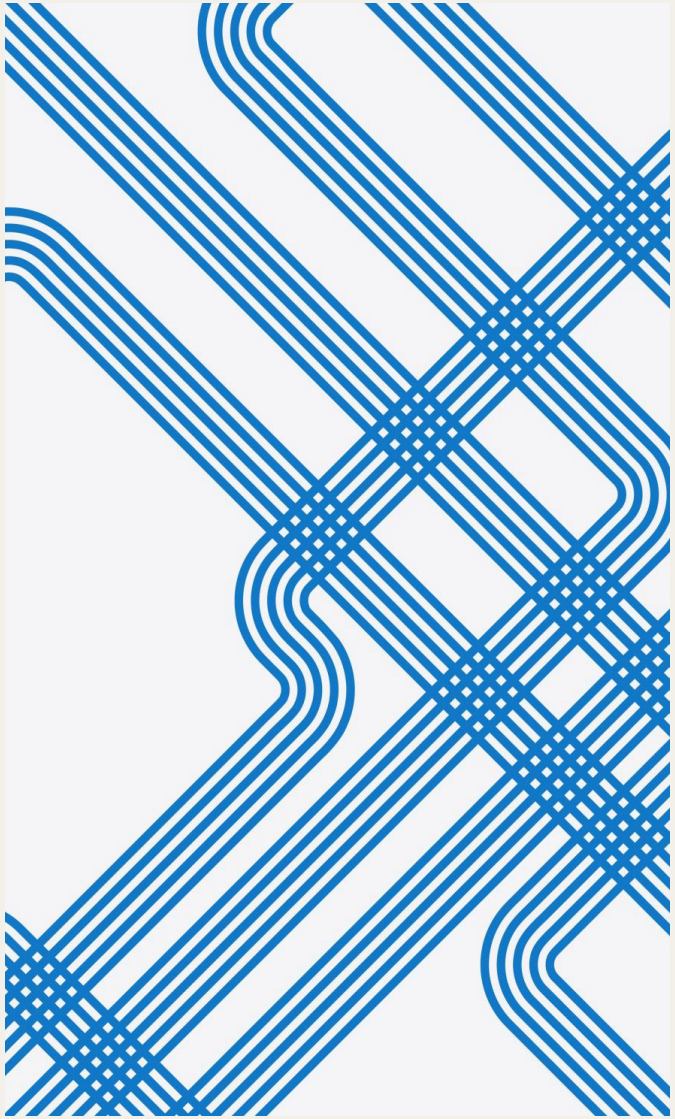
Translating Policies into Open Science and FAIR Data **Practices** in Chemistry:

Incentivizing Researchers and Reducing Barriers



Ye Li

Librarian for Chemistry, Chemical Engineering,
Materials Science and Engineering
MIT Libraries

yel@mit.edu

https://orcid.org/0000-0001-8361-6916

February 22, 2024

NATIONAL ACADEMIES Sciences Engineering Medicine

Future Implications of Open-Access & FAIR Data Practices on Chemistry and Chemical Engineering Publications – A Workshop
Session III: Supporting the Research Community

