frontiers
in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 14 April 2021
doi: 10.3389/frma.2021.656233

Scaling Scientometrics: Dimensions on Google BigQuery as an Infrastructure for Large-Scale Analysis

Daniel W. Hook[1,2,3]* and *Simon J. Porter*[1]

[1]Digital Science, London, United Kingdom, [2]Department of Physics, Washington University in St Louis, St Louis, MO, United States, [3]Centre for Complexity Science, Imperial College London, London, United Kingdom

Cloud computing has the capacity to transform many parts of the research ecosystem, from particular research areas to overall strategic decision making and policy. Scientometrics sits at the boundary between research and the decision-making, policy-making, and evaluation processes that underpin research. One of the biggest challenges in research policy and strategy is having access to data in a way that allows for analysis that can respond in an iterative way to inform decisions. Many decisions are based on "global" measures such as benchmark metrics that are hard to source and hence are often nonspecific or outdated. The use of cloud technologies may be promising in addressing this area of providing data for research strategy and policy decisions. A novel visualisation technique is introduced and used as a means to explore the potential for scaling scientometrics by democratising both access to data and compute capacity using the cloud.

Keywords: research cartography, cloud technology, Dimensions, Google BigQuery, data democratisation, centre of mass, unique identifiers, research policy

OPEN ACCESS

Edited by:
Yi Zhang,
University of Technology Sydney,
Australia

Reviewed by:
Dilek Cetindamar Kozanoglu,
University of Technology Sydney,
Australia
Xiaojing Cai,
Zhejiang University, China

****Correspondence:***
Daniel W. Hook
daniel@digital-science.com

Specialty section:
This article was submitted to Research Policy and Strategic Management, a section of the journal Frontiers in Research Metrics and Analytics

Received: *20 January 2021*
Accepted: *10 February 2021*
Published: *14 April 2021*

Citation:
Hook DW and Porter SJ (2021) Scaling Scientometrics: Dimensions on Google BigQuery as an Infrastructure for Large-Scale Analysis. Front. Res. Metr. Anal. 6:656233. doi: 10.3389/frma.2021.656233

1 INTRODUCTION

In recent years cloud technologies have become used more extensively in research. The combination of cost-efficient storage and on-demand compute capability have lowered barriers for many who are either not technically savvy or who lack the financial resources to create and maintain large-scale real-world computer infrastructure. In the academic discplines of bibliometrics and scientometrics, and in the related practical fields of research management, strategy and policy, the use of cloud-based tools are still naiscent. On one hand, data volumes are relatively small (at least compared with familiar big data fields such as particle physics) while on the other, the costs and complexity of arranging access to bibliometric data sources, processing raw data and maintaining analysis-ready datasets have been prohibitive for all but the best funded researchers, analysts and policymakers.

We argue that cloud technologies applied in the context of scientometrics do not only have the capacity to democratise access to data but also to democratise access to analysis. Here we define "analysis" to be the combination of data access together with the capacity to calculate. Data access is often thought to be constrained solely by licence agreements, but is also characterised by technical limitations. Recent progress has been made in improving access to research metadata (Waltman, 2020). Yet, data licence agreements typically do not make arrangements for the delivery of an often-updated analysis-ready database, but rather give access either to a raw flat-file data that needs to be processed, structured and mounted into a database format with regular updates that must be applied to keep the data relevant, or access to an API, which must go through a similar process to create an analysis-ready database. Beyond this logical data structuring activity, there has also historically been the need for physical hardware that effectively defines the computational capacity of the user. Cloud technologies have the capacity to remove both of these constraints by providing an analysis-ready database and computational capacity on a per-use basis.

Few research areas yet take the approach of providing a cloud-based central store of research data for researchers to query, manipulate, and compute with to support their investigations. However, this type of approach can be seen in the conception of "computable data" introduced by Wolfram (2010) as a result of the development of Wolfram Alpha.

In this article we seek to highlight the types of analysis that can be carried out if data is made accessible in the Cloud, as described above, as well as the implications for community ownership of research benchmarks, and the opportunity to place analytical capabilities with a far broader range of stakeholders.

To begin, we provide a working definition of accessibility and use Dimensions on Google Big Query to explore a simple example related to the field of "knowledge cartography," which was introduced and explored extensively by Börner et al. (2003), Boyack et al. (2005, 2007), Borner (2010), Börner et al. (2012), Börner (2015). We use this example as it has great narrative power and makes global use of a dataset. (Here, by global, we mean that to complete an analysis, every record in the dataset maybe required to contribute toward the result—a good example of a global calculation is a field-weighted citation normalisation, since this requires the citation counts of every publication in a set for a defined time period.)

This example brings together the use of a structured, analysis-ready dataset hosted on the Cloud, with unique identifiers to connect from metadata records to spatial information with on-demand computation to provide a visualisation that can readily be updated, iterated and provided regularly to stakeholders in a maintainable manner. We believe that the analysis presented here is entirely novel in a bibliometric or scientometric context. It is remarkable that results of this type have not been presented by other researchers, but we take this to be a hallmark of the limitations of prior computational approaches.

1.1 Defining Data Accessibility

The viability and trustworthiness of bibliometric datasources has been a matter of significant attention in the bibilometrics community over recent years (López-Illescas et al., 2009; García-Pérez, 2010; Mongeon and Paul-Hus, 2016; Bornmann, 2018; Herzog and Lunn, 2018; Martín-Martín et al., 2018; van Eck and Waltman, 2019; Huang et al., 2020). The emergence of new datasources has led to significant analytical efforts to understand the strengths and weaknesses of different approaches to collecting and indexing content (Powell and Peterson, 2017; Thelwall, 2018; Martín-Martín et al., 2020; Visser et al., 2021). The primary focuses of these works are in the assessment of coverage (completeness of journal/subject coverage, and accuracy and completeness of the citation network) together with technical issues around stable construction of field normalisations and other benchmarking details. Both of these areas are foundational in whether a database can be used in bibliometric and scientometric anaylsis, and whether it is appropriate to use these data in evaluative contexts. More recently, there has been innovative work which extends this standard approach to assess coverage in a different manner to examine suitability of datasets for "bibliometric archeology" Bornmann et al. (2020).

For the purposes of this paper, we characterise the majority of existing comparative analyses as being focusing on one or more of five key data facets:

- **Coverage:** The extent to which a body of metadata covers the class of objects and the relationships between objects that it sets out to catalogue. The question of coverage needs to take account of editorial decisions (e.g., inclusion of content based on quality criteria) or other limitations based on geography, timeframe, subject classification, or nature of output (e.g., books, preprints/working papers, datasets). In many systems an editorial choice is implicit in the nature of the repository (e.g., arXiv.org deals with preprints), or it may need to be made explicit as a complex set of criteria may be in action. An often overlooked part of coverage is the completeness of the linkage between objects in the database. In the case of a single object database (most commonly publication databases in the case of bibliometrics) this will typically correspond to citations between publications. However, in multiobject systems, this can be a more complex setting as there are potentially many different types of links in the graph to quantity and assess (for example, links from publications to people, publications to institutions (via people), grants to people, grants to institutions, grants to funders, grants to papers, and so on).
- **Structure:** the technical format and field structure of the metadata; this includes consideration of any relevant metadata standards that may be relevant as well as consideration of completeness of fields in the data. In particular, the extent to which appropriate data is present in any particular record and the extend to which PIDs are present in the data. An example of this would be the inclusion of funder acknowledgement data in bibliographic records: there is the need for consideration of a standardised structure to represent the information, as well as standarisation of identifiers for both funders and the grants that they award. Beyond these technical considerations we would include the extent to which this information is avialable and whether there are cultural issues in the provision of these data (for example, prior to the Open Access movement, some funders have not required acknowledgement).
- **Nature:** the parts of the scholarly record being retained and represented (for example, articles, journals, datasets, preprints, grants, peer reviews, seminars, conference proceedings, conference appearances, altmetric information, chemical molecules, genebank entries, affiliation to research institutions or industrial organisations, guest lectures, outreach work, world premiers, evidence of impact, and so on).
- **Context:** for data, we think of this as provenance. It should consider the source of the data and who has had the opportunity to interact with an change the data but also the

details of enhancement techniques that may have been applied. Specifically, it is valuable to know if the data have been enhanced, extended or "completed" via the use of AI or machine-learning algorithms and a record of the algorithm used.

