Enabling Registration, Identification, Integration, and Replication on OSF

Eric Olson

eric@cos.io

https://orcid.org/0000-0002-5989-8244

MagLab Issues Faced

- Data management planning
- Org wide policy adherence and adoption
- Disparate levels of awareness
- Lack of transparency once data leaves the facility

Mission:

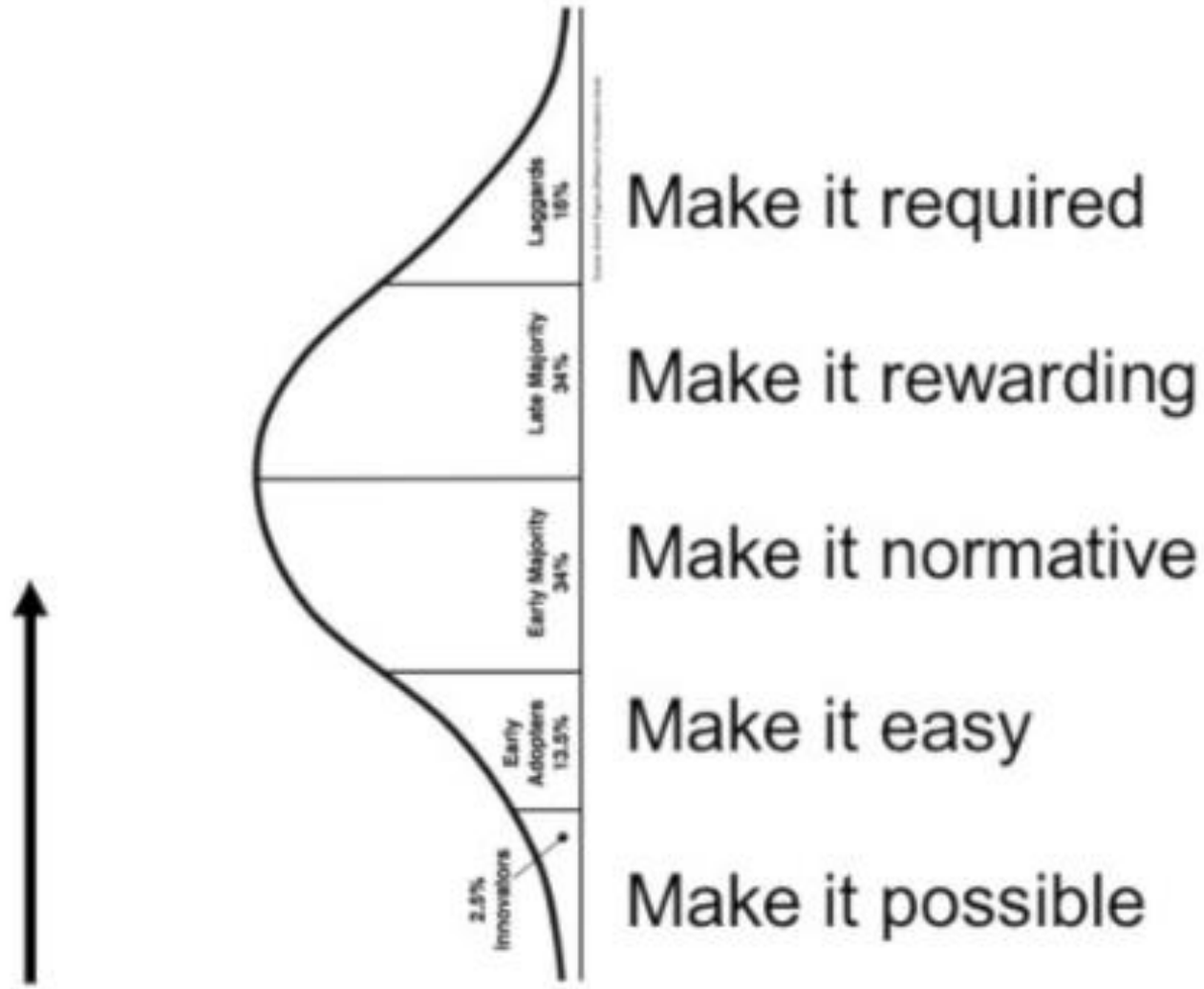
To increase the openness, integrity and reproducibility of research.

Non-profit, Open-source, Free for researchers

How do we accomplish the mission?

