Digital Science White Paper

Introducing Dimensions Research Integrity

Powered by Ripeta

Leslie D. McIntosh, Ruth Whittam, Simon Porter, Cynthia Hudson-Vitale, and Misha Kidambi

FEBURARY 2023

DIGITAL science

Dimensions

About Digital Science

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

About the Authors

Leslie D. McIntosh is the CEO and founder of Ripeta
https://orcid.org/0000-0002-3507-7468

Ruth Whittam is the Lead Data Scientist at Ripeta
https://orcid.org/0000-0002-6151-8423

Simon Porter is the VP Research Futures at Digital Science
https://orcid.org/0000-0002-6151-8423

Cynthia Hudson Vitale is the co-Founder of Ripeta
https://orcid.org/0000-0001-5581-5678

Misha Kidambi is the Science Communications Specialist at Ripeta

Digital Science is part of Holtzbrinck, a global media company dedicated to science and education.

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

DOI: dx.doi.org/10.6084/m9.figshare.21997385
This report was created in Overleaf

Contents

1 Executive Summary **1**

2 Introduction **2**
2.1 From Ripeta models to Dimensions Research Integrity 2
2.2 Trust Markers and Dimensions Research Integrity 2

3 Trust Markers: Current metrics and approach **4**
3.1 Understanding the Ripeta Models behind Research Data Integrity 6

4 Early Findings **7**
4.1 Observation 1. Evolving Science Trust Markers 2011-2021 7
4.2 Observation 2. Research disciplines, publishers, funders, and institutions all play a significant role in influencing trust marker uptake 8
4.3 Observation 3. Repository Share: Data Availability Statement Repository Mentions 11
4.4 Observation 4. The way researchers communicate author contributions is rapidly changing 12

5 What can trust markers be used for? Opportunities for new business processes **13**
5.1 How to access Dimensions Research Integrity 14

6 Conclusion/Discussion **14**

1 Executive Summary

Trust markers - the explicit statements on a paper such as funding, data availability, conflict of interest, author contributions, and ethical approval - represent a contract between authors and readers that proper research practices have been observed. Trust markers highlight a level of research transparency within a publication and reduce the reputational risks of allowing non-compliance to research integrity policies to go unobserved.

Dimensions Research Integrity introduces a new ability to measure the uptake and usage of trust markers across the global published landscape, based on the analysis of 33M full text articles.

This paper outlines how the Research Integrity dataset was created from algorithms developed at Ripeta. Also we show why it is important for publishers, funders, and institutions to systematically measure trust markers across their articles.

Early Observations using Dimensions Research Integrity show:

- The use of trust markers in scientific literature has increased dramatically over the last decade.
  - Trust markers related to ethics approval, funding, and competing interest statements have become well established in the research community.
  - Processes around ensuring data availability and providing author contribution statements have also improvemed over the past few years.
- Different publishers have prioritised the inclusion of different trust markers at different rates.
- The adoption of trust markers differs depending on the fields of research, suggesting the need for different outreach strategies based on research disciplines.

The Dimensions Research Integrity dataset is available as a module extension to the Dimensions Google BigQuery offering. Consultancy reports for individual funders, publishers, and institutions are also available.

2 Introduction

2.1 From Ripeta models to Dimensions Research Integrity

Research integrity and trust in science are at the forefront of scientific communications. As early as 2013, the World Economic Forum cited the growth of misinformation and disinformation as a global risk, especially in high-stakes and volatile situations, where false information or inaccurately presented imagery can cause damage before it is possible to communicate accurate information. More recently, in the United States, the Biden-Harris administration has signalled their commitment to increasing the integrity of government and federally funded research as a mechanism to mitigate misinformation and strengthen public trust in science. Similar measures have been adopted by governments in the United Kingdom, the European Union, the African Union, and other regions. All organisations involved in scientific publishing are aware that the integrity and trustworthiness of a piece of research is of comparable importance to the attention and citation it receives. To ensure the quality of research, international organisations and bodies have been established to develop guidelines and best practices; these include the Hong Kong Principles and the Singapore Statement on Research Integrity.

Ripeta developed sophisticated methodologies and tools to improve the integrity and reproducibility of scientific research to ensure trust in science. Ripeta's tools examine the content of published papers and look for **trust markers**. - the hallmarks of responsible science: a clearly stated study objective, a statement of how the research is funded, guidance about how to obtain a copy of the study data, and many others. These trust markers are increasingly prevalent across scientific research, often mandated by publishers and funders. They shed light both on research integrity and research practices. For example, they can be used to **understand where researchers are storing their data** allowing universities to provide better support for certain tools. They **help funders visualize the impact of policy decisions**, and they **help publishers monitor the adoption of article templates**.