- **Quality:** the definition that we offer here is not synonymous with coverage (which is sometimes an assumption that is implicit in studies of data: more data equals higher quality). Rather, we prefer to define quality in terms of errors found in the data (for example, misassociation of two objects or misattribution). We also believe that consideration of data homogeneity is important—a highly complete dataset that lacks good structure can be extremely difficult to use. There is also an aspect to quality that goes to robustness of the data, which we define to be the percentage of records need to be corrected in each update to the data. It a data source is highly stable (i.e., few data changes) then this may be a measure of robustness (although this implicitly assumes that the maintainer of the data source is working to improve quality actively). Ultimately, the hallmark of quality, in our opinion, is the extent to which an analysis can be reproduced identically on a dataset that may have been updated.

The first four of these aspects of a dataset define the extent of a "data world" that may be explored and analysed to deliver insight. If we wish to push out the boundaries of this world, then we can do that by improving each of these facets: Extending the coverage of the database, deepening sophistication of the facetting, expand the different types of data that we include for analysis, or by broadening the links between different parts of the data to improve context. Data quality determines one key element of the confidence and trust that we can place in analyses.

It may be argued that more established data sources have sought to optimise coverage, structure and quality of their data. But, newer databases have brought a new focus on nature and context (Hook et al., 2018; Herzog et al., 2020). By expanding the types of data they that cover, or by creating better linkages between those new data types to improve our ability to contextualise data, they improve the variety and subtlty of the insights that the scientometrics community may generate. We have attempted to make our list of facets of data facets comprehensive at a high level, however, we also recognise that this is a large and complex subject about which apparently little has been written. As a result, we view this suggestion as an initial framework to be improved upon and iterated by others. Several features that could be included in a hollistic approach to defining the value that may be derived from analysis go beyond the dataset. Examples include: the affiliation of the analyst, and the robustness of a statistical treatment. We argue that data accessibility is a different type of feature of a dataset that should be considered more actively, especially in the rise of cloud technologies.

Data accessibility is a complex and multifaceted topic. The key facets that we believe to be important in the context of scientometric data and analysis are:

1. Timeliness: the extent to which a user can access sufficiently up-to-date data;
2. Scale: the extent to which it is possible to access the whole dataset for calculational purposes;
3. Equality: the extent to which the resources to process the data and perform a calculation are technologically practical;
4. Licence: the legal terms that define the extent to which the data may be used and published.

The example that we use here does not attempt to illustrate or address all these facets, but rather focuses on Scale and Equality—two facets that we believe to be best addressed by cloud technologies. We have recently explored timeliness of data access in a recent article where we proposed the idea of "Real-time bibliometrics" (Hook et al., 2020). Although data licencing does play a role in the use of cloud technologies, this goes significantly beyond the scope of the current article and should be addressed by those with more expertise in the area.

Specifically, we examine classes of calculation for which data access is required for scale and look at how Cloud technologies can facilitate both scale and equality of access to data. Our example will use Digital Science's Dimensions on BigQuery infrastructure. We note that this paper is specifically designed not to be a comparative study of the accessibility of different data sources, but rather as an opportunity to showcase the types of analysis that can be carried out if technological choices are made that democratise data access.

This paper is organised as follows: In **Section 2** we describe the Dimensions on Google BigQuery technical stack, and the specific queries used for the analysis presented in the following section. In **Section 3** we show the results of several different calculations of the centre of gravity of research production using the method described in **Section 2** and discuss the context of those results. In **Section 4**, we consider the potential of Cloud technologies to meet a broad set of use cases.

2 METHODS

2.1 Technical Infrastructure

Many Cloud technologies are already used across research, especially in technical subjects requiring large-scale computation or storage, or those who engage in large-scale collaborations. Indeed, Cloud technologies are becoming more widespread in research as they prove to be highly cost-effective for some types of research activity. Typical use cases involve storage and transfer of data or obtaining computational power on demand.

For those with structured data, the cloud technologies that allow users to not only store and distribute access to a dataset but also to perform complex calculations with an on-demand infrastructure are now coming of age. Technologies such as Amazon Redshift, Snowflake, and Google BigQuery all have the potential to meet the use cases mentioned above (Zukowski, 2018).

In addition to their technical capabilities, these technologies are opening up new business models through the ability to share secure data in a fine-grained and controlled manner. Any of the technologies mentioned allows a data holder to share data from their cloud database with others on a permissioned basis, opening up access specifically or generally based on many different criteria. From a business model perspective, a critical differentiator (not used in the current example), is that two parties can add their data to the cloud completely securely, one can keep their data private while the other can open their data up on some mix of open access and commercial basis. The second actor's data can then be used by the first actor, on whatever the appropriate contractual terms are, mixing the data with their private data in a completely secure manner. The only requirement is that each dataset should have a sufficient overlap in persistent unique identifiers to allow the datasets to be compatible. Hence, this technology is a strong reason for all stakeholders in the community to adopt and ensure that the data that they expose is well decorated with open identifiers. For large, frequently updated datasets where there is significant overhead in just storing and updating the data, this new way of working compeletely changes the basis of engagement.

From the perspective of the current article, the availabilty of Dimensions data in the Google BigQuery Cloud environment allows users to access and compute directly with the data without having to invest in either building or maintaining local infrastructure. It also allows users to manipulate and calculate with data across the whole Dimensions dataset. The only technical expertise that is required is an ability to program with SQL.

It is easy to see how the calculation explained below could easily be replaced to calculate other metrics and indicators that require access to a "global" dataset. Such calculations include journal metrics such as Journal Impact Factor (Garfield and Sher, 1963), EigenFactor (Bergstrom, 2007), SJR (González-Pereira et al., 2010) or CiteScore (Van Noorden, 2016), as well as the production of journal citation distributions (Larivière et al., 2016), field-based normalisations such as RCR (Hutchins et al., 2016), as well as geographical benchmarks, trend analysis or examples of knowledge cartography, such as the example that we have chosen to explore.

2.2 Calculation

To illustrate how the new technologies described above may be used, we perform a simple global calculation. As noted above, the word "global" here is not intended to refer to a geographical context, but rather implies that each record in the database will potentially contribute to the calculation.

We calculate the centre of mass of global research output year by year. This calculation has several noteworthy features that demonstrate the capabilities that we've discussed earlier. The calculation: 1) involves every publication record in our dataset; 2) makes use of a unique identifier to connect publication outputs to geographical locations (in our case through GRID); 3) makes use of the time-depth of the publications records in the database to give a trend analysis.