Federal Funder Requirements

NIH – 2003
https://grants.nih.gov/grants/policy/data_sharing/

NIH – 2023
https://sharing.nih.gov/

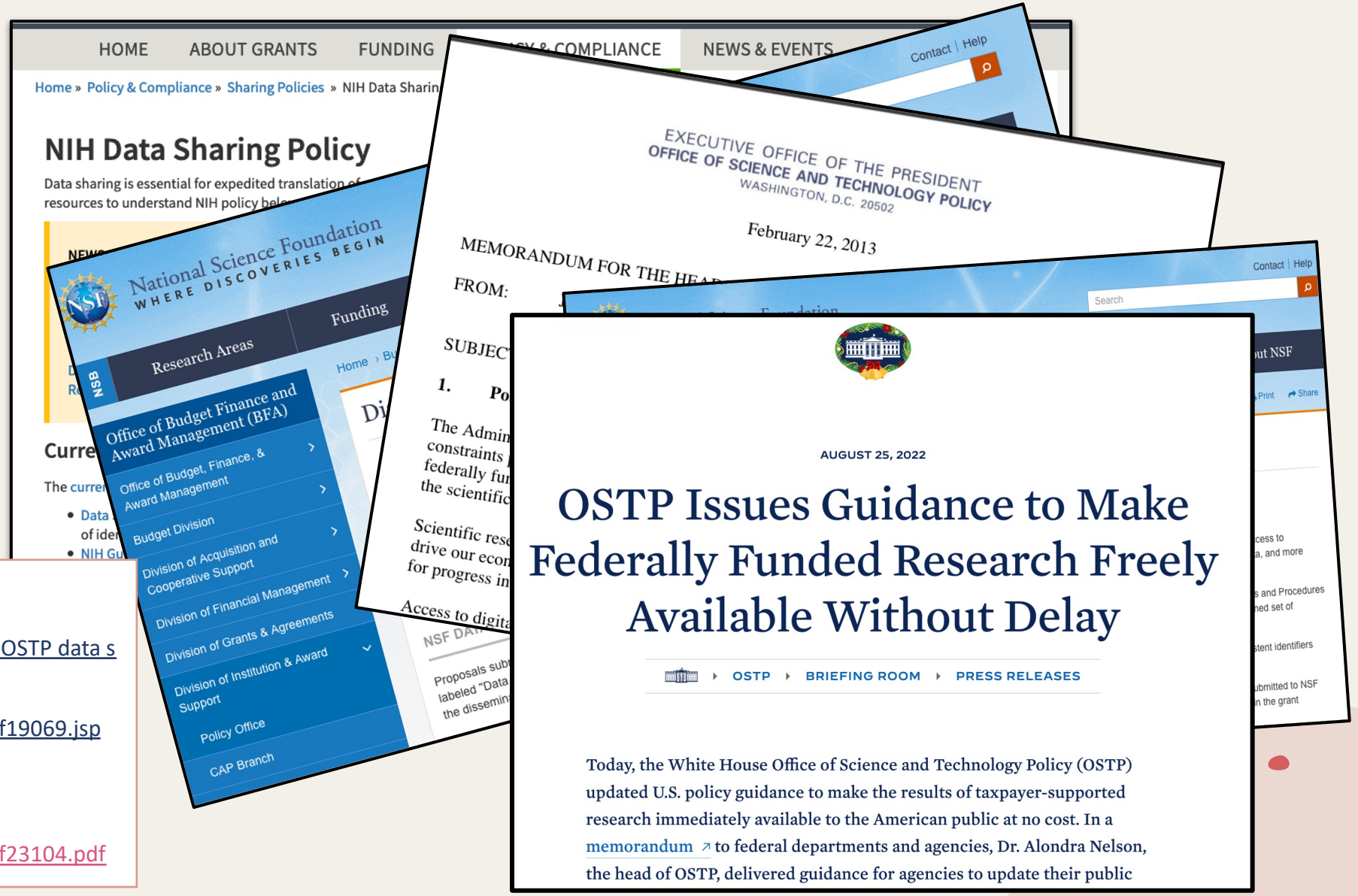
OSTP - 2013
https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

OSTP – 2022
https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

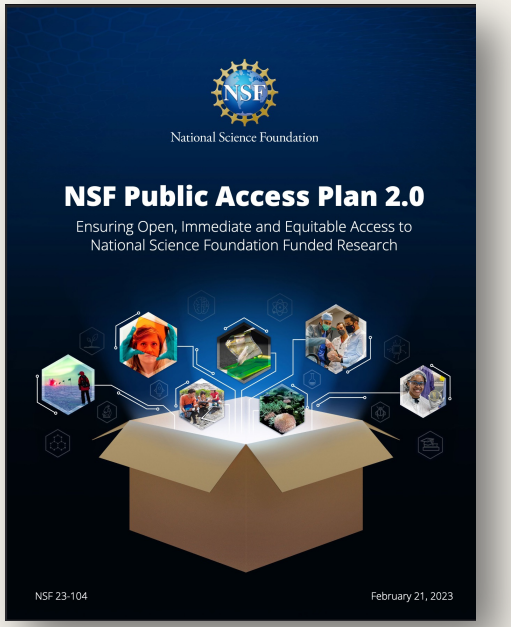
NSF - 2010
https://www.nsf.gov/bfa/dias/policy/dmp.jsp OSTP data s