2.2 Trust Markers and Dimensions Research Integrity

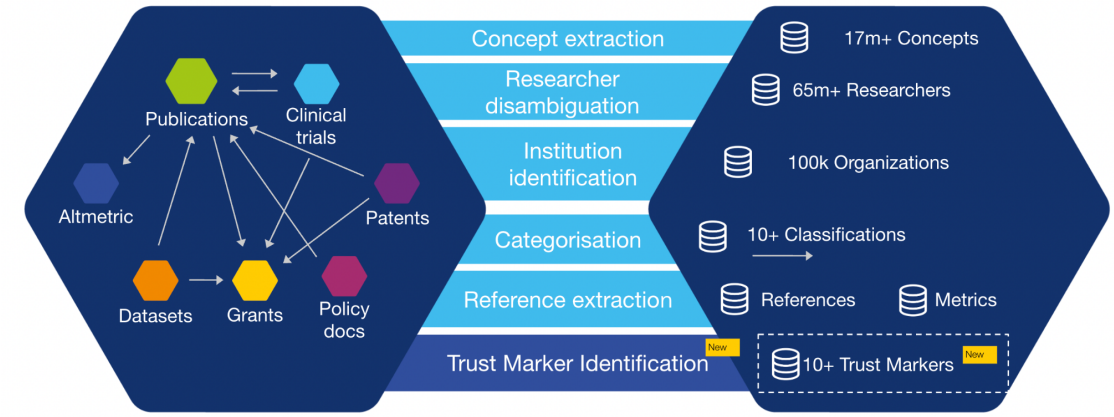

Figure 1: Trust Markers included into the linked information network within Dimensions

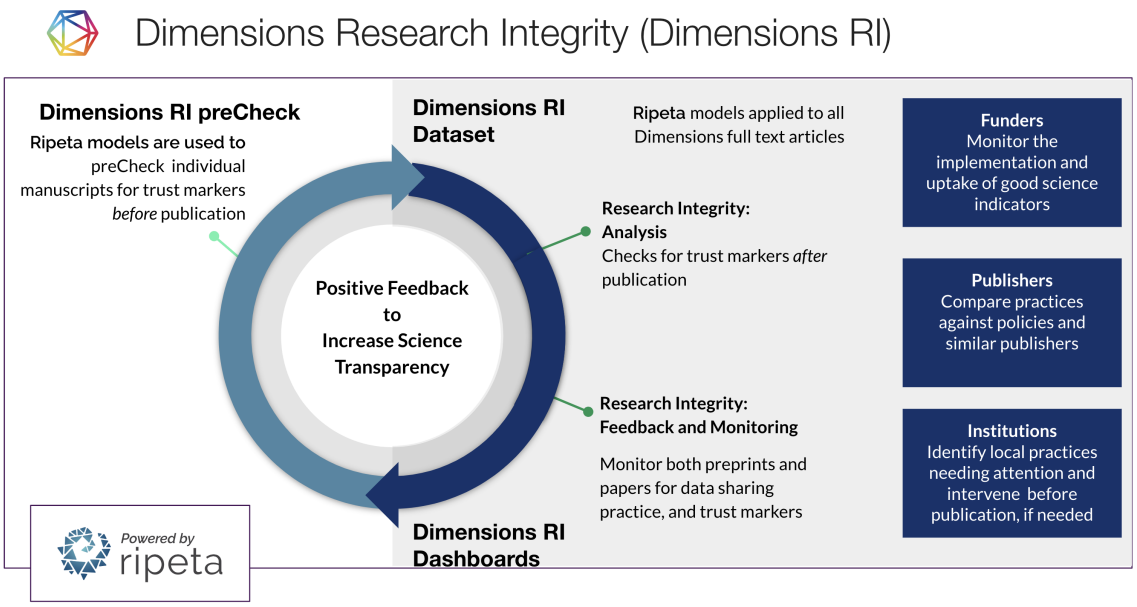

Figure 2: Dimensions Research Integrity: Creating a positive feedback loop to improve Research Integrity

Dimensions Research Integrity brings the Ripeta methodology into the Dimensions ecosystem, detecting the presence of **trust markers across 33M research articles, conference proceedings, book chapters, and preprints.** Coverage of Dimensions Research Integrity is from 2010, and includes all articles with full text available in Dimensions.