Using non-cloud infrastructure to perform this calculation such as a standard relational database hosted on physical infrastructure would make this calculation time consuming and resource intensive. By leveraging cloud infrastructure we can quickly iterate the detail of this calculation to test different hypotheses. For example, we can easily shift from a centre of mass calculation that focuses on publications to one that focuses on awarded grants, patents or policy documents. We can trivially change the weighting factor from an unweighted calculation to a citation weight in the case of publications, grant size in USD for grants, the funded associated with a publication, the altmetric attention associated with a patent and so on. We can also easily restrict our analysis to a specific research topic, country, institution, a specific class of grants, a particular type of funding or a larger-scale policy initiative such as open access. To take this even further, one can imagine even subtler weighting schemes that take the CRediT taxonomy (Allen et al., 2014) into account.

In the examples contained in this paper we focus on publication output and either unweighted or citation-weighted formaulations. The core of the centre of mass calculation is a simple weighted average of spatial positions that all students of classical mechanics meet early in their studies—it is equivalently known as a centre of gravity calculation or centroid.

In our example, each "mass" is an affiliated research institution and the location of that mass is the geographical location of the principle campus as recorded in GRID. For each individual paper, there is a centre of mass the position of which is proportional to the contribution of the affiliations of the researchers who have contributed to the paper. For exmaple, if a paper were to be entirely written from researchers at a single institution then the centre of mass for the paper in our calculation would be the location of the principle campus in GRID. If a paper were to be written by two co-authors, one at the University of Cambridge and the other at the University of Oxford, then the centre of mass would be computed to be midway between the two Senate House buildings of the two institutions. To find the centre of mass of global output in any year, we average the spatial location of all the papers produced in that year. We can think of this position as the "average centre of global research production" or the "centre of mass/gravity of global research output".

We also introduce a citation-weighted version of this calculation which may be interpreted as a measure of centrality of global research attention to research output.

Formally, we define the centre of mass of a set of research objects to be the spatial average (or centroid) of the affiliations of the co-creators of the output. On a paper with n co-authors, each co-author is associated with $1/n$ of the paper. If a given co-author is affiliated with m institutions, then each institution will have a weight of $1/m$ of that co-author's part of the paper, and $1/nm$ of the overall paper. Thus, each author-institution pairing has a weight a_{nm} where

$$\sum_n \sum_m a_{nm} = 1. \tag{1}$$

We do not need to explicitly sum over authors to get the overall contribution of a specific institutions nor do we need to worry about repetition of institutions since, in our calculation, we reduce an institution to the longitude and latitude of its principal campus. Hence, there is a natural accumulation of weight to a geographical location.

This reduction to longitude and latitude is made possible through the use of GRID. The longitude and latitude of research institutions is not held natively within the Dimensions dataset. However, each institution in Dimensions is associated with a persistent unique identifier that allows us to connect to other resources. In the case of Dimensions the institution identifier is the GRID identifier. GRID not only includes some helpful data about institutions such as the longitude and latitude that we use here but also acts as a gateway to resources such as ROR (the Research Organisation Registry) that will in turn facilitate access to other pieces of information.

This means that we can simply calculate the average longitude, $\overline{long}$ and latitude $\overline{lat}$ of a single research output using:

$$\overline{lat} = \frac{1}{T}\sum_i\sum_j lat_{ij}; \qquad \overline{long} = \frac{1}{T}\sum_i\sum_j long_{ij}, \tag{2}$$

where T is the total number of publications.

We can then extend this to a group of outputs by introducing an index, k, that ranges over each output in the relevant set to create the average longitude $\overline{Long}$ and average latitude $\overline{Lat}$ of the whole set:

$$\overline{Lat} = \sum_k\frac{1}{T_k}\sum_i\sum_j lat^k_{ij}; \qquad \overline{Long} = \sum_k\frac{1}{T_k}\sum_i\sum_j long^k_{ij}, \tag{3}$$

where T_k is the total number of institutional affiliations on the kth paper in the average.

Longitude and latitude are defined as angles on the surface of a sphere with longitude in the range $[-90, 90]$ and latitude in the range $[-180, 180]$. The construstion in **Eq. 3** guarantees that the final results of these calculations are also in these ranges.

Further weighting factors can also be added to the calculation to highlight issues of particular interest. For example, if we were to consider an example using research publications and we wished to calculate not just the centre of the output rate but rather the centre of the combination of output weighted by the attention given to that output, then we might introduce a weighting by the number of citations received by each paper.

In that case **Eq. 3** would need to be updated and the form for the centroid would be:

$$\overline{Lat} = \frac{1}{C}\sum_k\frac{C_k}{T_k}\sum_i\sum_j lat^k_{ij}; \qquad \overline{Long} = \frac{1}{C}\sum_k\frac{C_k}{T_k}\sum_i\sum_j long^k_{ij}, \tag{4}$$

where C_k is the number of citations of kth paper and C is the sum of all citations across papers in the set.

Likewise, if we were interested in the level of non-scholarly attention we might replace citations by some relevant altmetric data.

The code snippet below is the implementation of **Eq. 4** using Google BigQuery's implementation of SQL on the Dimensions dataset. In addition to the calculation explained above, the code below takes into account cases where creators may miss an affiliation by ensuring that the normalisation is consistent in the case of null data.

One assumption that may not at first appear obvious with the weighted approaches used here is that the sum of all citations in time has been used. As a result, papers in 1671 have had 350 years to garner citations whereas more recent publications have had much less time. Of course, the average in each case is performed on a homogeneous basis (i.e., only publications of the same year are averaged together), however, this does introduce an implicit bias in the analysis in that a citation bias may have a comtemporary skew. A further analysis could be performed that only considered the citations in an n-year window following the date of publication of the paper. Of course, introducing such a parameter also makes a value judgement about the lifetime of a piece of research.

In **Section 3** we use this method to showcase three analyses: 1) a standard unweighted calculation of the centre of mass of research output from 1671 to present day; 2) a calculation of the centre of mass of research weighted by citation attention over the same time period; 3) a calculation of the citation-weighted centre of mass of research based just on data from the freely available COVID-19 dataset that is available on the Google BigQuery environment.

2.3 Data Specifics

The details of the high-level data schema in Dimensions, including information about coverage and the treatment of unique identifiers is described in several recent publications, for example, Hook et al. (2018, 2020).

Once the data are produced from a script such as the one above they were downloaded from the interface and are initially analysed in Mathematica. The graphics shown in **Section 3** are produced using Datawrapper.de.

At the Mathematica analysis stage, we plotted every year of data from the system. However, this gave an unsatisfactory picture as the data are quite messy. In the early years of the dataset (approximately from 1671 to 1850) the number of publications with a GRID-listed institutions number in the single digits. A confluence of reasons contribute to this picture: 1) the low number of overall publications; 2) the low level of stated academic affiliations of authors in early work; 3) affiliations to institutions that are not part of GRID. **Figure 1** shows the number of publications with at least one recognisable (GRID-mapped) affiliation in each year in the Dimensions dataset.

From 1900, the data begins to settle as it begins to be appropriate to treat it statistically in the context of a statistical calculation such as the one outlined in **Section 2.2**. Between 1900 and 1970, the year-on-year variability of the data decreases, and from the 1970s the data describes a fairly consistent path with few significant derivations. As such, we have denoted points in the figures in grey where they contain “less robust” data and in red when the data are “more robust”.