NSF - 2019
https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp

NSF – 2023
(Public Access Plan 2.0)
https://www.nsf.gov/pubs/2023/nsf23104/nsf23104.pdf



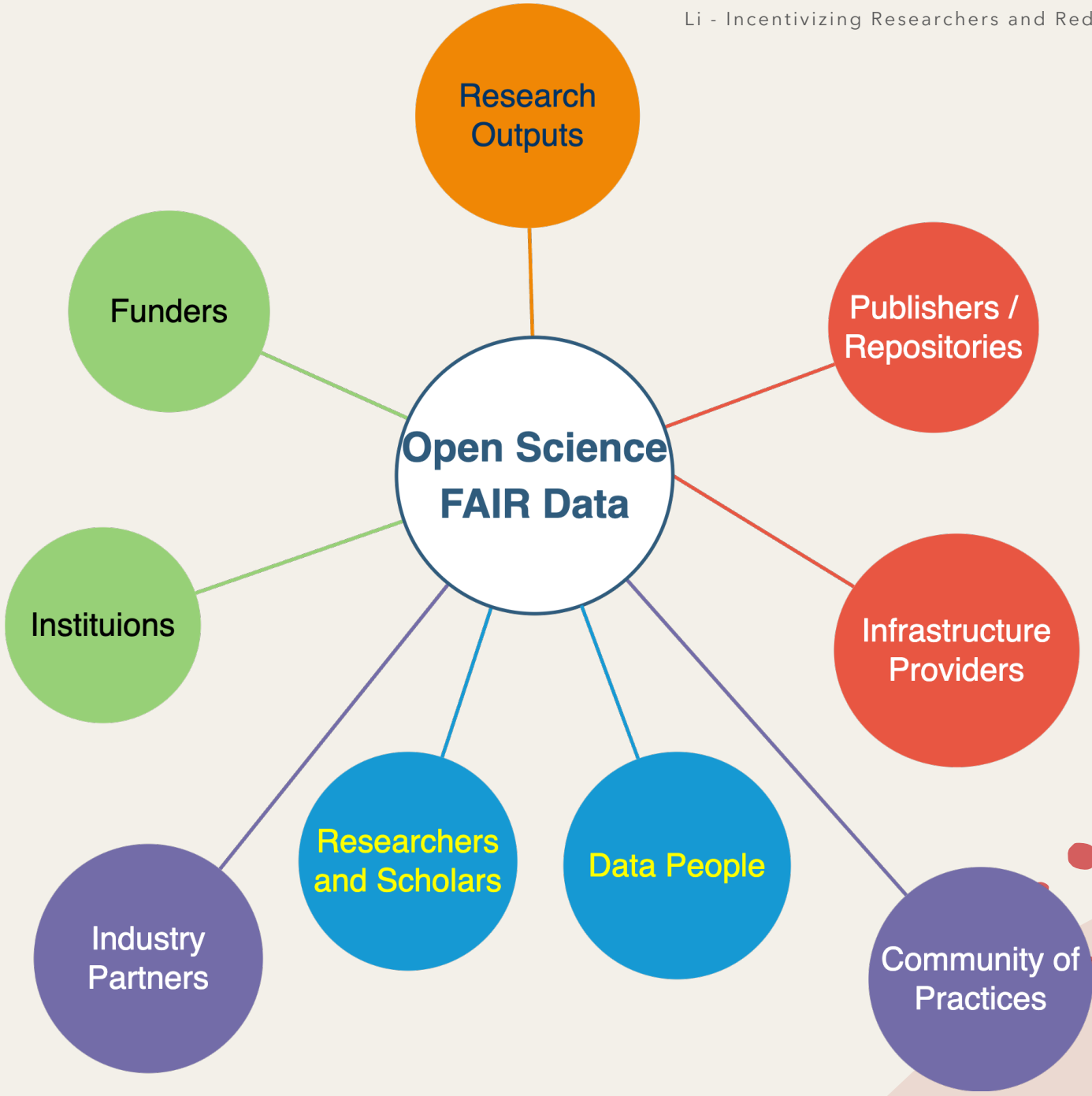
Open, Immediate and equitable access applies to both peer-reviewed publications and scientific data