This data is offered as a Google Big Query (GBQ) Dimensions module to publishers, funders, and institutions.

Dimensions Research Integrity is also available as an API for use in manuscript submission workflows, providing a holistic solution that both aims to improve as well as measure research integrity practice (Figure 2).

3 Trust Markers: Current metrics and approach

Trust markers are a new type of article metadata representing the integrity and reproducibility of scientific research. Found in Dimensions GBQ, trust markers are of interest to customers within the publishing, academic, and funding spaces. Trust markers represent a contract between authors and readers that proper research practices have been observed. Trust markers highlight the level of research transparency within the document and reduce reputational risks by checking and flagging non-compliance to research integrity policies.

Trust Markers are individual elements that allow us to understand, classify, and categorise trust in science. Ripeta applies the term to elements of written scientific communications that help to build trust. For example, the Trust Markers identified by Dimensions Research Integrity address **reproducibility** and **transparency**.

- Trust in **reproducibility** is centred around the elements of a paper which may facilitate the ability to achieve the same results when replicating or reproducing the original study.

- Trust in **transparency** is based around ascertaining the legitimacy of the authors and whether their reporting adheres to established standards of scientific communication.

Reproducibility and transparency trust markers sit alongside other indicators of how an article should be read, such as whether or not an article has been peer-reviewed. These document type markers are already a standard component of Dimensions, and Dimensions Research Integrity uses document type classifications to select which articles to analyze.

The table below lists the current trust markers developed by Ripeta and indicates those that will be available in the first release of Dimensions Research Integrity.

Dimensions Research Integrity Trust Markers

Category	Trust Marker	Description	Dimensions Research Integrity 1st Release
Transparency	Funding statement	States if the author(s) of the paper were granted funding in order to conduct their research.	Y
	Ethical approval statement	Statement affirming that the conducted research has been carried out in an ethical fashion with proper consent from all participating parties.	Y
	Competing interests statement	Declares possible sources of bias, based on personal interests of the author(s) in the findings of the research. For example, the source of funding, past or present employers of the author(s), or the author(s) financial interests.	Y
	Author contribution statement	Details of each author's role in the development and publication of the manuscript.	Y
Reproducibility	Repositories	The names of any research data repositories used by the author(s) to preserve, organize and facilitate access to study data.	Y
	Data locations	Locations where research data (raw or processed) can be accessed.	Y
	Data availability statement	A dedicated section of a scientific work indicating whether data from the research is available and where it can be found.	Y
	Code availability statement	States if and how one could gain access to the code used to conduct the study/research.	Y (in development)
	Analysis software	Softwares used to conduct the experiment (includes version).	N (in development)

Created with Datawrapper

Table 1: **Trust Markers in Dimensions**
The Trust Markers identified by Dimensions Research Integrity address reproducibility and transparency. Trust in reproducibility is centred around the elements of a paper which may facilitate the ability to achieve the same results when replicating or reproducing the original study. Trust in transparency is based around ascertaining the legitimacy of the authors and whether their reporting adheres to established standards of scientific communication.

3.1 Understanding the Ripeta Models behind Research Data Integrity

Data

The training, evaluation, and validation dataset spans many fields – from Medical and Health Science, to Studies in Creative Arts and Writing – to ensure that the datasets used for the model were not concentrated or biased towards a single field. On average, the model for each trust marker incorporated in Dimensions Research Integrity was trained, evaluated, and validated across 10 different fields of study. Because a significant difference in transcripts was observed between papers published at the start of the decade and those published at the end of the decade, Ripeta's pipeline included papers published in 2011 as well as papers published in 2021.

Classification results

Preprocessing includes converting the PDF to text strings, in order to extract and isolate segments that resonate with the definition of each trust marker. Each isolated text segment is then labeled using https://spacy.io/universe/project/prodigy/ then used for training, evaluation, and validation purposes. At the time of writing this paper, 19.3K text segments (e.g., sentences and paragraphs) had been used in training, 4.1K text segments in evaluation, and 1.6K text segments in validation (see Table 2).

Resulting Models were then applied to Dimensions to create Dimensions Research Integrity.