CO

ntegrity, and



red

arding

mative

easy

it possible

mmunities,
rigor and

Transparency and Openness Promotion (TOP) Guidelines, 8 Standards

- Data citation
- Data, Materials, and Code Transparency
- Design and Analysis Transparency (RGs)
- Preregistration (with analysis plans)
- Replication (with Registered Reports)

TOP Guidelines, 3 Levels

- Not compliant
- Disclose
- Require
- Verify

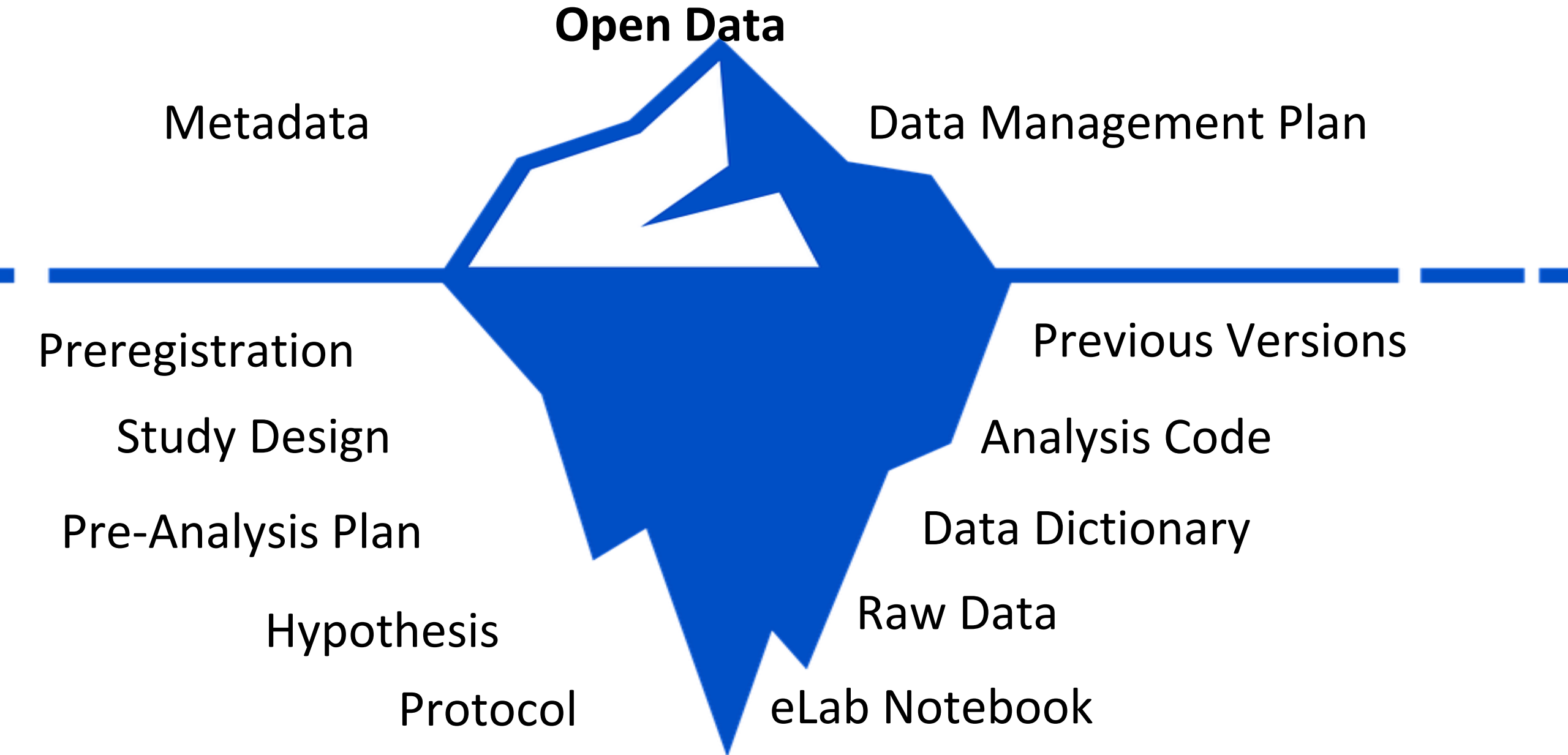
Communities

Communities enabling open practices

Researchers are more likely to adopt reproducible practices when they are backed by community leadership and support. Learn about the groups, institutions, and funders partnering with COS to equip their researchers with open infrastructure, methods, and training.

Why Reproducibility?

Anatomy of the Open Data Iceberg



CORRESPONDENCE

LINK TO O

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

A recent report by Arrowsmith noted that the success rates for new development projects in Phase II trials have fallen from 28% to 18% in

to 'feasible/marketable', and the financial costs of pursuing a full-blown drug discovery and development programme for a particular tar-

results that are publish
duce. However, there is
this apparently widespre
public recognition (for
and the surprisingly fe
tions dealing with this
knowledge, so far there
in-depth, systematic a
reproduced results with
wet-lab experiments rela
tion and validation.

Early research in the
try, with a dedicated bud
mainly work on target

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored

reproducibility of publish
ate our incidental obser
reports are frequently
quantitative data, we p
of our early (target iden
tion) in-house projects i
fields of oncology, wom

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2×2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P.A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

SOURCES OF ISSUES IN REPRODUCIBILITY

- Methodological, statistical, and reporting practices
- Structural and organizational practices
- Rarely, intentional scientific misconduct

WHAT IS REPRODUCIBILITY?

Computational Reproducibility:
If we took your data and code/analysis scripts and reran it, we can reproduce the numbers/graphs in your paper

Methods Reproducibility:
We have enough information to rerun the experiment or survey the way it was originally conducted

Results Reproducibility/Replicability:
We use your exact methods and analyses, but collect new data, and we get the same statistical conclusion

WHY SHOULD YOU CARE?

