Team Introduction

Our team combines expertise in data standards development, neurophysiology research, and scientific software engineering to address the challenge of incentivizing and measuring effective data sharing.

Benjamin K. Dichter, Team Captain - *Founder & CEO, CatalystNeuro*
Dr. Dichter received a Ph.D. in bioengineering from UC Berkeley and is the founder of CatalystNeuro, a software consulting company facilitating collaboration in neuroscience. Dr. Dichter is a co-PI for the Neurodata Without Borders (NWB) project, which is a data standard for neurophysiology recognized with an R&D 100 Award in 2019. His company, CatalystNeuro, is also part of the development team for DANDI, an archive for neurophysiology data, where they focus on facilitating data contributions and reuse. Through the support of foundations such as the Michael J Fox, Simons and Kavli Foundations, CatalystNeuro has engaged in projects with over 50 neurophysiology labs, building custom software to standardize and/or publish their data. Dr. Dichter is also the lead organizer of NeuroDataReHack, a week-long workshop that trains neuroscientists in the reuse of open datasets. NeuroDataReHack has been running for 4 years and has trained over 100 neuroscientists in the reuse of open data.

Ryan Ly - *Scientific Data Engineer, Lawrence Berkeley National Laboratory*
Dr. Ly received a PhD in neuroscience from Princeton and is a neuroscientist and software engineer with expertise in developing open-source, community-driven scientific software and data infrastructure. As the technical lead of the NWB data standardization project, Dr. Ly has developed sustainable, user-friendly tools to help researchers convert their data to standardized formats without requiring extensive programming expertise. His experience in making data standards accessible to researchers with varying technical backgrounds will be invaluable for developing tools and metrics that can be broadly applied across scientific domains.

Stephanie Prince - *Scientific Data Engineer, Lawrence Berkeley National Laboratory*
Dr. Prince received a PhD in neuroscience from Emory University. She is a neuroscientist and software engineer with experience developing open-source scientific software tools, leading biomedical research projects, and engaging with the research community to promote reproducible and open science. As part of the NWB team, Dr. Prince has organized several workshops to facilitate data standardization, reuse, and sharing in neurophysiology. Dr. Prince's experience generating large-scale datasets and working with data generators to standardize their data gives her unique insight into the challenges researchers face when sharing data and the incentives that would motivate them to share data more effectively.

Oliver Rübel - *Staff Scientist, Lawrence Berkeley National Laboratory*
Dr. Rübel received a PhD in Computer Science from the University of Kaiserslautern and is the lead of the NIH grant for integration and dissemination of NWB as an official neurophysiology data standard for the BRAIN Initiative. His research has focused on visualization, analysis, and data management for large-scale data. Dr. Rübel has led the development of multiple data standardization frameworks and has received an R&D 100 Award for his work on NWB. His experience developing standards for complex scientific data and facilitating their adoption across research communities will help ensure the proposed S-index captures the quality and impact of shared data.

Complementary Expertise. Our team's complementary expertise uniquely qualifies us to develop the S-index. All team members have contributed to the development and implementation of the NWB data standard, giving them insight into what makes data truly reusable. We have experience as researchers generating and analyzing complex scientific datasets, providing first hand understanding of both the challenges and benefits of data sharing. Our extensive experience working with research communities to facilitate data sharing, has given us insight into the incentives and barriers of data sharing. Additionally, all team members have experience developing software tools to support data standardization and sharing, providing the technical expertise needed to develop implementable metrics.

The FAIR-Enhanced Data Sharing Index (S-index): Enhancing Data Sharing Through Comprehensive Evaluation of Reuse and Rigor

Introduction. Data sharing advances scientific discovery, enhances research reproducibility, and accelerates biomedical research. However, *there is currently a critical lack of incentives for scientists to invest the time and effort to effectively share their data.* The NIH Data Management and Sharing Policy [1] is a crucial step towards increased data dissemination, but lacks robust metrics to evaluate compliance and impact. As a result, the scientific community is still missing a mechanism to recognize the valuable scientific contribution of high-quality data sharing. To address this gap, we propose the *FAIR-Enhanced Data Sharing Index (S-index)* that combines data reuse impact via citations with data sharing rigor via automated FAIR assessment. Our approach assigns an index not only to individual researchers, but also to journals and institutions, creating a multi-stakeholder system that incentivizes data sharing across the scientific enterprise. Our S-index thus provides a quantifiable measure that will reward researchers for their data quality and impact and has the potential to motivate both individuals and institutions to invest their time into effective data sharing to advance open science.