Trust Marker	Training Dataset	Evaluation Dataset	Validation Dataset	f1 - score
Data Availability Statement	6K	1.7K	233	0.98
Data Locations	1.7K	437		0.84
Ethical Approval Statement	1.8K		246	0.93
Funding Statement	785	191	245	0.95
Open Access Statement	501	192	227	0.87
Code Availability Statement	1.7K		216	0.96
Competing Interest Statement	284	207	247	0.97
Repositories	2K	953		0.86
Author Contributions Statement	425	98	232	0.97

Table 2: **Trust Marker Validation Scores**

4 Early Findings

4.1 Observation 1. Evolving Science Trust Markers 2011-2021

Source: Dimensions Research Integrity via Google BigQuery

Trust Markers in research have increased dramatically over the last 10 years, with each Marker on a path to become an established component of research community practice.

Prevalence of funding statements, already established as community practice in 2011, has risen steadily from just over 30% to just under 50% in 2021.

Over the same period, the presence of competing interest statements and ethical approval statements have seen a rapid uptake in practice, rising from under 8% in 2011 to 38% and 32% respectively. Author contributions have also increased steadily to just under 26% in all research articles selected.

Research community practice for the addition of data availability statements has taken longer to develop but is now rapidly seeing adoption. In 2020, data availability statements were only observed on 10% of papers. In one year, this percentage has more than doubled to slightly above 22%.

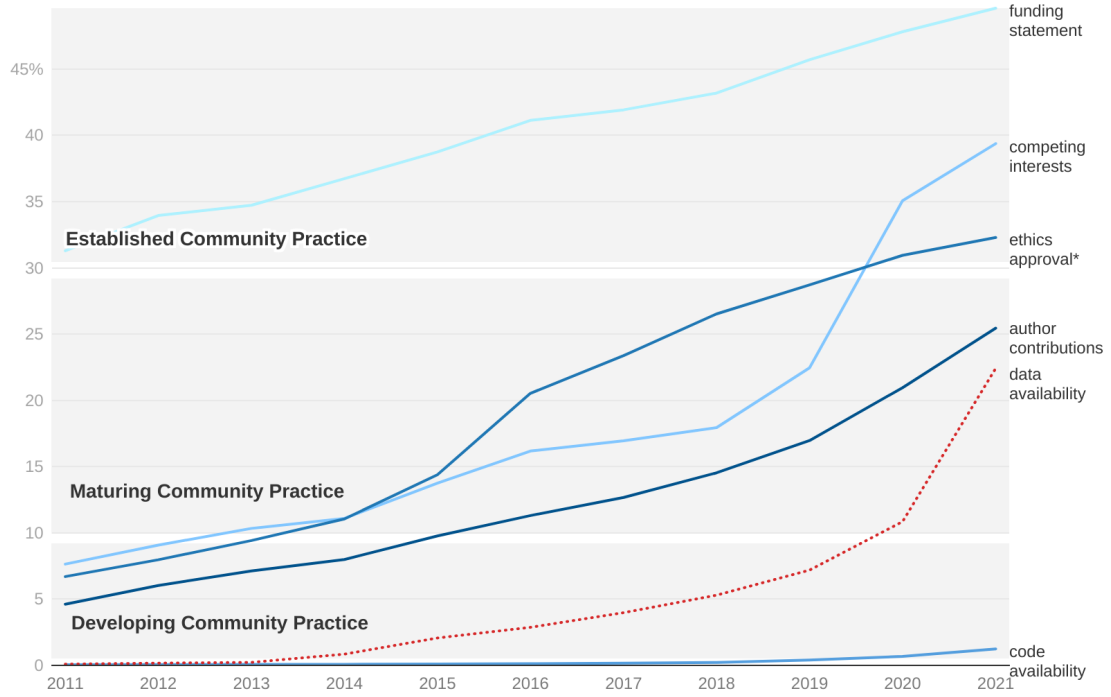

Figure 3: **Evolving Science Trust Markers 2011-2021**
The percentage of ethics papers are calculated over publications a mesh classification of Humans or Animals. The ethics trust marker looks at those papers that include a specific ethics section (as opposed to mentioning ethics approval somewhere in the text).

4.2 Observation 2. Research disciplines, publishers, funders, and institutions all play a significant role in influencing trust marker uptake

Journal publishers are effecting change in data sharing practices. In 2014, Lin and Strasser published a set of recommendations for the role of publishers in access to data and a call to action to implement policies that make data sharing a fundamental practice of scientific communication. In 2015, the Transparency and Openness Promotion (TOP) guidelines were published, which further encouraged journal publishers to adopt data availability and sharing statements. While some journals, such as Nature, have had policies since 2013, these "calls-to-action" accelerated the adoption of data availability statement policies, including PLOS(2016), AAAS (2016), and The Royal Society (2016), to name a few.