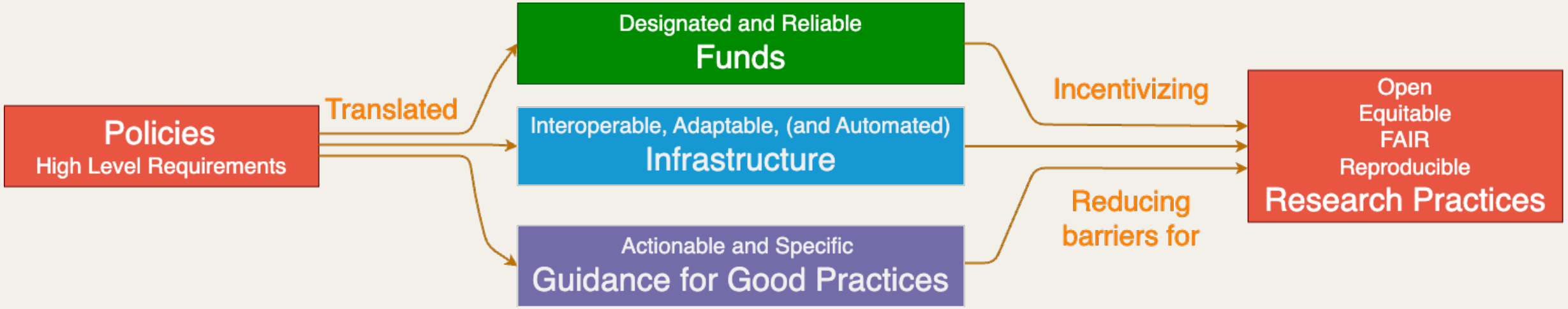
https://www.nsf.gov/pubs/2023/nsf23104/nsf23104.pdf

The broad range of stakeholders for Open Science and FAIR Data brings complexity but also power

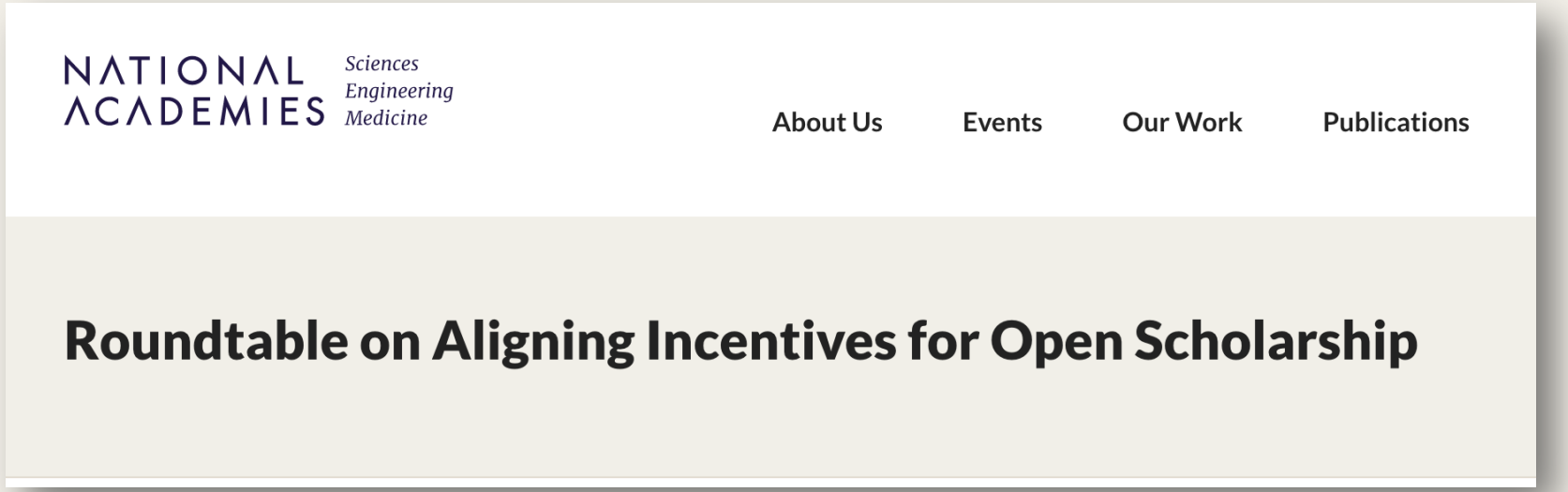
FAIR:
Findable, Accessible, Interoperable, Reusable