In the final analysis presented, we focus on the COVID-19 dataset and perform a month-by-month analysis. In this situation, we are again in the law of relatively small

Listing 1. Listing to produce a citation-weighted centre of mass year-by-year using SQL on Google BigQuery with Dimensions data.

```sql
WITH pubs_reweighted AS (SELECT p.id,
         p.year,
         a.first_name,
         a.last_name,
         a.initials,
         /* count the distinct number of organisations per author */
         COUNT(distinct g.id) num_orgs,
         /* list of all the GRIDs per author */
         ARRAY_AGG(grid_id) grids,
         /* count the number of authors on the paper that have
   affiliations in GRID */
         COUNT(p.id) over(partition by p.id) authors
FROM
     `dimensions-ai.data_analytics.publications` p
      INNER JOIN unnest(authors) a
      INNER JOIN unnest(a.affiliations_address) aff
      INNER JOIN `dimensions-ai.data_analytics.grid` g
         ON g.id = aff.grid_id
GROUP BY
        p.id,
        p.year,
        g.name,
        a.first_name,
        a.last_name,
        a.initials
),
/* get the location for each GRID. Each row that is being summed here
    represents a single author.  If they have more than one
   affiliation then the contribution of the author is split equally.
   */
pub_center_mass AS
    (SELECT pr.id,
          SUM((g.address.latitude/pr.num_orgs)/pr.authors)  latitude,
          SUM((g.address.longitude/pr.num_orgs)/pr.authors)
   longitude
    FROM pubs_reweighted pr,
        UNNEST(grids) grid_id
        INNER JOIN `dimensions-ai.data_analytics.grid` g
             on g.id = grid_id
    GROUP BY pr.id)
SELECT p.year,
/* sum the centre mass for all publications / the number of
   publications; replacing p.metrics.times_cited in the respective to
   an explicit value of "1" recovers a weighting-free from the
   calculation */
      (sum(cm1.latitude * p.metrics.times_cited)) /   (sum(p.metrics.
   times_cited)) latitude,
      (sum(cm1.longitude *  p.metrics.times_cited)) / (sum(p.metrics.
   times_cited)) longitude
  FROM pub_center_mass cm1
   INNER JOIN `dimensions-ai.data_analytics.publications` p
      ON p.id = cm1.id
  GROUP BY p.year
  HAVING sum(p.metrics.times_cited) > 0
  ORDER BY year
```

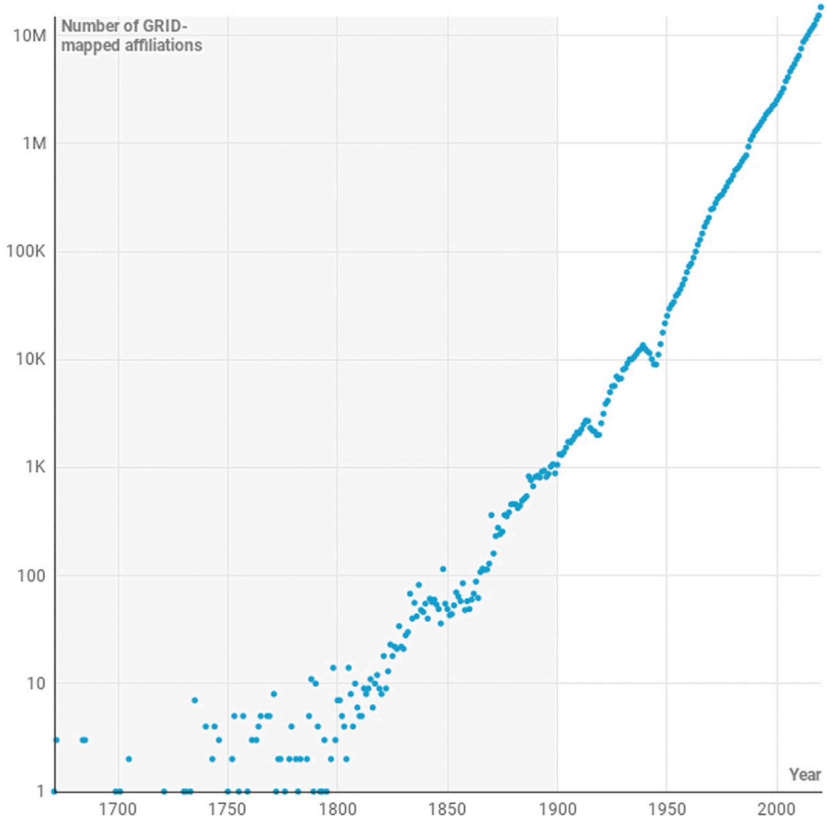

FIGURE 1 | Logarithmic-scaled plot of the number of GRID-mapped institutions associated with papers in the dimensions database by year from 1671 to 2020. The two notable dips in the data in the first half of the 20th Century co-incide with the two world wars. The grey background highlights the region between 1671 and 1990 in which the number of contributing records is taken to be too small to give a stable basis for statistical analysis.

TABLE 1 | Number of COVID-19 research publications including journal articles, preprints, monographs and book chapters by month during 2020 in the dimensions database.

Month	Number of publications
January	289
February	751
March	3,140
April	9,999
May	15,502
June	15,377
July	16,706
August	15,645
September	16,191
October	18,304
November	15,170
December	15,153

numbers where we have to be careful about statistical effects. However, the COVID-19 dataset has grown quickly during 2020 with a few hundred papers in January growing to several thousand papers per month in November (see **Table 1**).

3 RESULTS

From a historical perspective, the calculation of a variety of difference centres of mass can be revealing. At the least, they may confirm accepted doctrine, but in the best situation they can reveal features that allow us to quantify and understand how aspects of our society are developing in a very relatable manner.

Bibliometric analyses such as those presented here have previously been difficult to undertake due to the challenges of arranging data access, having the capacity to process data into an appropriate format, having the computation capacity to perform calculations and having a good reason to do put effort into generating this kind of output. With the arrival of cloud-based technologies the technical challenges are removed. A mere 40 lines of code, with a runtime of significantly less than 1 min, is required to produce the data that underlies the analysis presented here based on the Dimensions dataset.

By comparison, such plots are relatively more common in other areas of research, such as economics or geography. The recent work of Dobbs et al. (2012) examined the movement of the centre of mass economic activity in the world from 1CE to the present day, showing that the economic centre of mass two millennia ago lays on a line between Rome and China. During

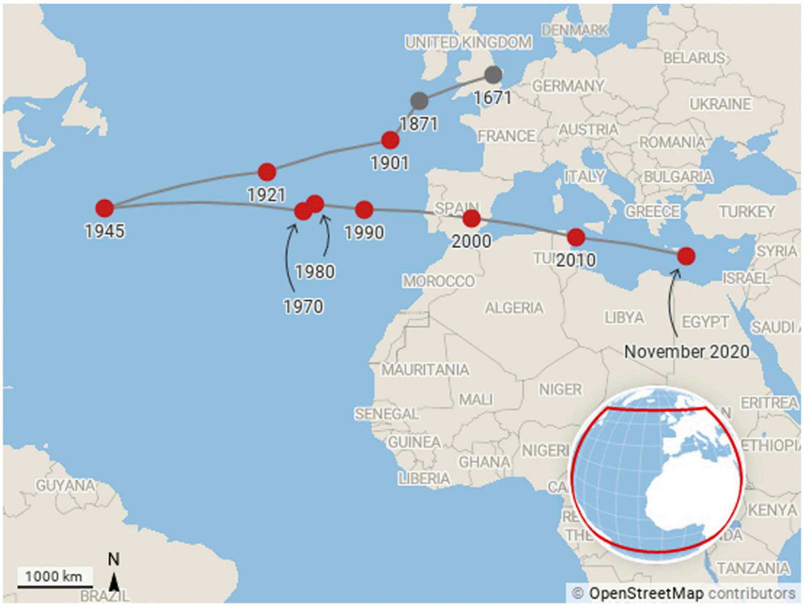

FIGURE 2 | Motion of the centre of mass of research production from 1671 to present day. The centre of mass calculation is unweighted by citations or other measures and is based solely on the outputs of papers by institutions that appear in the GRID database.