At a high level, the presence of Trust Markers reveal different publisher approaches. As can be seen in figure 4, newer, open access (OA) publishers have been quicker to adopt Trust Markers in their papers. The reasons for this are complex, but anecdotally, OA publishers have typically grounded their publishing mission and values in open scholarship and open science, which as defined by UNESCO, refers to open access to scientific publications, research data, metadata, open educational resources, software, and source code and hardware. This commitment to open scholarship and open science has catalysed not only their adoption of policies but also the implementation of Trust Markers as a compliance mechanism for those policies. Open Access (OA) publishers also grew exponentially in the early years of the internet, which allowed for the development of simpler, lower cost, online workflows. Being somewhat newer to publishing further allowed OA publishers the flexibility to quickly adapt to increasing expectations for more robust reporting requirements and federal-based policies for research integrity.

publisher	first published year	publications	data availability statement percentage	authors contribution statement	competing interests statement	funding statement percentage	ethics approval
Public Library of Science (PLoS)	2,004	19,977	98	99	100	89	30
Frontiers	1,992	64,928	97	99	32	80	76
AIP Publishing	1,932	13,061	96	20	8	61	5
Hindawi	1,978	31,728	92	24	17	58	23
MDPI	1,999	154,687	84	97	90	97	14
BMJ	1,841	13,935	60	17	99	75	59
Springer Nature	1,845	342,960	41	48	41	68	68
Oxford University Press (OUP)	1,670	28,642	27	23	25	59	20
Wiley	1,822	189,776	25	16	11	50	17
IOP Publishing	1,890	74,795	14	1	1	34	14

Figure 4: **Trust Marker Coverage by Publisher 2021**
Publishers are ordered by data availability coverage. *The percentage of papers with ethics approval are calculated over publications a mesh classification of Humans or Animals. the ethics trust marker looks at those papers that include a specific ethics section (as opposed to mentioning ethics approval somewhere in the text)

At a deeper level, Trust Markers also reveal patterns of researcher behaviour, be it the need to encourage more researchers to deposit data in online repositories, or identify the repositories that researchers are using so that they can be better supported.

There has been a significant body of research focused on understanding disciplinary differences and expectations for adopting research practices that align with the Trust Markers. In the oft-cited book, Big Data, Little Data, and No Data, Dr. Christine Borgman presents a number of case studies of data practices and research methods for researchers in the sciences, the social sciences, and the humanities. In general, what "research data are/is" varies by discipline, research methods, type

of data collected, and study type. This may also have a significant effect on how data availability statements are written by researchers and if data is shared.

Figure 5 provides an example of how Dimensions Research Integrity can be used to track these discipline differences at a broad field level, whilst Figures 6 and 7 provide examples of more fine-grained analysis, by drilling down into a single field of research (Chemical Sciences). One can then look at the share of articles by publisher that involve data availability statements, then, within these publications, what percentage of papers provide links to data repositories.