- Increases efficiency of your own work
- Can reduce false leads
- Data sharing citation advantage

> It takes some effort to organize your research to be reproducible...the principal beneficiary is generally the author herself.
>
> – Jon Claerbout

Making Scientific Contributions Reproducible
sepwww.stanford.edu/oldsep/matt/join/redoc/web/iris.html

RESEARCH / DATA MANAGEMENT PLANNING

1. What are you going to store?
2. Where and how are you going to store it?
3. Who will have access to it?
4. When will they have access to it?

RESEARCH / DATA MANAGEMENT PLANNING

1. Wha
2. Whe
3. Who

PLAN AHEAD

Checklists and common structures are your best friend!

RESEARCH / DATA MANAGEMENT PLANNING

Could do this internally on a lab server, personal computer or website, but:

- Makes eventual sharing more work
- Unclear how stable/accessible that will be in the long run
- Cross lab/institution collaborations harder

PREREGISTRATION

Documenting your research plan in a read-only public repository before you conduct the study.

Practice originated in clinical research and is now expanding more broadly.

PREREGISTRATION

Benefits of preregistering your study depend on how much information you include. At a minimum a preregistration should include the "what" of a study:

- Research question
- Population and sample size
- General design
- Variables to be collected, or dataset you'll be using

Here is a great example of a registered report: https://osf.io/2ds52

WHY PREREGISTER?

- Preregistration helps reduce the “file drawer effect” by increasing discoverability of unpublished studies.
- Preregistered analysis plans help improve study accuracy and replicability by guarding against unintended false positive inflation.

STEPS

1. Create a structured workspace
2. Preregister study
    - Document research plan
    - Make public snapshot
3. Add materials from study
4. Add and document analyses
5. Share study data, code, and materials

“Ok, that seems like a good idea, but how do I do all of that stuff you just said?”

-Anyone who hears this talk

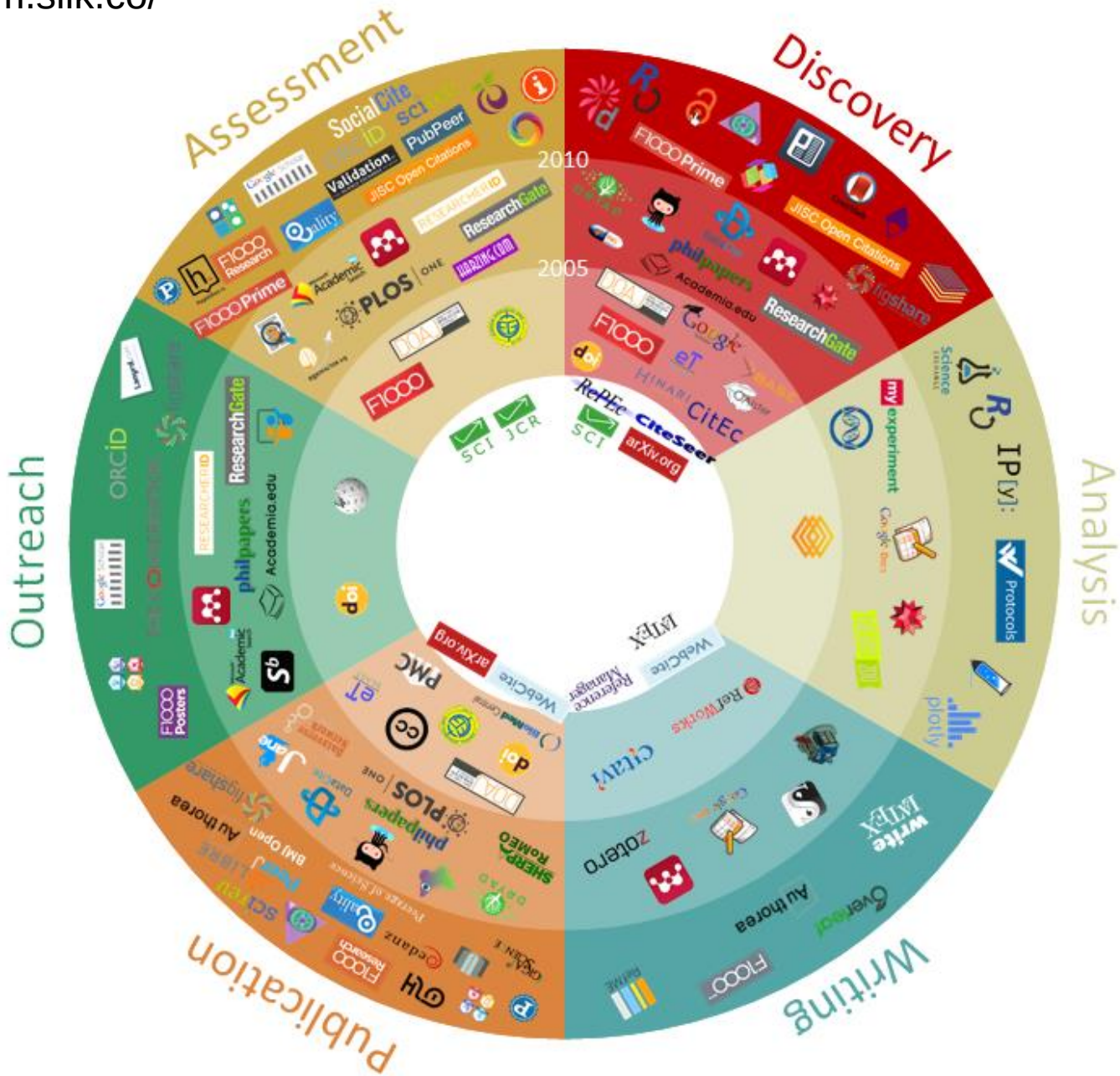
How do we accomplish the mission?