this period, the Silk Roads was the commercial axis between the two largest empires in the world: the Roman Empire and the Eastern Han Empire. It is unsurprising that the economic centre of gravity is closely linked to these ancient centres of commerce. The centre of mass was solidly grounded in the same region until at least 1500. However, following the englightenment in the 18th century, science and technology began to transform the economies of Europe and for a century from 1820 to 1913 the centre of mass of the world's economy moved rapidly West and North as the Industrial revolution transformed first the United Kingdom and then the wider Western world. Interestingly, in the McKinsey analysis, despite America's increasing world status and riches during the 20th Century, the centre of economic mass never quite left the Eurasian continent, reaching its zenith in 1950, just over Iceland, before beginning its journey Eastward and, again, Southward as first Europe emerged from war, Japan developed economically during the 1980s and finally China reached economic preeminence as we entered the Asian Century (Rachman, 2017).

Most in academia agree that formal research publication dates from 1665 with the first issue of the Philosophical Transaction of the Royal Society (Hurst, 2010). Hence, the data that we have around research activity only spans a few hundred years and does not share the time-depth available in the work of Dobbs et al. (2012). As a result, from a data perspective, we miss much of the detail around the development of older societies such as those in Egypt and China. Anaecdotally, it is particularly interesting that the Chinese did not develop a research community with the associated communication structure despite significant technologies through the Ming and Qing periods. Indeed, many of the principles that led to the Enlightment in Europe had parallels in Qing China and there is even evidence in European writings that they were aware of enlightnment-style developments in China (Wood, 2020). Yet, this does not appear to have resulted in the emergence of formal research publication culture. Miodownik offers a material scientist's view in Miodownik (2014) on the relative rate of development of Chinese science - it may be that the development and wide adoption of glass in preference to porcelain is the small change that shaped the development of history for several centuries.

The scholarly communications community has associated today's digital infrastructure (such as persistent unique identifiers) with pre-digital-era publications and this gives us an ability to piece together a much fulller picture than would otherwise be the case. Nevertheless, **Figure 1** makes it clear that data are not sufficient to be treated in a reasonable statistical manner until much more recently. For the purposes of our example, we have chosen to keep the more statistically questionable points on our plot for aesthetic reasons, but have coloured these points in later figures in grey to denote the intrinsic uncertainty and arbitrariness of the choice of the data point.

Figure 2 shows the motion of the unweighted centre of mass of global publication output between 1671 and the present day. The start point of the path is an easy one to calculate since only one publication in that year is associated with a DOI and a GRID-resolved institution. The paper concerned is a Letter that appeared in the Philosophical Transactions of the Royal Society of London. It is written by "Mr. Isaac Newton, Professor of the Mathematicks in the University of Cambridge; containing his new theory about light and colors". The path is highly volatile in the years following 1671 as the number of papers is small (those interested in this detail can review the annual calculation in the supplementary material). However, by 1901, there is are sufficiently many papers with well-identified institutions that the path settles somewhat.

Many of the great academic institutions in the US had been established in the late 18th Century. Through the 19th Century the "Robber Baron" industrialists such as Mellon, Carnegie and Rochefeller had continued the trend of setting up academic institutions and by the 20th Century, these institutions were pulling the centre of mass of research (eratically at first, but then with increasing speed) away from Europe. The First and Second World Wars saw significant disruption in Europe and the wealth that had taken the British Empire a century to accumulate travelled to the US in just four years as Britain underwrote the costs of the First World War between 1914 and 1918. And so, the movement of the centre of mass of research production makes complete sense from 1900 to 1945.

If anything, it is remarkable that 1945, the year that Vannevar Bush wrote his famous Endless Frontier report (Bush, 1945), marks the turning point of the transit of the centre of mass back toward Europe. While the end of disruption in Europe meant that academics could return to their research and publication could begin again, Germany was in ruins and the economy of the United Kingdom was in tatters. Despite the success of United States-based programs such as the Manhattan Project during the war, research focus had yet to come to the fore in United States universities.

Following Bush's report, the National Science Foundation was created and the formal basis for a period of United States-centred scientific pre-eminence was established. In Europe, the reorganisation of research was also under way, the Kaiser Wilhem Institute was renamed to the Max Planck Institute in 1948 and in 1949 the Frauhofer Institute was established. By the 1960s, the Royal Society of Great Britain would coin the term "Brain Drain" to describe the movement of British Scientists from the Old World to the New (Cervantes and Guellec, 2002; Balmer et al., 2009). In the United Kingdom, Wilson's White Heat of Technology of the 1960s (Morris and Noble, 1993) served to help to keep the centre of mass moving torward Europe.

Overall the balance of publication volume remained in Europe's favour from 1945 until 1970, with a slow draft in the centre of mass of publication toward Europe. During the final decade of this period US spending on research as a proportion of its discretionary budget reached an all-time high (House, 2020) with the that, between 1970 and 1980, the centre of mass looked as thought it might turn around and head back toward the US once more. The high level of investments in research had begun to pay off and science was riding high in the public psyche in the US in this period.

Yet, despite the payoff from the space race and the beginning of the computer age, spearheaded by silicon valley in the US, the path of the centre of mass resumed its trajectory toward Europe in the 1980s. The speed of transit of the centre of mass has remained about the same since 1990s, but this conceals a complex set of forces behind this motion: The rise of Japan as an industrial and research power; the emergence of the professionalisation of research in the United Kingdom; the creation of a Europe-wide research strategy embodied in the creation of the European Research Council and centralised strategic funding from the framework program grants and the Horizons 2020 program; and, since 2000, the rise of China as both a major economy and research power. Indeed, in decades to come we are likely to see the centre of mass travel further as China establishes further and India scales up its research economy.

An unweighted calculation shows the clear average centre of production, but it is interesting also to think about different types of weighting. This should be done with care since the interpretation of such weightings is not trivial. **Figure 3** shows a similar picture to **Figure 2**, but this time with each institution's contribution weighted by the fraction of the number of citations associated with the papers written by their affiliated authors. The addition of citation data stablises the path overall, as there is a bias toward the most established research economies. In this figure, the centre of mass continues to be closest to the United States in 1945, but it returns to Europe initially more slowly, and actually turns around, heading back toward the United States in the 1980s, before moving once more toward Europe, moving faster than ever, by 2000.

The speed of movement toward the east has increased significantly over the last 20 years, which is indicative not only of increasing research volumes in China as well as Japan, India, Australia and New Zealand but also the increased citation garnered by those publications.

Additionally, while the range of movement of the centre of mass from east to west is significant, its movement to the south, while being monotonic and more limited in range than the longitudinal motion, is notable by its consistency in the latter half of the 20th Century. The majority of the world's large cities, and hence most abundant research economies, are in the northern hemisphere. Yet, the trend is to the South and tracking this motion is sure to be interesting in the future.

Our third and final narrative is contained in **Figure 4**, which shows the motion of the citation-weighted centre of mass of COVID-19 research on a monthly basis during 2020. The number of publications that contribute to each point on the plot is shown in **Table 1**.