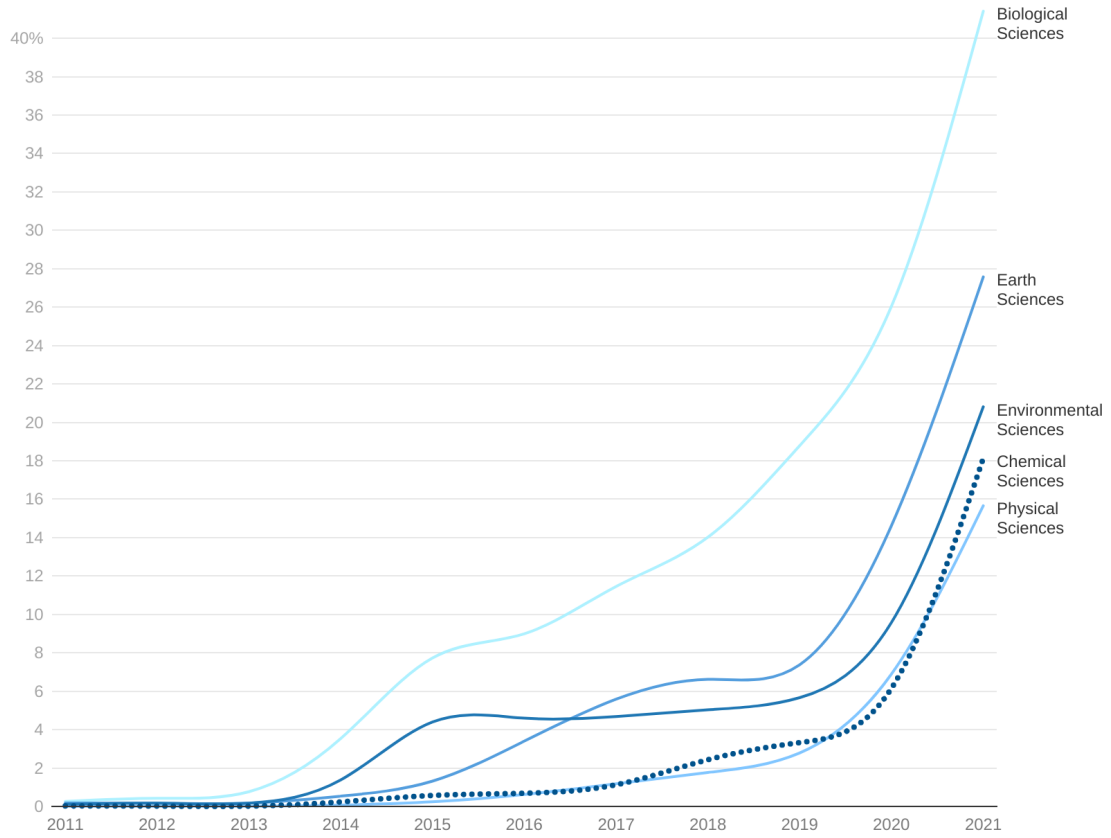

Figure 5: **Data Statement Coverage by Selected Research Categories 2011-2021**

publisher	publications	publications with data statements	data availability statement percentage
Elsevier	62,746	2,839	4.52%
American Chemical Society (ACS)	39,789	490	1.23%
Wiley	28,663	8,578	29.93%
Royal Society of Chemistry (RSC)	27,201	810	2.98%
MDPI	20,014	17,292	86.40%
Springer Nature	19,461	5,517	28.35%
Pleiades Publishing	4,193	4	0.10%
Taylor & Francis	3,276	112	3.42%
AIP Publishing	1,714	1,600	93.35%
The Electrochemical Society	1,563	48	3.07%

Figure 6: **Chemical Sciences Data Statement Coverage by Publisher**

publisher	publications	publications with links to online repositories	online repository percentage
Elsevier	62,746	224	0.36%
American Chemical Society (ACS)	39,789	329	0.83%
Wiley	28,663	209	0.73%
Royal Society of Chemistry (RSC)	27,201	170	0.62%
MDPI	20,014	1,028	5.14%
Springer Nature	19,461	732	3.76%
Pleiades Publishing	4,193	0	0.00%
Taylor & Francis	3,276	15	0.46%
AIP Publishing	1,714	85	4.96%
The Electrochemical Society	1,563	3	0.19%

Figure 7: **Chemical Sciences Data Statement Coverage by Publisher**

4.3 Observation 3. Repository Share: Data Availability Statement Repository Mentions

There is a marked difference between having data availability statements, and making data available in an appropriate repository. Within those papers that make data available, GitHub has become a code/data 'repository' of choice for many researchers. But GitHub does not provide a persistent version of record, leaving the data and code vulnerable to future deletions.