COS Mission: to increase openness, integrity, and reproducibility of research

COS Product Vision: To empower communities, institutions, and funders to advance rigor and transparency of research.


The Research Lifecycle
Search and Discover
Develop Idea
Design Study
Acquire Materials
Collect Data
Store Data
Analyze Data
Interpret Findings
Write Report
Publish Report

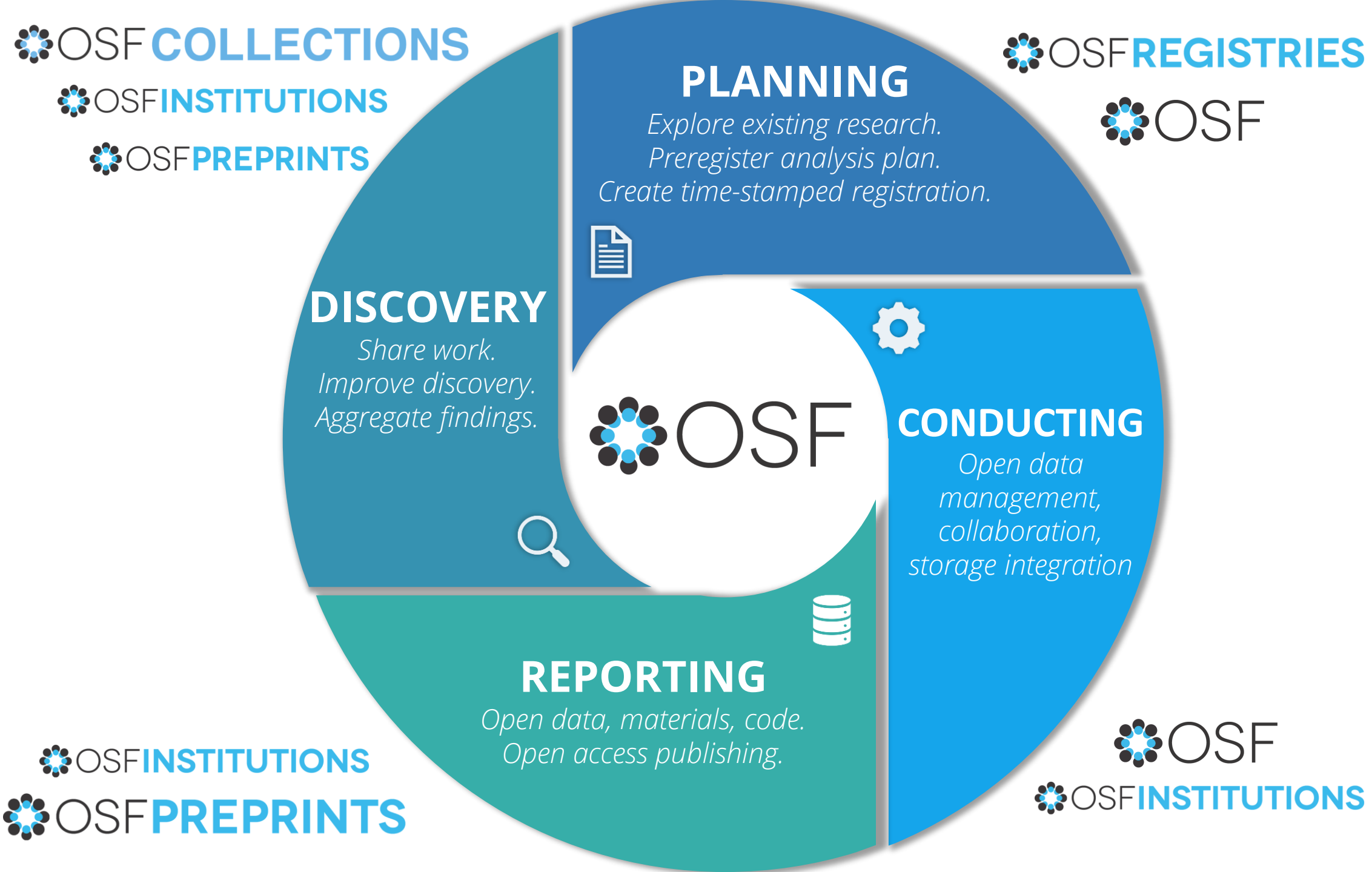
Source: https://innoscholcomm.silk.co/



Source: Kramer B, Bosman J. Innovations in scholarly communication - global survey on research tool usage. F1000Res. 2016 Apr 18;5:692. doi: 10.12688/f1000research.8414.1.