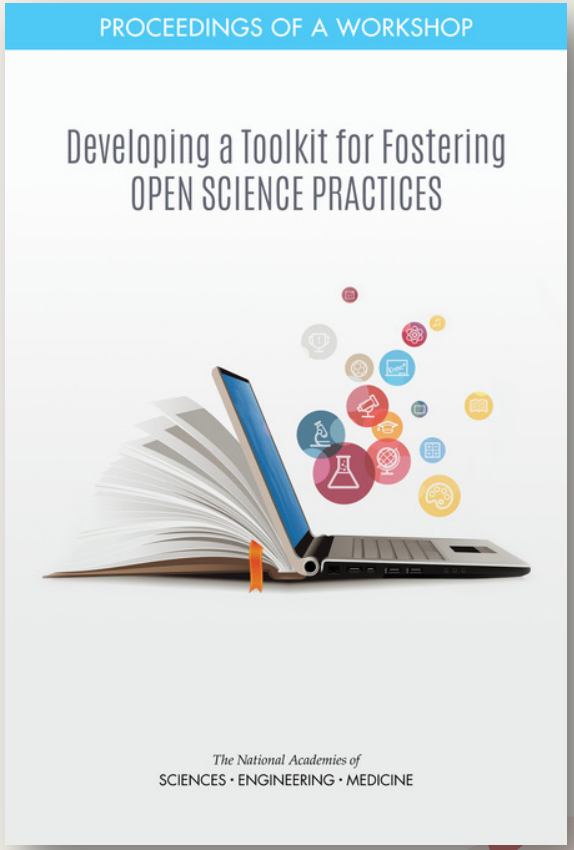
All stakeholders collaborate to enable Open Science and FAIR data practices



Incentives, such as funds, recognition, and promotion, matter for all disciplines



https://www.nationalacademies.org/our-work/roundtable-on-aligning-incentives-for-open-science , Accessed Feb 14, 2024



National Academies of Sciences, Engineering, and Medicine. 2021. Developing a Toolkit for Fostering Open Science Practices: Proceedings of a Workshop. Washington, DC: The National Academies Press. https://doi.org/10.17226/26308.

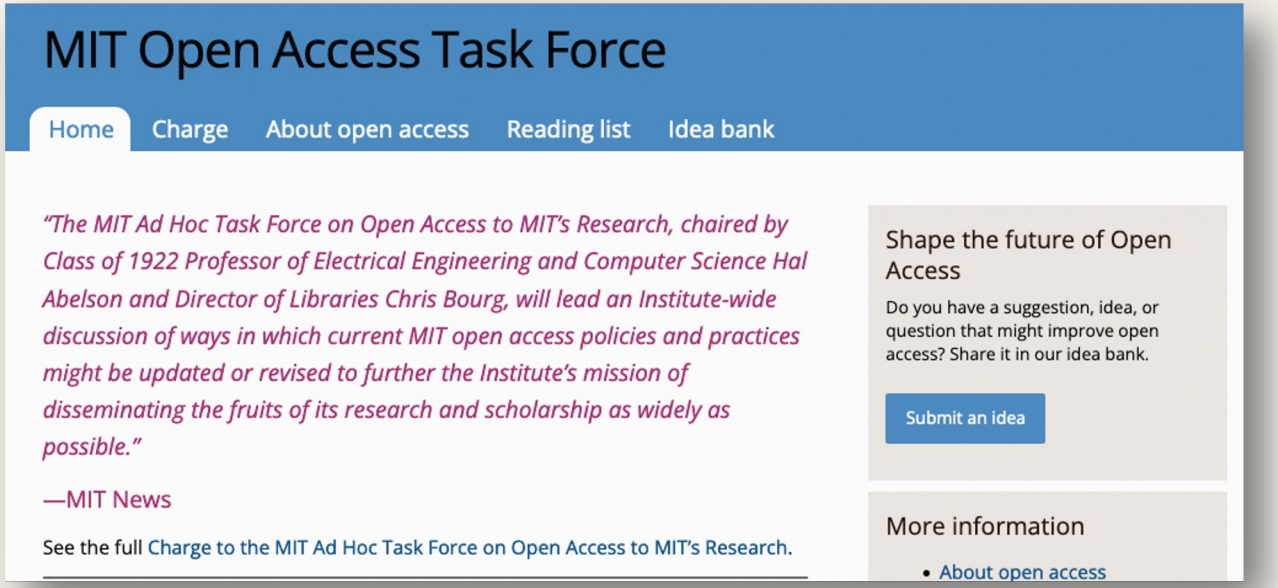
Institutional Policy – Empower and Support