As news of COVID-19 emerged from Wuhan in China during at the beginning of the year, China's researchers quickly turned their attention to studying the disease. The fact that the centre of mass of COVID-19 research in January 2020 is located on the Tibetan plateau (paradoxically, quite near to the centre of mass of global economic output in 1CE as calculated in the McKinsey report that originally inspired this line work in this paper) rather than closer to China's research centres is a clear indication that research was already taking place in the international community. As the year progressed and the virus spread to pandemic migrated West, more and more research organisations in the West turned their attention to COVID-19 research. The shift in the centre of mass of global research production and the speed at which this happened is easy to see from **Figure 4**.

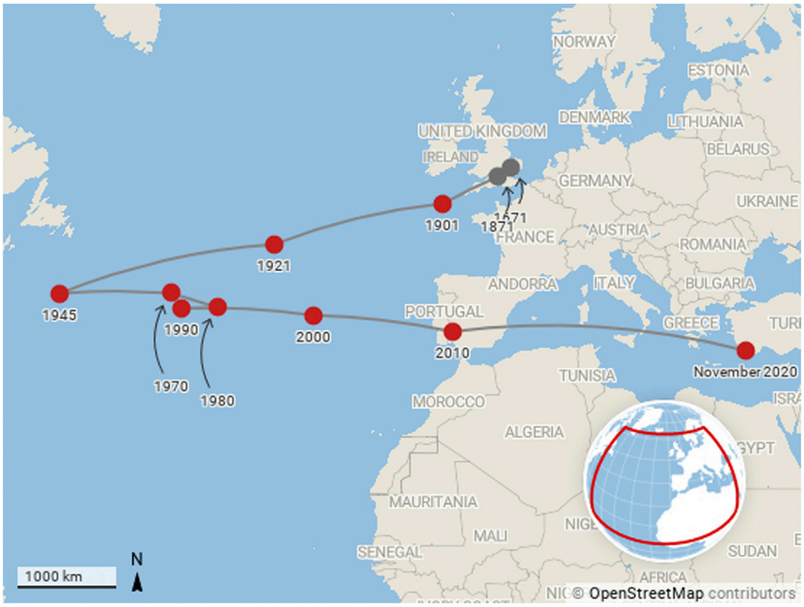

FIGURE 3 | Motion of the centre of mass of research production from 1671 to present day. The centre of mass calculation is weighted by citations to outputs as described by the Code Listing 1 and **Eq. 4**.

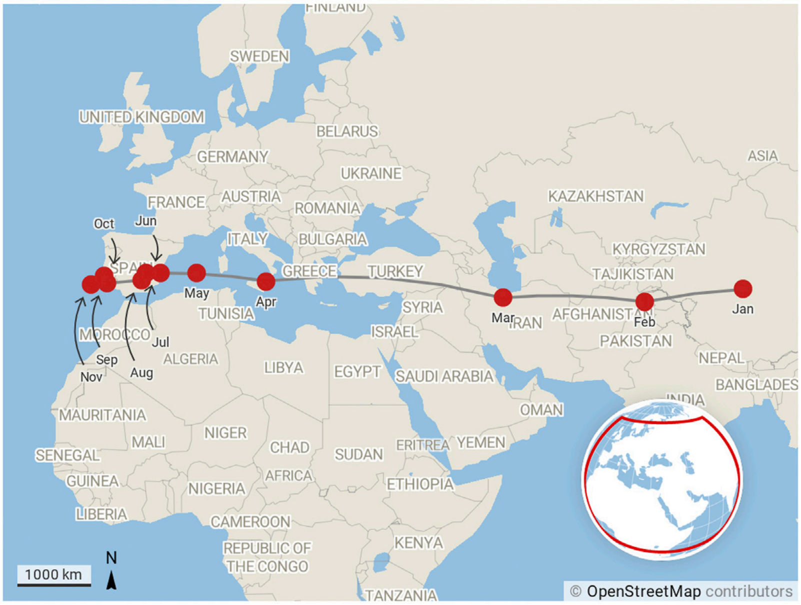

FIGURE 4 | Motion of the centre of mass of research production month by month for COVID-19 publications from January 2020 to November 2020. The centre of mass calculation is weighted by citations as described by **Eq. 4**.

Maps hold a special place in human storytelling and hence are a powerful means by which we can relate to data. The use of such maps does not come without baggage—such visualisations hide many facets. However, they are impactful and, we believe that the simplicity of the technology that we've demonstrated in this short article shows great promise as a tool to illustrate trends in academic research.

4 DISCUSSION

4.1 A New World of Analysis

In a recent book Goldin and Muggah (2020) produce a set of compelling maps with associated narratives. We have tried to take the same approach in our Results section in order to showcase how these maps may lead to inquiry and contextual interpretation

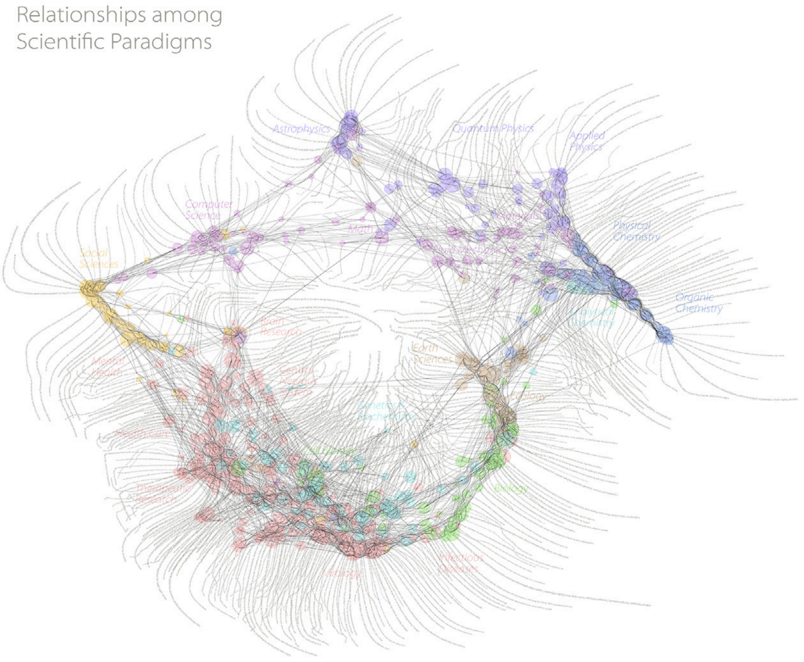

FIGURE 5 | One of the first visualisations of research that made use of a full global dataset. While "calculated", a significant amount of manual work was needed to make this beautiful visualisation, which ensures that detailed data is married with a meaningful visualisation. *Reproduced with kind permission of W B Paley.*

beyond the standard work of analysts. We have also shown how responsive and immediate these analyses can be—not only adding an interesting thread to historical discourse but allowing us to see emergent trends in real time. We believe that this type of thinking is well understood by many in the scientometric community, as evidenced by the attention received by the work of W. B. Paley (**Figure 5**) and others who originally pioneered research cartography. One of the enduring challenges of automated data visualisation is the ability to optimise layout and preserve information. In general, it is not possible to reach the level at which this is done in Paley et al.'s work. However, in making it easier to create visualisations on the fly, while we give up the data transparency that Paley aspires to, we are able to add speed of iteration so that a visualisation can be used in an actionable manner.