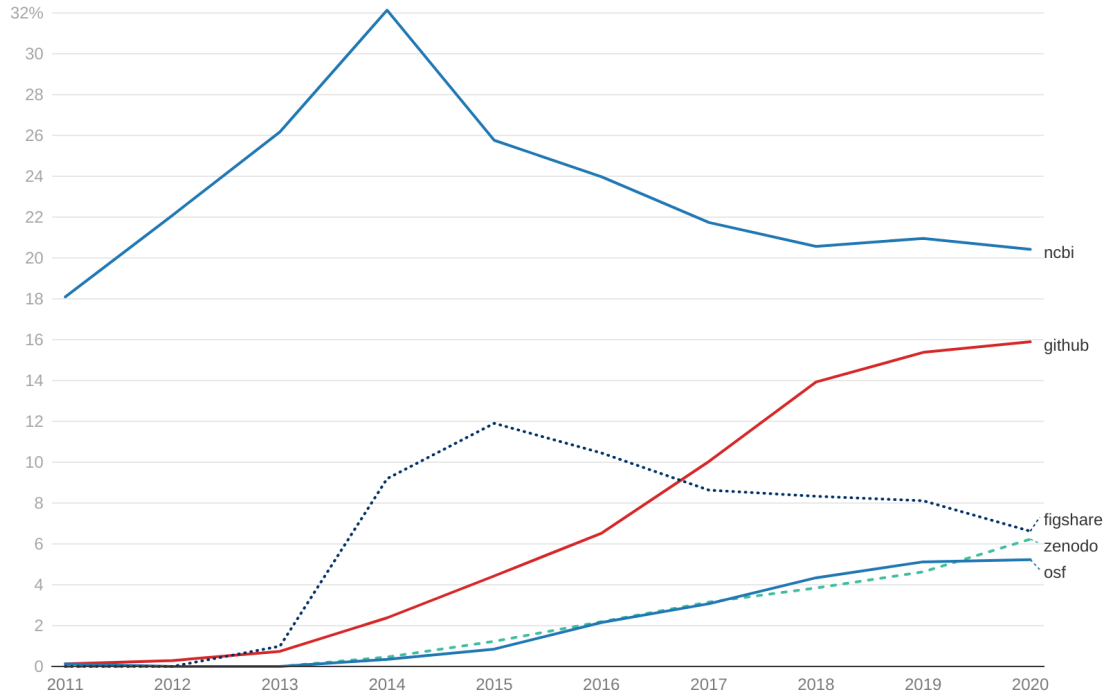

Figure 8: **Repository Share: Data Availability Statement Mentions**

4.4 Observation 4. The way researchers communicate author contributions is rapidly changing

In addition to tracking author contribution statements, Dimensions Research Integrity also tracks the verbs describing these contributions. As Figure 9 illustrates, the use of different verb groups has evolved at different paces. Over the last 10 years, there has been a steady increase in the verbs 'performed', 'wrote', and 'designed', with the verbs 'approved' and 'contributed' reaching similar frequencies in recent years. The verbs 'analyzed', and 'conceived' share similar, more modest adoption paths, whilst the rapid rise in 'agreed', and 'published' potentially signal the evolution of new language being added to Author Contribution Statements: "All authors have read and **agreed** to the **published** version of the manuscript."