MIT Open Science Policies

- Recommendation 1: All-campus OA policy
- Recommendation 3: Clarify and consolidate thesis holds policies
- Recommendation 6: Develop a data registry
- Recommendation 9: Code and software
- Recommendation 10: DLC plans to encourage open sharing
- Recommendation 12: Institutional advocacy for open science
- Recommendation 14: Form OATF recommendations implementation team

Owing to a variety of reasons, including financial and resource constraints, reordering of Institute priorities, and diverted attention due to COVID-19 response management demands, work on the following recommendations has been delayed:

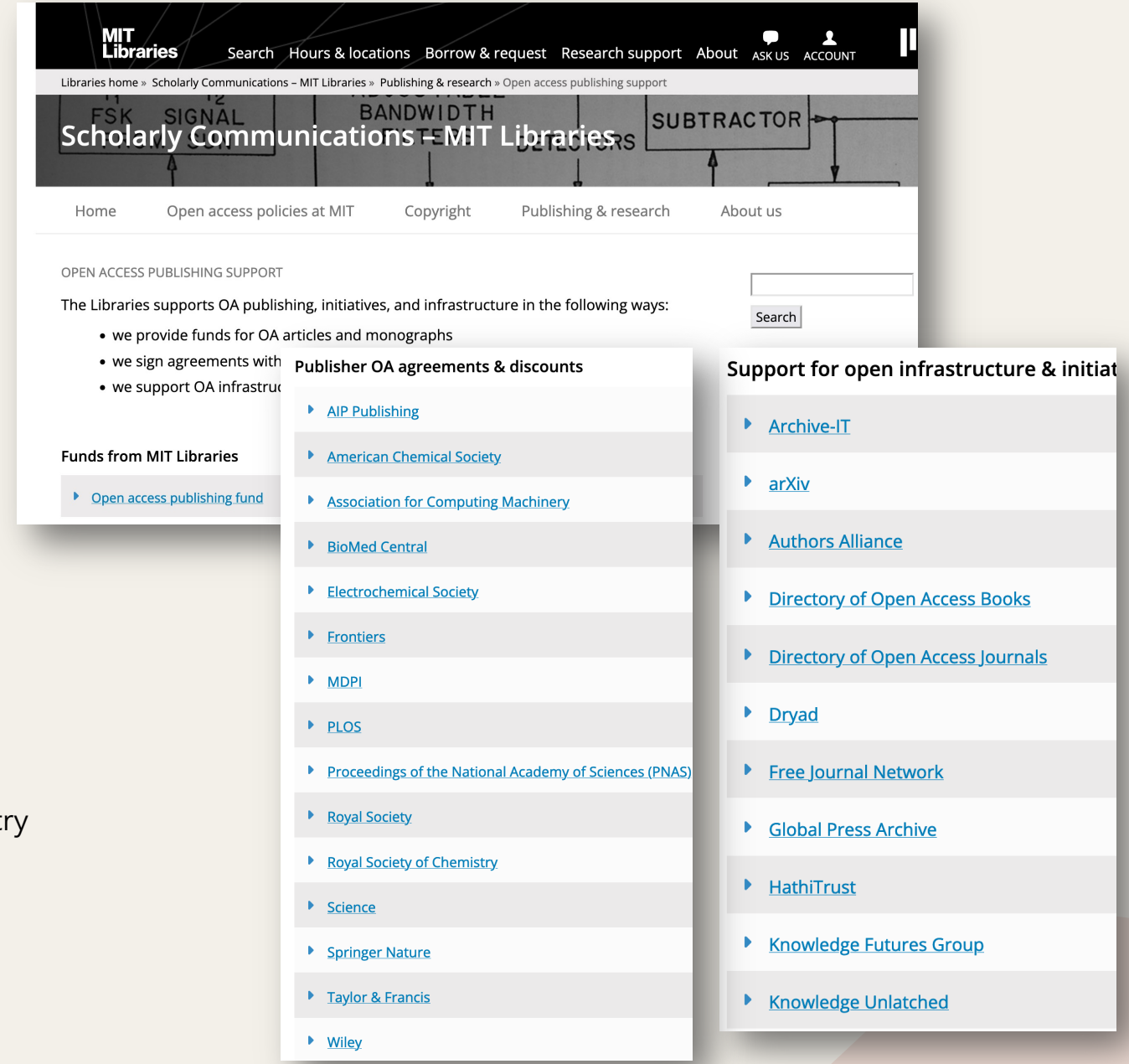
- Recommendation 2: OA policy for monographs
- Recommendation 4: OA requirement for internal MIT grants
- Recommendation 5: Open Data Fund
- Recommendation 7: Funds for OA initiatives
- Recommendation 8: Increase the impact of open educational content
- Recommendation 11: Recognition for data sharing
- Recommendation 13: MIT's International Advisory Committee