We have not engaged in a comparison of the merits of the visualisation used in our example here, merely noting that it can be a powerful narrative tool. It is, however, important to note that the type of visualisation that we have showcased in this example is specifically designed to showcase the use of data in the way that we have discussed. We do not assert that it is a better visualisation than others, indeed, it is a tremendously reductive visualisation that hides many features that other visualisations bring to the fore. It does not, for example, give a clear picture of where the reserach is being performed or give a sense of the research that is being performed. However, if one is willing to sacrifice some of these specificities in order to connect to historical context or narrative we continue to believe that this visualisation has some merits. As with all visualisations, there are specific situations where specific visualisations work well.

One technical point that is important to discuss is choice of coordinates. Many visualisations used in bibliometrics (including, for example, **Figure 5**, VOSviewer and CiteSpace) make use of abstract visualisation spaces, which is to say that they do not anchor to a physical map. In the case of the visualisation that we have explored here, the geography of the Earth is implicit in the visualisation. Implicit in the analysis that we have performed is a choice of coordinate system - we have chosen the one that originates from the International Meridian Conference in 1884, where Greenwich, United Kingdom was defined to be at 0° longitude. This choice is implicitly embedded in our calculation in the following sense: If a piece of research is co-authored between two colleagues, one in Beijing (around 39° N, 116° E), one in San Francisco (around 37° N, 122° W), then the centroid of the research would be in south Spain, by our coordinate definition, rather than in the Pacific ocean, which might be a more natural average in this case. A choice of coordinates naturally gives advantage to specific points in a map. We do not propose a solution here, but note that it is important to state the assumptions that are being made with all data visualisations.

It is widely recognised that data visualisation can be a powerful tool for contextualisation and interpretation (Tufte, 2001; Rendgen, 2018; Dick, 2020). The analysis presented in this paper aims to make three points: Firstly, that data accessibility is a partner to data quality and an important part of how data may be deployed to gain insight; Secondly, that data certain visualisation styles and appraoches have been previously overlooked due to the lack not only of the data accessibility, but also the need for data connectivity through persistent identifiers; Thirdly, that tools like these should not be limited only to the most well funded researchers and that cloud infrastructure may be an effective mechanism to democratise access to these types of data, tools and interpretation, and hence be a route to superior strategic decision making across the sector (Herzog et al., 2018).

4.2 A New World of Data

By introducing the scientific method in his book Novum Organum in 1620, Bacon codified the deep relationship between science and data. The importance of data is not solely limited to the scientific disciplines, rather data defined by a broad definition has always been part of research, regardless of topic. However, until relatively recently in human history, data has been rare. In the last half century we have seen an explosion in the amount of data made available not only by physical and biological experiments, but also by social experiments and also the emergence of the digital humanities. We have gone from a poverty of data to an amount of data that cannot be handled by any individual human mind.

As in the wider world of research, scientometrics has seen a rise in data availability over the last twenty to thirty years as the research community has grown and professionalised. The need for metadata that describes not only the outputs of research but also the process by which they are produced, the broad scholarly record, is now widely acknowledged.

In the next few years, we are likely to see the amount of metadata collected about a research output increase manyfold, so that the metadata about an object exceeds the data contained within the object. The ability to scale data systems, share and manipulate data and to summarise it for human consumption in visualisations is becoming critical, as is understanding the biases that are inherent to different visualisation styles.

In moving forward, we argue that critical consideration needs to be given to data accessibility. Others such as Mons (2020) have argued cogently that investment should be made into research data. We believe that investment could be helped by introducing a framework such as the one proposed here to support a working definition of data accessibility and good practice. The facets of coverage, structure, nature, context and quality, could form the basis of a helpful rubric for making research data more valuable and accessbile to the community. There is already a precedent for gaining cross-community collaboration in projects such as I4OC and I4OA as well as structures for use of metrics in DORA and the Leiden Manifesto (Hicks et al., 2015)—is data access another similar area where the community should seek to build principles to ensure the most even playing field?

4.3 Future Explorations

The methods explored in this paper can be extended and applied in many different scenarios. It is easy to see how this analysis could be repeated and customised for a variety of geographies (e.g., specific countries or regions), subject areas (e.g., COVID as shown here or Sustainable Development Goals) and timescales. Weighting schemes could include altmetric-based approaches, funding weighting, journal metric-led weighting or any number of different approaches to suit specific needs. In addition, using Dimensions, parallel analyses could be performed based on grant data, clinical trials data, patent data, pollicy documents or data. As noted previously, equivalent problems that could make use of similar capabilities and technologies include global heatmapping of specific research activities, the creation of specific custom benchmarks or other metrics to specification and on demand.

We have discussed context as a critical part of research analysis in this paper. Thus, it is important to highlight the context of the data used in our analyses. Despite the foundational principals behind Dimensions of not editorialising its data holdings, it is still not a universal dataset. At the current time, not all funding organisations make their data openly available and the publications associated with some geographies and some fields are not held in the DOI registries that have yet been integrated into Dimensions. As a result, the analysis presented here has flaws and will naturally show an english-language centred view of the world.

In this paper, we have focused on a particular analysis and visualisation style that we have not seen in the scientometric literature before. We beleive that the lack of use of this style is due to the constraints that we have outlined. However, we believe that our underlying argument around data access can be applied also to the production of visualisations such as those offered by VOSviewer, CiteSpace and similar technologies (Chen, 2006; Colavizza et al., 2021).

We close by commenting that, if adopted broadly, we believe that the cloud techniques applied in this article can lead to better decision making across academia as analysis can become more iterative and more available across the sector.

DATA AVAILABILITY STATEMENT

The datasets for **Figures 1–4** of this study can be found in the Figshare at https://doi.org/10.6084/m9.figshare.13611230.

AUTHOR CONTRIBUTIONS

DH developed the idea for this paper and drafted the manuscript and carried out the visualisation. SP developed the implementation of the code, determined the business rules and methodology for the data extraction. Both co-authors edited and reviewed the manuscript.

FUNDING

This research was supported by Digital Science and Research Solutions Limited.

REFERENCES

Allen, L., Scott, J., Brand, A., Hlava, M., and Altman, M. (2014). Publishing: credit where credit is due. *Nat. News* 508, 312–313. doi:10.1038/508312a

Balmer, B., Godwin, M., and Gregory, J. (2009). The royal society and the 'brain drain': natural scientists meet social science. *Notes Rec. R. Soc. J. Hist. Sci.* 63, 339–353. doi:10.1098/rsnr.2008.0053

Bergstrom, C. (2007). Eigenfactor: measuring the value and prestige of scholarly journals. *College Res. Libr. News* 68, 314–316. doi:10.5860/crln.68.5.7804

Börner, K. (2015). *Atlas of knowledge: anyone can map*. Illustrated Edn. Cambridge, MA: MIT Press.

Borner, K. (2010). *Atlas of science: visualizing what we know*. Illustrated Edn. Cambridge, MA: MIT Press.

Börner, K., Chen, C., and Boyack, K. W. (2003). Visualizing knowledge domains. *Ann. Rev. Inf. Sci. Technol.* 37, 179–255. doi:10.1002/aris.1440370106

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: the UCSD map of science. *PLoS One* 7, e39464. doi:10.1371/journal.pone.0039464

Bornmann, L. (2018). Field classification of publications in dimensions: a first case study testing its reliability and validity. *Scientometrics* 117, 637–640. doi:10.1007/s11192-018-2855-y

Bornmann, L., Mutz, R., and Haunschild, R. (2020). *Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases*. arXiv:2012.07675 [physics] ArXiv: 2012.07675.