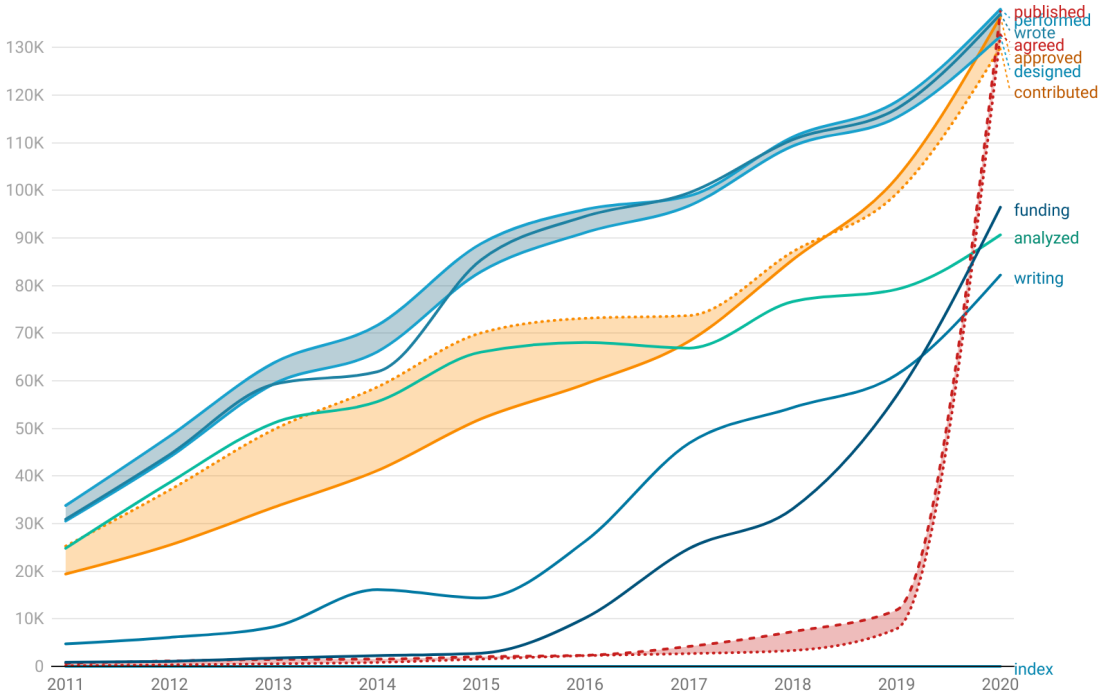

Figure 9: **Author contribution verbs all articles 2021**

5 What can trust markers be used for? Opportunities for new business processes

Uses for Dimensions Research Integrity are described below.

Because Trust Markers represent integrity and reproducibility of research, Trust Markers act as a contract between authors and readers to guarantee proper research practices have been observed. The Trust Markers sit alongside other indicators of how an article should be read, such as whether or not it has been peer-reviewed.

Publishers, funders, and institutions all have strong incentives to encourage the increased use of trust markers across all research papers with which they are associated. Further, each of these bodies has different levers that can be implemented to encourage change. Funders can set policy recommendations and mandates, publishers can set local policies and facilitate more thorough data collection, and institutions can provide targeted education on best practices.

For each of these interventions, Dimensions Research Integrity provides a way to assess whether researcher behaviour has changed as a result. Table 3 provides an overview use cases by organization type .

Dimensions Research Integrity Use Cases	Example	Publishers	Funders	Institutions
Benchmark transparent science practices across publishers and journals.	Are your journals meeting your benchmarks and community standards?	✔	✔	
Monitor the effect of policy on transparent science practice	How many of my funded papers have data availability statements	✔	✔	
Monitor data sharing practices, and repository preferences across research.	Which repositories are being used?	✔	✔	✔
Identify areas of research that require further process attention	Find areas of research with MeSH classifications of Animals/Humans but no separate ethics statement		✔	✔
Monitor competing interest statements	Are there papers with authors from companies that do not have competing interest statements?	✔	✔	✔
Identify research areas that are 'change opportunities'	Are there large pockets of 'data available upon request,' where similar research is being made available in repositories?		✔	✔
Audit repository deposit practice	Are repository links in the right format? (Github repositories also backed up to a repository with doi, no links to private sharing urls or google-drive... for preprints - intervene before publication.	✔	✔	✔
Identify areas of good practice				✔
Through the analysis of Author Contribution Statements, to identify acknowledgement patterns	Identify areas/journals/publishers that could be change agents for enhanced metadata practices (such as the credit ontology)	✔	✔	✔

Table 3: **Dimensions Research Integrity use cases**
For each use case, it is indicated whether it applies to funders, publishers, or research institutions.

5.1 How to access Dimensions Research Integrity

Dimensions Research Integrity is available as a separate Dimensions GBQ module. In early 2023, we will engage with our scientometric partners to review, validate, and improve the data. Their feedback will allow us to improve our processes and the resultant data. In the spirit of research integrity, we will continually make improvements to the product based on the needs and feedback of the community.

6 Conclusion/Discussion

Today, research integrity and trust in science are at the forefront of the scientific communications field. On one hand, implementing transparent research practices and the sharing of research data, protocols, and code have become crucial because they accelerate scientific progress and solve real-world problems in fields ranging from health and medicine to the environment and society. On the other hand, open science with its broad and open research sharing practices has revealed issues in trust, particularly with limited checks on scientific integrity. Open access, an increasing number of manuscript submissions, and a growing list of reporting guidelines all contribute to the tremendous stress faced by the traditional structures of the publishing ecosystem such as the peer review process.

As a result, questionable scientific veracity and mis- or disinformation has spread among the publishing industry, institutions, funding agencies, researchers, and also the general population. Simultaneously, the growing importance on the impact, metrics, or citations of a research article or output means that research assessment is 'attention-based' instead of quality-based. With an overemphasis on the article review process and the attention through citations, there is a real danger that attention generated by a research article or output is confused with its quality. There is a critical need to assess the trustworthiness, quality, and integrity of research in a rapidly expanding international system, and indeed the need for non-attention-based metrics is of increasing importance - especially to mitigate misinformation and increase research integrity.

Dimensions Research Integrity developed on top of Ripeta's innovative methodologies and tools establish a means to uphold integrity in scientific research and strengthen trust in science.

Part of **DIGITAL** science

digital-science.com