Open Science Framework (OSF)

- A cloud-based, open source collaborative management service that ensures researchers will never lose their own work.
- Researchers deposit research data, materials, and docs into flexible project spaces for their teams.
- They control who has access to what and when, and can decide whether to make some or all of the content public.
- **Interoperability with additional research tools** and multiple interfaces can accommodate dynamic research workflows across all research disciplines and throughout the lifecycle.


OSF COLLECTIONS
OSF INSTITUTIONS
OSF PREPRINTS
OSF REGISTRIES
OSF
PLANNING
Explore existing research.
Preregister analysis plan.
Create time-stamped registration.
DISCOVERY
Share work.
Improve discovery.
Aggregate findings.
OSF
CONDUCTING
Open data management, collaboration, storage integration
REPORTING
Open data, materials, code.
Open access publishing.
OSF INSTITUTIONS
OSF PREPRINTS
OSF
OSF INSTITUTIONS

Landscape alignment

Research Lifecycle

NASEM's vision for repositories

Open Science By Design

NIH Generalist Repository

NIH

Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

BMIC Home | CDE Resource Portal

Home > BMIC Home > NIH Data Repositories

Generalist Repositories

While NIH encourages the use of domain-specific repositories where possible, such repositories are not available for all datasets. When investigators cannot locate a repository for their discipline or the type of data they generate, a generalist reposi
accept data regardless of data type, fo
specific generalist repository and the l
generalist repositories.

- Dataverse
- Dryad
- Figshare
- Mendeley Data
- Open Science Framework
- Vivli
- Zenodo

Framework Foundation: Three Data States

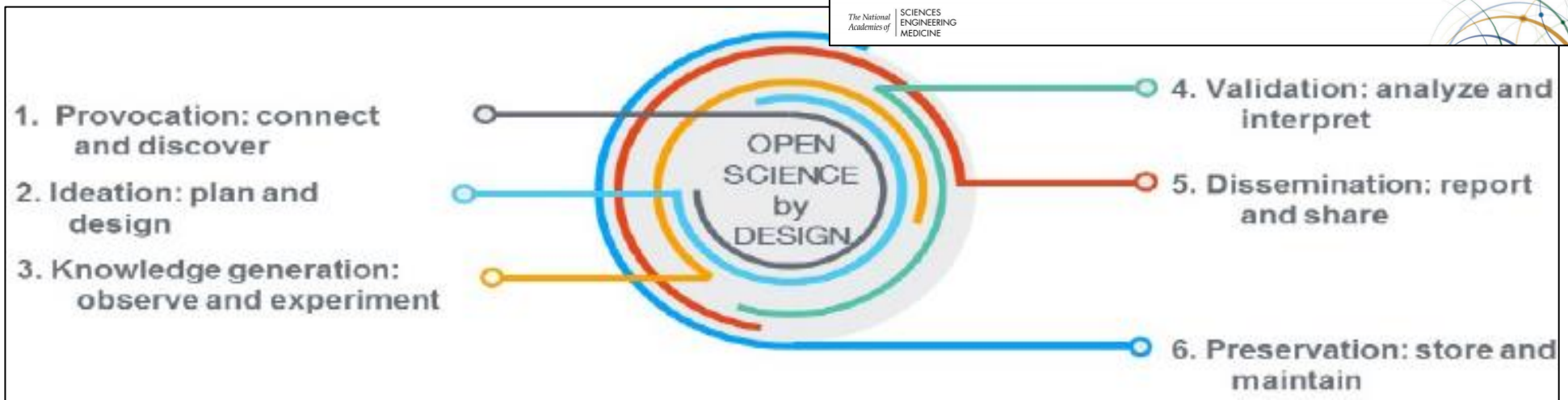
State 1: **Primary research**/data management environment; data are captured and analyzed

State 2: **Active repository** and platform; data may be acquired, curated, aggregated, accessed, and analyzed

State 3: **Long-term preservation** platform

(Box 2.1 in text)