Boyack, K. W., Klavans, R., and Börner, K. (2005). Mapping the backbone of science. *Scientometrics* 64, 351–374. doi:10.1007/s11192-005-0255-6

Boyack, K. W., Klavans, R., Paley, W. B., and Börner, K. (2007). "Mapping, illuminating, and interacting with science", in SIGGRAPH '07: ACM SIGGRAPH 2007 sketches, San Diego California, August, 2007 (New York, NY: Association for Computing Machinery) 2–es. doi:10.1145/1278780.1278783

Bush, V. (1945). The endless frontier, report to the president on a program for postwar scientific research. Office of Scientific Research and Development Washington DC, Technical Report.

Cervantes, M., and Guellec, D. (2002). The brain drain: old myths, new realities. *OECD Observer* 230, 40–41.

Chen, C. (2006). CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inf. Sci.* 57, 359–377. doi:10.1002/asi.20317

Colavizza, G., Costas, R., Traag, V. A., van Eck, N. J., van Leeuwen, T., and Waltman, L. (2021). A scientometric overview of CORD-19. *PLoS One* 16, e0244839. doi:10.1371/journal.pone.0244839

Dick, M. (2020). *The infographic: a history of data graphics in news and communications*. Cambridge, MA: MIT press.

Dobbs, R., Remes, J., Manyika, J., Roxburgh, C., Smit, S., and Schaer, F. (2012). Urban world: cities and the rise of the consuming class. McKinsey Global Institute, Technical Report.

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: a case study for the computation of h indices in Psychology. *J. Am. Soc. Inf. Sci.* 61, 2070–2085. doi:10.1002/asi.21372

Garfield, E., and Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *Amer. Doc.* 14, 195–201. doi:10.1002/asi.5090140304

Goldin, I., and Muggah, R. (2020). *Terra Incognita: 100 maps to survive the next 100 years*. 1st Edn. Century.

González-Pereira, B., Guerrero-Bote, V. P., and Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: the SJR indicator. *J. Informetr.* 4, 379–391. doi:10.1016/j.joi.2010.03.002

Herzog, C., Hook, D., and Konkiel, S. (2020). Dimensions: bringing down barriers between scientometricians and data. *Quant. Sci. Stud.* 1, 387–395. doi:10.1162/qss_a_00020

Herzog, C., and Lunn, B. K. (2018). Response to the letter 'field classification of publications in dimensions: a first case study testing its reliability and validity'. *Scientometrics* 117, 641–645. doi:10.1007/s11192-018-2854-z

Herzog, C., Hook, D., and Adie, E. (2018). "Reproducibility or producibility? metrics and their masters," in STI 2018 conference proceedings (Leiden, Netherlands: Centre for Science and Technology Studies (CWTS)). 685–687.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., and Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nat. News* 520, 429–431. doi:10.1038/520429a

Hook, D. W., Porter, S. J., Draux, H., and Herzog, C. T. (2020). Real-time bibliometrics: dimensions as a resource for analyzing aspects of COVID-19. *Front. Res. Metr. Anal.* 5, 25. doi:10.3389/frma.2020.595299

Hook, D. W., Porter, S. J., and Herzog, C. (2018). Dimensions: building context for search and evaluation. *Front. Res. Metr. Anal.* 3, 23. doi:10.3389/frma.2018.00023

House, W. (2020). Historical table, 9.1 - total investment outlays for physical capital, research and development, and education and training, 1962–2020. [Dataset].

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., et al. (2020). Comparison of bibliographic data sources: implications for the robustness of university rankings. *Quant. Sci. Stud.* 1, 1–34. doi:10.1162/qss_a_00031

Hurst, P. (2010). Trailblazing-350 years of royal society publishing. *Notes Rec. R. Soc.* 64, 85–89. doi:10.1098/rsnr.2009.0077

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

Larivière, V., Kiermer, V., MacCallum, C. J., McNutt, M., Patterson, M., Pulverer, B., et al. (2016). A simple proposal for the publication of journal citation distributions. *bioRxiv*, 062109. doi:10.1101/062109

López-Illescas, C., de Moya Anegón, F., and Moed, H. F. (2009). Comparing bibliometric country-by-country rankings derived from the Web of Science and Scopus: the effect of poorly cited journals in oncology. *J. Inf. Sci.* 35, 244–256. doi:10.1177/0165551508098603

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., and Delgado López-Cózar, E. (2018). Google scholar, web of science, and scopus: a systematic comparison of citations in 252 subject categories. *J. Informetr.* 12, 1160–1177. doi:10.1016/j.joi.2018.09.002

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., and Delgado López-Cózar, E. (2020). Google scholar, microsoft academic, scopus, dimensions, web of science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126, 871–906. doi:10.1007/s11192-020-03690-4

Miodownik, M. (2014). *Stuff matters: the strange stories of the marvellous materials that shape our man-made world*. 1st Edn. London: Penguin.

Mongeon, P., and Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics* 106, 213–228. doi:10.1007/s11192-015-1765-5

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature* 578, 491. doi:10.1038/d41586-020-00505-7

Morris, J., and Noble, D. (1993). Science and economic policy in the United Kingdom. *Physiology* 8, 136–140. doi:10.1152/physiologyonline.1993.8.3.136

Powell, K. R., and Peterson, S. R. (2017). Coverage and quality: a comparison of web of science and scopus databases for reporting faculty nursing publication metrics. *Nurs. Outlook* 65, 572–578. doi:10.1016/j.outlook.2017.03.004

Rachman, G. (2017). *Easternisation: war and peace in the asian century*. 1st Edn. New York, NY: Vintage.

Rendgen, S. (2018). *The minard system: the complete statistical graphics of Charles-Joseph Minard*. New York, NY: Princeton Architectural Press.

Thelwall, M. (2018). Dimensions: a competitor to scopus and the web of science?. *J. Informetrics* 12, 430–435. doi:10.1016/j.joi.2018.03.006

Tufte, E. R. (2001). *The visual display of quantitative information*. 2nd Edn. Cheshire, CT: Graphics Press.

van Eck, N. J., and Waltman, L. (2019). Accuracy of citation data in web of science and scopus. arXiv:1906.07011 [cs] ArXiv: 1906.07011.

Van Noorden, R. (2016). Controversial impact factor gets a heavyweight rival. *Nature* 540, 325–326. doi:10.1038/nature.2016.21131

Visser, M., van Eck, N. J., and Waltman, L. (2021). Large-scale comparison of bibliographic data sources: scopus, web of science, dimensions, crossref, and microsoft academic. *Quant. Sci. Stud.*, 1–22. doi:10.1162/qss_a_00112

Waltman, L. (2020). Open metadata: an essential resource for high-quality research intelligence. [Dataset]. doi:10.5281/zenodo.4289982

Wolfram, S. (2010). Making the world's data computable. *16:58:17; Conference: transcript of Stephen Wolframs keynote talk from the first Wolfram Data Summit, in Washington, DC*, September 9, 2010. https://writings.stephenwolfram.com/2010/09/making-the-worlds-data-computable/ (Accessed January 01, 2021).

Wood, M. (2020). *The story of China: a portrait of a civilisation and its people*. London, United Kingdom: Simon & Schuster UK.

Zukowski, M. (2018). "Cloud-based SQL solutions for big data," in *Encyclopedia of big data technologies*. Editors S. Sakr and A. Zomaya (Cham: Springer International Publishing), 1–7. doi:10.1007/978-3-319-63962-8_318-1

Conflict of Interest: The authors declare that this study was funded by Digital Science. Both authors and investigators are employees of Digital Science and the investigations presented here were undertaken as part of the normal operation of the business. No special funding or conditions were attached to this work.