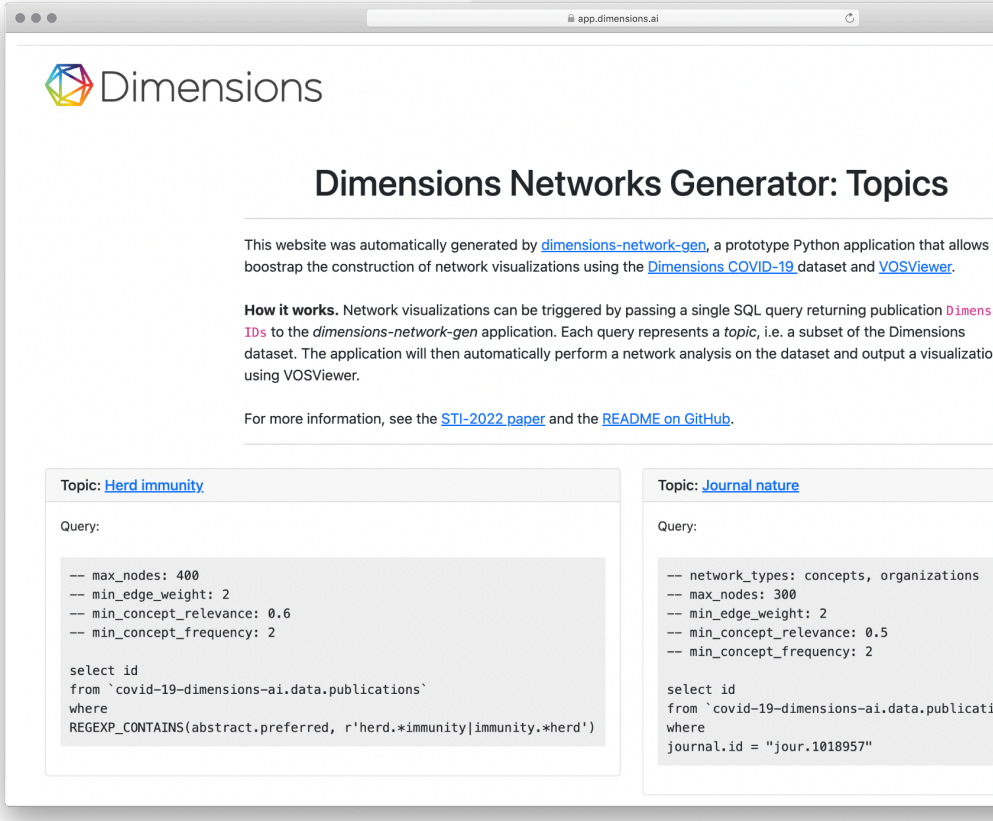
Static HTML pages generated.



Results

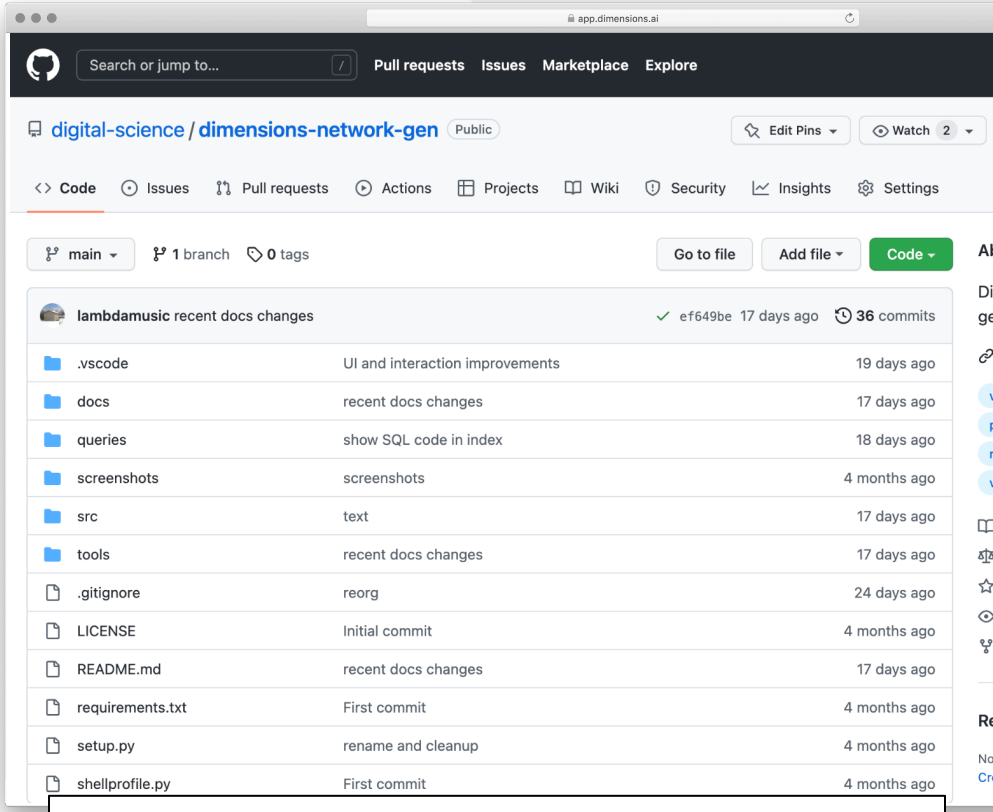
Try it out at:

https://digital-science.github.io/dimensions-network-gen



Wrapping up

- Introducing an **open source** prototype using Python and VOSViewer
- Simplifies the process of generating **scientific networks**
- Uses Dimensions' freely available COVID-19 dataset on **Google BigQuery**
- Shows challenges and proposes a solution for cloud-based analytics generation
- **Audience**: analysts, researchers and anyone who deals with scientometrics data processing



https://digital-science.github.io/dimensions-network-gen

Dimensions

Next steps

- Adding more **network** generation queries
- Adding more output **visualizations**
- Supporting **Jupyter** & Colab