The National Academies of SCIENCES ENGINEERING MEDICINE

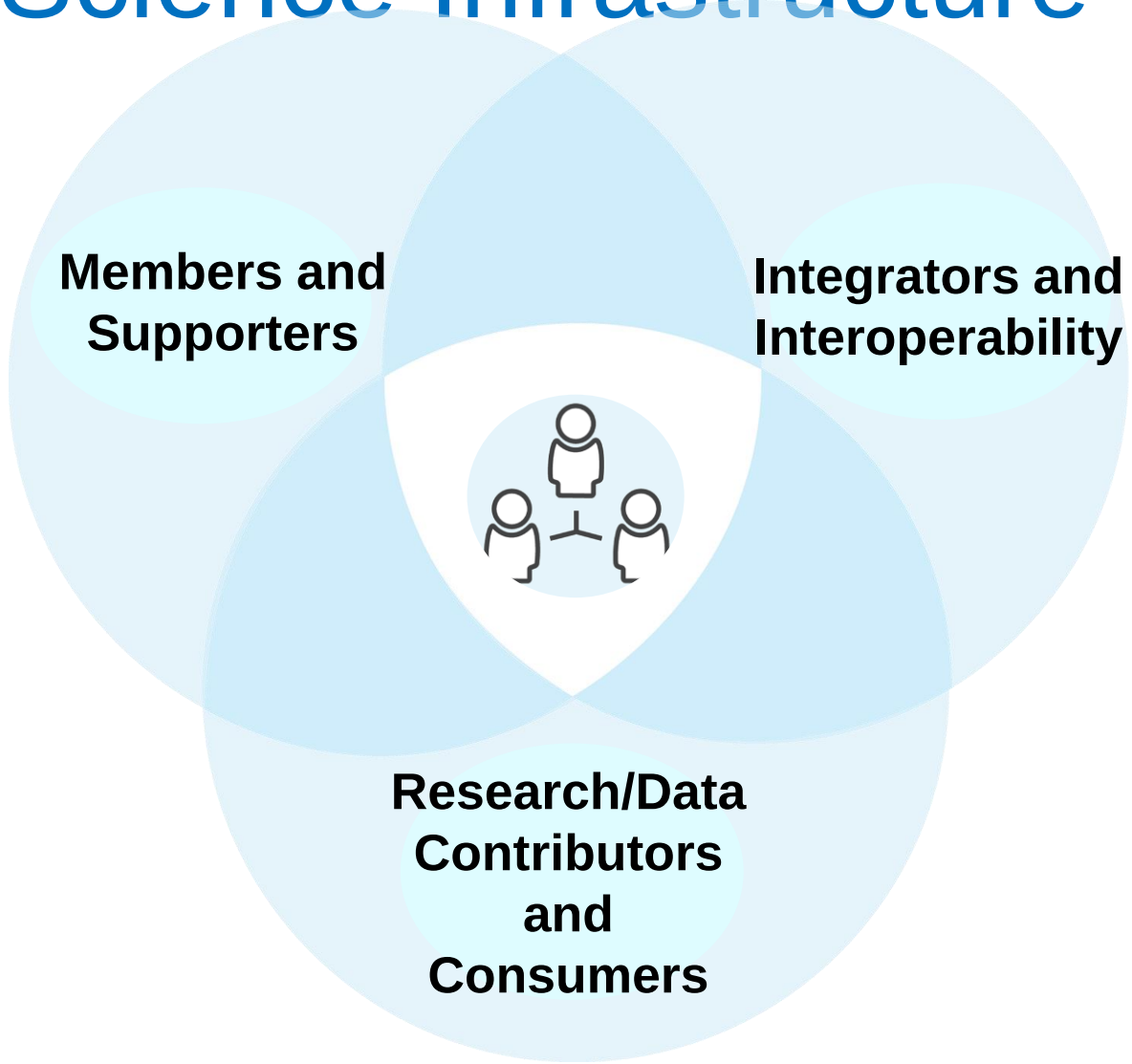


Partners in Open Science Infrastructure

How do we develop and maintain an ecosystem that can respond to many research community needs? Not alone!

We are constantly listening to and adapting for our 350,000 users and dozens of institutional partners. They in turn contributed 18 thousand preprints and 1 million projects that were downloaded over 23 million times in 2020 alone.

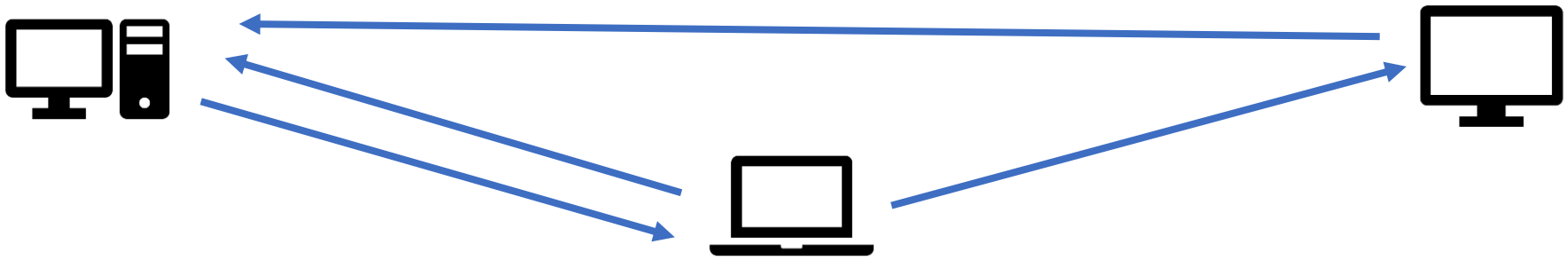
But key to the vision of OSF is to not create features that other services are already providing for researchers, so we seek integration at every opportunity!