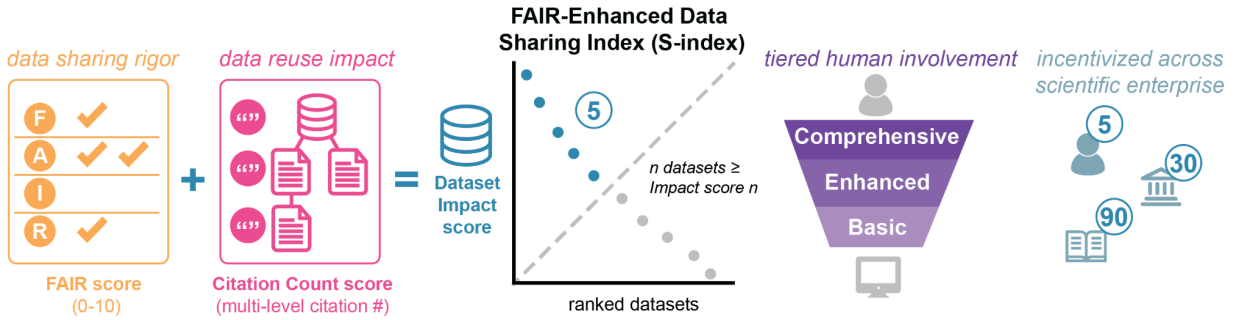

Figure 1. FAIR-Enhanced Data Sharing Index (S-index) innovations. *A) The Dataset Impact Score captures data sharing rigor and data reuse impact. B) Tiered levels of automation ensure implementation is feasible. C) Multi-level application incentivizes both individuals and institutions.*

Conceptual Innovation

Unique aspects and differences from existing metrics. Previously proposed metrics like the data-index [2] reward data impact but fail to capture data quality. Our S-index builds upon the data-index and introduces three key innovations: (1) dual-component evaluation combining data reuse impact through citations with data sharing rigor through automated FAIR assessment; (2) tiered implementation with increasing sophistication from automated assessment to human review; and (3) multi-level application extending beyond individual researchers to journals and institutions (Fig. 1). Unlike the h-index (publication impact) or data-index (reuse focus), the S-index recognizes rigorous data sharing in addition to citation count, combining approaches into a single comprehensive metric.

Methods and algorithms. The S-index is calculated as follows:

1. For each dataset, determine the FAIR-Score (0-10) to assess data sharing rigor
2. For each dataset, calculate the Citation Count Score to assess data reuse impact
3. A dataset's Impact Score is its FAIR-Score + Citation Count Score
4. A researcher's S-index is the largest number n where they have at least n datasets with an Impact Score of at least n (similar to the h-index)

The S-index of a journal or institution is calculated the same way as that of a researcher, aggregating across all datasets associated with that institution or journal. This approach creates a cohesive ecosystem of metrics that incentivize data sharing at multiple levels.

The S-index incorporates a comprehensive FAIR [3] assessment rubric (Table S1) that evaluates datasets across four key dimensions, each worth up to 2.5 points for a total maximum score of 10. Findability is

assessed through persistent identifier verification, metadata richness, and searchable registry inclusion. Accessibility evaluates standardized protocol usage, authentication appropriateness, and metadata persistence. Interoperability examines knowledge representation standards, vocabulary usage, and qualified references to other data. Reusability assesses license clarity, provenance documentation, and community standards compliance. This balanced scoring system ensures datasets are evaluated not just on technical accessibility but on their complete readiness for reuse. This FAIR-Score enhances the evaluation of data sharing by providing immediate recognition for high-quality data sharing practices even before citation metrics accumulate.

The Citation Count Score, based on the data-index citation metric in [2], captures both direct and downstream reuse of a dataset. It is calculated by summing direct citations of the dataset, citations of associated papers describing the data, and citations of papers that reuse the dataset. As a result, datasets that are highly cited will achieve a high score, as will datasets that are reused in a highly cited paper or in many moderately cited papers. This approach reflects both the dataset's immediate impact and its broader contribution to subsequent research, capturing the cumulative effect of data reuse across the research ecosystem.

Table S2 illustrates how our metric will be calculated. In this example, the researcher has an S-index of 7, as they have 7 datasets with an Impact Score of at least 7. Without the FAIR-Score component, their data-index would only be 5 (5 datasets with at least 5 data-index citations). This example demonstrates how our S-index provides recognition for high-quality datasets (Papers 7, 9, and 10) that have few or no citations yet, rewarding rigorous data sharing practices even before the datasets are widely reused.