Integrators and Interoperability

OSF maintains a free and ***open API, an Application Programming Interface,*** *that can be used to extend OSF capabilities into other custom software development projects.*

APIs enable different systems to “talk” to one another, exchanging information that would be otherwise inaccessible.



https://www.cos.io/communities/software-developers

Integrators and Interoperability

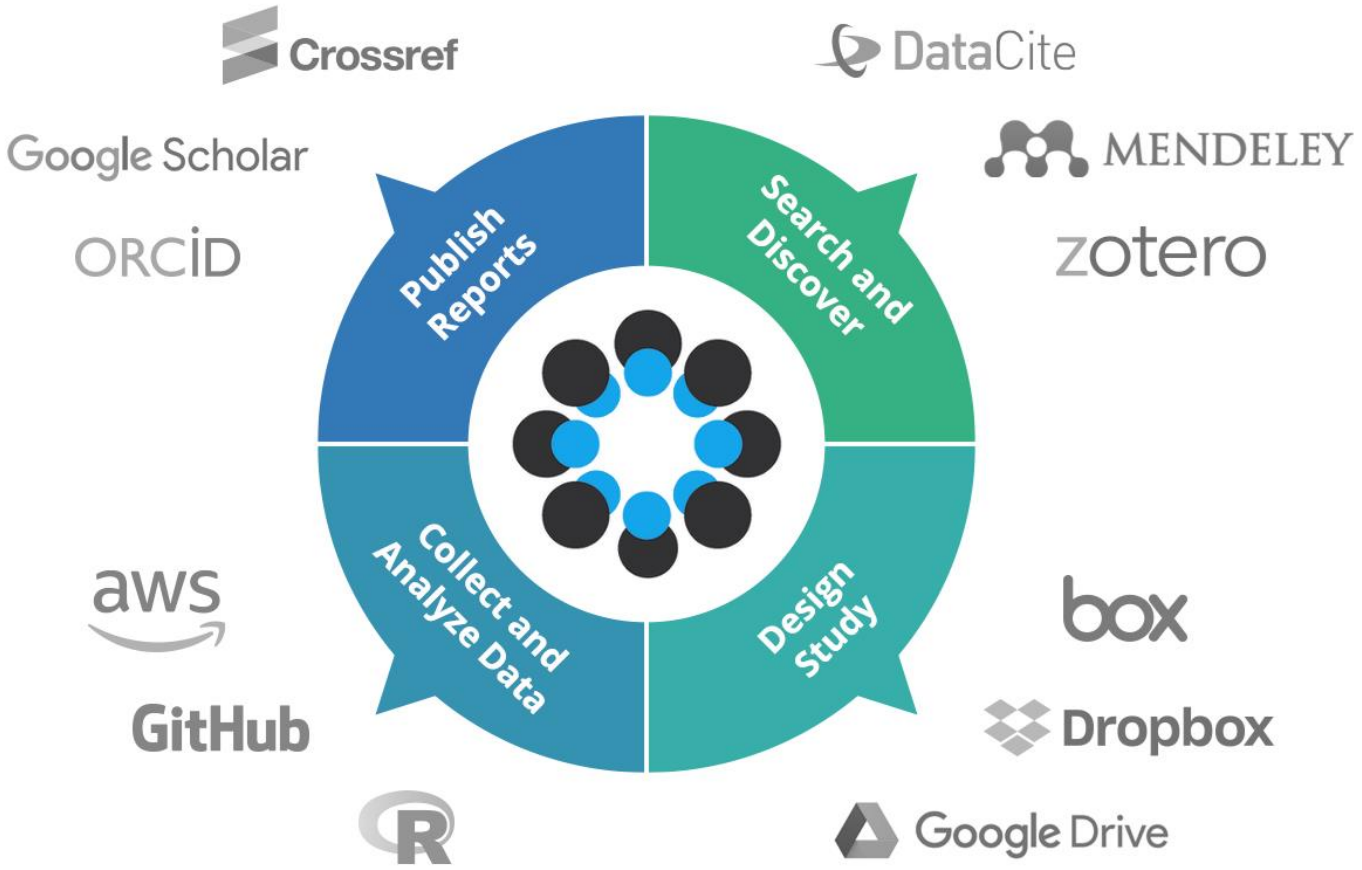
We want to enable any tool to connect with OSF users and features.
For example…

- osfclient – Command-line client for uploading and downloading files to and from OSF
- osfr – R interface to the OSF
- PresQT – An open-source tool suite with RESTful services to improve preservation and re-use of research data and software
- (New!) protocols.io – Move protocols and documents to and from OSF
- And more!

https://www.cos.io/communities/software-developers

Integrators and Interoperability

We also want to meet researchers where they are by enabling integrations with other tools and services that they utilize at each stage of the research lifecycle. This allows a user can store their data or citations in the places that they already use and trust, but still connect them with their OSF projects and collaborators without any duplication of effort.



What now? Integration frenzy!

COS and the OSF partner with research communities in many ways, including the development of add-ons and other integrations. I want to chat with anyone who has an interest in integrating their tools, or have a potential integration with tools that their communities value and use!

Things that are coming soon:

- An OSF member organizations has supported the enhancement of OneDrive integration with write, copy, and move features, as well as connections to institutional OneDrive accounts.
- Local and even regional/consortial repository integrations
- Another OSF Institutions member developing and merging more storage and even computing resource add-ons
- And more partnerships to come!

https://www.cos.io/products/osf-institutions

OSF for Orgs

Community operated spaces are customized to meet the standards set by the stakeholders in your community. Reflect the high expectations of transparency and rigor of your research area.

View and enable open policy compliance

Track outputs and provide an interface where your researchers can quickly adopt the rigorous workflows required by your open policies.

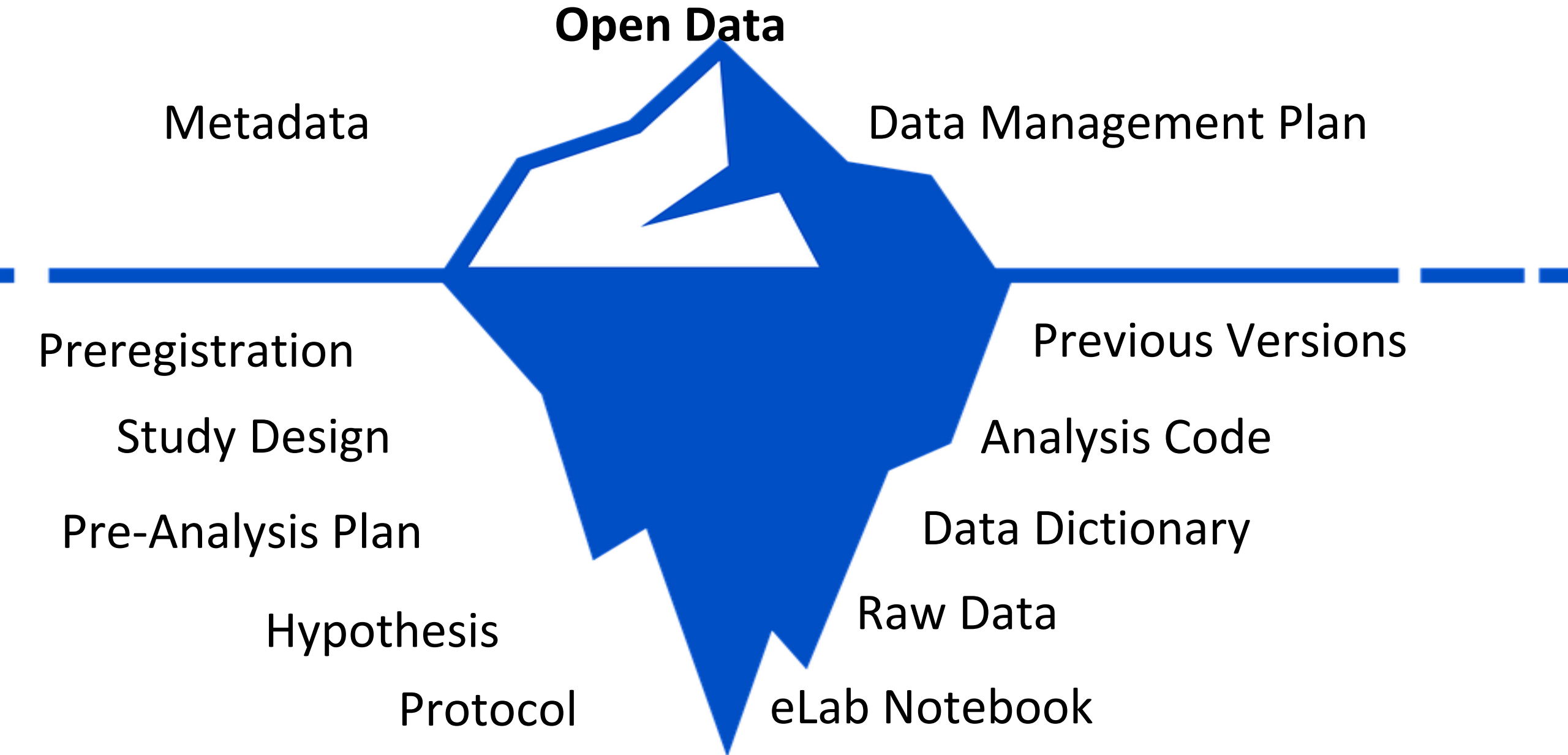
Gather open research outputs

Provide a community repository for exploring open data, preregistrations, and other transparent materials in your field.

Demonstrate rigor and transparency

Cultivate new norms for open sharing and collaboration among your community.

Anatomy of the Open Data Iceberg



Now what?

Let’s look at an example:

https://test.osf.io/registries/maglab/discover

HOW CAN WE STAY INVOLVED?

COS and the OSF partner with research communities in many ways, which is why we dedicate our resources to making tools and services that respond to your needs. Let's keep the conversation going!

- Integrate with OSF
  - OSF API
  - OSF Institutions
- Attend Future Events
  - Watch for new events here
- Have something that we might be able to collaborate on? Let's talk!
  - eric@cos.io