Feasibility

Implementation strategy. We propose three implementation tiers:

(1) Basic - fully automated system using existing citation tracking APIs, repository access, and ML models
(2) Enhanced - annual recalculation with validation checks and version tracking
(3) Comprehensive - human review with dispute resolution

The Basic tier would analyze public repository metadata to extract FAIR indicators and use citation APIs to track dataset usage. This implementation leverages existing infrastructure including DataCite and CrossRef for citation tracking, repository APIs from platforms like Figshare and Zenodo, and machine learning models for automated FAIR assessment. The Enhanced tier would add scheduled accessibility checks to verify persistent availability and interface with version control systems to accommodate dataset evolution. The Comprehensive tier would implement a workflow where researchers could appeal automated assessments with supporting evidence for human review.

Interactive dashboards will be developed for different stakeholders with customized views for researchers, university departments, journals, and funders. These dashboards will show the FAIR-Score and Citation Count Score for each dataset, providing transparency into how S-Indexes are calculated across individuals, institutions, and journals.

Required resources. There are currently 148 specialist NIH-supported Scientific Data Repositories and 9 additional generalist repositories identified by the NIH [4,5]. Required resources to implement and scale our approach include software infrastructure (web crawlers, NLP models, validation tools), data resources (citation databases, repository APIs), and human resources (engineers, domain experts). Most technologies already exist as open-source or accessible APIs, making implementation realistically attainable with modest investment. Development can build on existing FAIR assessment tools like the FAIR Evaluation Services [6], along with established citation tracking infrastructures. We estimate the implementation of the Basic tier would take two full-time engineers (FTE) two years to develop and 0.5 FTE for ongoing maintenance (e.g., adapting to changing APIs and new repositories). Implementation of the Enhanced tier would require another 0.5 ongoing FTE to develop and maintain automated processes,

and the Comprehensive tier would require an additional 1 ongoing FTE to manage community engagement. This tiered approach allows us to maintain sustainability through fluctuations in funding.

Stakeholder acceptance and alignment with current needs. Researchers, particularly early-career scientists, can benefit from recognition of data sharing contributions in hiring and promotion decisions. Institutions can demonstrate commitment to open science principles through high institutional S-indices, potentially attracting funding and talent. Funding agencies can use S-indices to evaluate compliance with data management and sharing policies and assess return on investment. Journals can differentiate themselves through their commitment to data accessibility, as measured by their journal S-index.

Impact

Incentivizing data sharing and behavioral changes. The S-index incentivizes data sharing by acknowledging high-quality preparation effort before citations accumulate, providing metrics for career advancement and funding proposals, while creating multi-level metrics for journals and institutions. We expect its adoption will drive researchers to improve dataset documentation with better metadata, adopt community standards over proprietary formats, and share data at publication rather than upon request. Institutions will be motivated to provide dedicated resources and training for effective data sharing to maximize their S-index. This ecosystem-wide approach encourages both individuals and organizations to invest in practices that accelerate scientific progress through better data accessibility.

Identifying widely used and impactful data resources. The dataset Impact Score underlying the S-Index builds upon data-index citations, which are designed to capture datasets that are not only widely used, but that have been also used in highly cited papers. By combining this approach with FAIR quality assessment, it ensures recognition for datasets with high current impact as well as those with strong potential for future reuse. The Impact Score thus captures the full downstream impact of valuable data resources while providing appropriate recognition to well-structured datasets even before widespread citation. Public leaderboards will highlight datasets with the highest Impact Scores across a research area, filterable by date, journal, institution, and researcher. Honors can be awarded for datasets with the highest scores to further encourage researchers and institutions to publish valuable, reusable datasets.

Conclusion

The *FAIR-Enhanced Data Sharing Index (S-index) will be a significant advancement in the incentivization of data sharing* by 1) rewarding exemplary researchers immediately (data sharing rigor) and in the long-term (data reuse impact), 2) providing a tiered approach to enable progressive adoption using existing infrastructure and 3) encouraging the entire scientific ecosystem to change its behavior to more effectively share and reuse data.

Advancements and sustainability measures. To ensure that the S-index remains an effective metric that aligns with the state of the art in open data, our implementation strategy includes forming a governing body of key stakeholders. This S-index consortium will provide scientific oversight and be responsible for refining the FAIR rubric, identifying new repositories, and mediating disputes about scores. As the research community begins to adopt the S-index and integrate it with existing systems, this governance model will support continuous improvement based on community feedback.

Future directions. Future enhancements include ORCID integration automating calculation through researcher identifiers, enhanced assessment algorithms, and educational resources for improved data sharing practices. These enhancements will improve the ease of adoption and ensure the S-index extends beyond early adopters to the entire scientific community.

The S-index will drive cultural change toward more open, collaborative science, maximizing data sharing value for scientific progress and public benefit. By providing concrete metrics that recognize and reward effective data sharing, we can help transform scientific practice to prioritize transparency, reproducibility, and collaboration in ways that accelerate discovery and maximize return on research investments.

Citations

[1] *Data Management and sharing policy*. National Institutes of Health. https://sharing.nih.gov/data-management-and-sharing-policy

[2] Hood, A. S. and Sutherland, W. J., 2021. The data-index: An author-level metric that values impactful data and incentivizes data sharing. *Ecology and Evolution*, 11(21), pp.14344-14350. doi: 10.1002/ece3.8126

[3] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. doi: 10.1038/sdata.2016.18

[4] *Repositories for Sharing Scientific Data.* National Institutes of Health. https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data

[5] *Generalist Repositories.* National Institutes of Health. https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/generalist-repositories

[6] *FAIR Evaluation Services.* FAIRSharing.org. https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/

Supplemental materials

FAIR Assessment Rubric Table

FAIR Component	Assessment Criteria	Scoring Levels
Findability (0-2.5 points)	Persistent Identifier (0-1 point)	0: No persistent identifier 0.5: Non-standard identifier without guaranteed persistence 1: Standard persistent identifier (DOI, Handle, ARK) that properly resolves
	Metadata Richness (0-1 point)	0: Minimal or absent metadata 0.5: Basic metadata (title, authors, date) but limited descriptive content 1: Comprehensive metadata including detailed description, keywords, related identifiers
	Searchable Registry Inclusion (0-0.5 points)	0: Not indexed in any searchable registry 0.25: Indexed in institutional/local repository only 0.5: Indexed in domain-specific or general-purpose registry with search functionality
Accessibility (0-2.5 points)	Standardized Protocol (0-1 point)	0: No standard retrieval protocol 0.5: Requires proprietary or specialized protocols 1: Accessible via standard open protocol (HTTP/S, FTP)

	Authentication Appropriate (0-0.75 points)	0: Inaccessible or no clear access mechanism 0.25: Unclear or overly restrictive authentication 0.5: Clear authentication process appropriate for data sensitivity 0.75: Open access with appropriate licenses for non-sensitive data
	Metadata Persistence (0-0.75 points)	0: Metadata and data access linked without separate persistence 0.5: Some metadata persists independently 0.75: Complete metadata remains accessible independent of data availability
Interoperability (0-2.5 points)	Knowledge Representation (0-1 point)	0: Proprietary or non-standard formats 0.5: Open but uncommon formats 1: Community-standard formats for the domain
	Vocabulary Usage (0-0.75 points)	0: No controlled vocabulary or ontology references 0.25: Some standard terms but inconsistent application 0.5: Consistent use of domain-specific controlled vocabularies 0.75: Comprehensive ontology references with semantic relationships
	Qualified References (0-0.75 points)	0: No connections to other datasets or metadata 0.5: Basic references without relationship qualification 0.75: Qualified references specifying relationship types
Reusability (0-2.5 points)	License Clarity (0-1 point)	0: No license information 0.5: Ambiguous or non-standard license 1: Clear standard license (e.g., Creative Commons)
	Provenance Documentation (0-0.75 points)	0: No provenance information 0.25: Minimal information about data origin 0.5: Basic methodology description 0.75: Comprehensive documentation of data collection, processing, and transformation
	Community Standards Compliance (0-0.75 points)	0: No adherence to community standards 0.25: Partial compliance with some standards 0.5: Substantial compliance with domain standards 0.75: Full compliance with all relevant community standards

Table S1: Tentative FAIR rubric for computing the FAIR-Score component of a dataset's Impact Score.

Paper	Original Data	FAIR Score	Citations	Data Used in Other Papers	Citations of Those Papers	Citation Count Score	Dataset Impact Score
Paper 1	✓	8.5	18	✓	21	39	47.5
Paper 2	✓	9	11	✓	13	24	33
Paper 3	✓	7	8	✓	15	23	30
Paper 4	✓	6.25	14	X	-	14	20.25
Paper 5	✓	4	1	✓	7	8	12
Paper 6	✓	5.5	4	X	-	4	9.5
Paper 7	✓	8	1	X	-	1	9
Paper 8	✓	7.5	0	X	-	0	7.5
Paper 9	✓	5	0	X	-	0	5
Paper 10	X	-	4	-	-	-	-

Table S2: Example S-Index calculation, which extends the example data-index calculation in [2]. This researcher has an S-Index of 7 because they have 7 datasets that each have an Impact Score of at least 7.