https://open-access.mit.edu/

OATF Implementation Team Progress Report. *MIT Open Access Task Force*. (2020) https://doi.org/10.21428/4090026e.884d8726

Libraries provide OA publishing funds/ agreements and invest in open infrastructure & initiatives



Challenges for chemists:

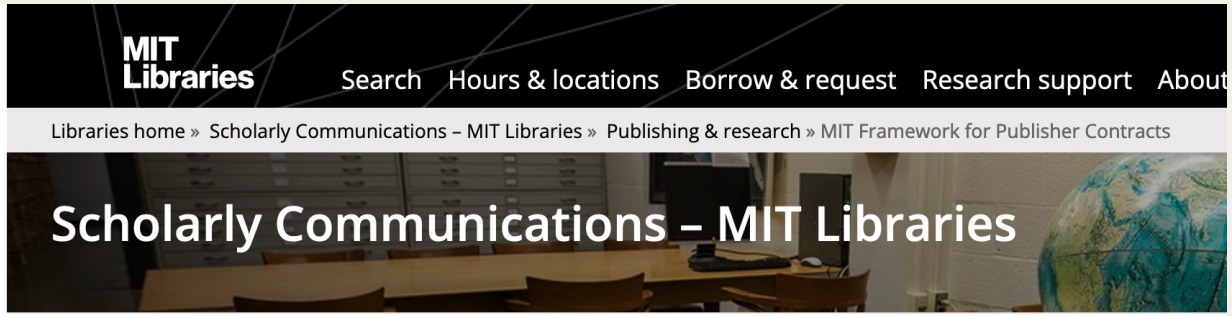
- Funds may not cover the full costs for journal publications in chemistry
- Agreements
  - May not have it with all major publishers in Chemistry
  - **Article Processing Charge (APC-) based models may not be equitable and sustainable**
- Open infrastructure investment may not prioritize discipline specific ones

https://libraries.mit.edu/scholarly/publishing/oa-publishing-support/ , Accessed Feb 14, 2024

MIT Framework for Publisher Contracts may provide a more equitable and sustainable path forward

- Author rights protection
- Value-added services
  - Automated deposit
  - Long-term digital preservation
  - **Computational access and use**
- Transparent and cost-based pricing models

https://libraries.mit.edu/scholarly/publishing/framework/ , Accessed Feb 14, 2024



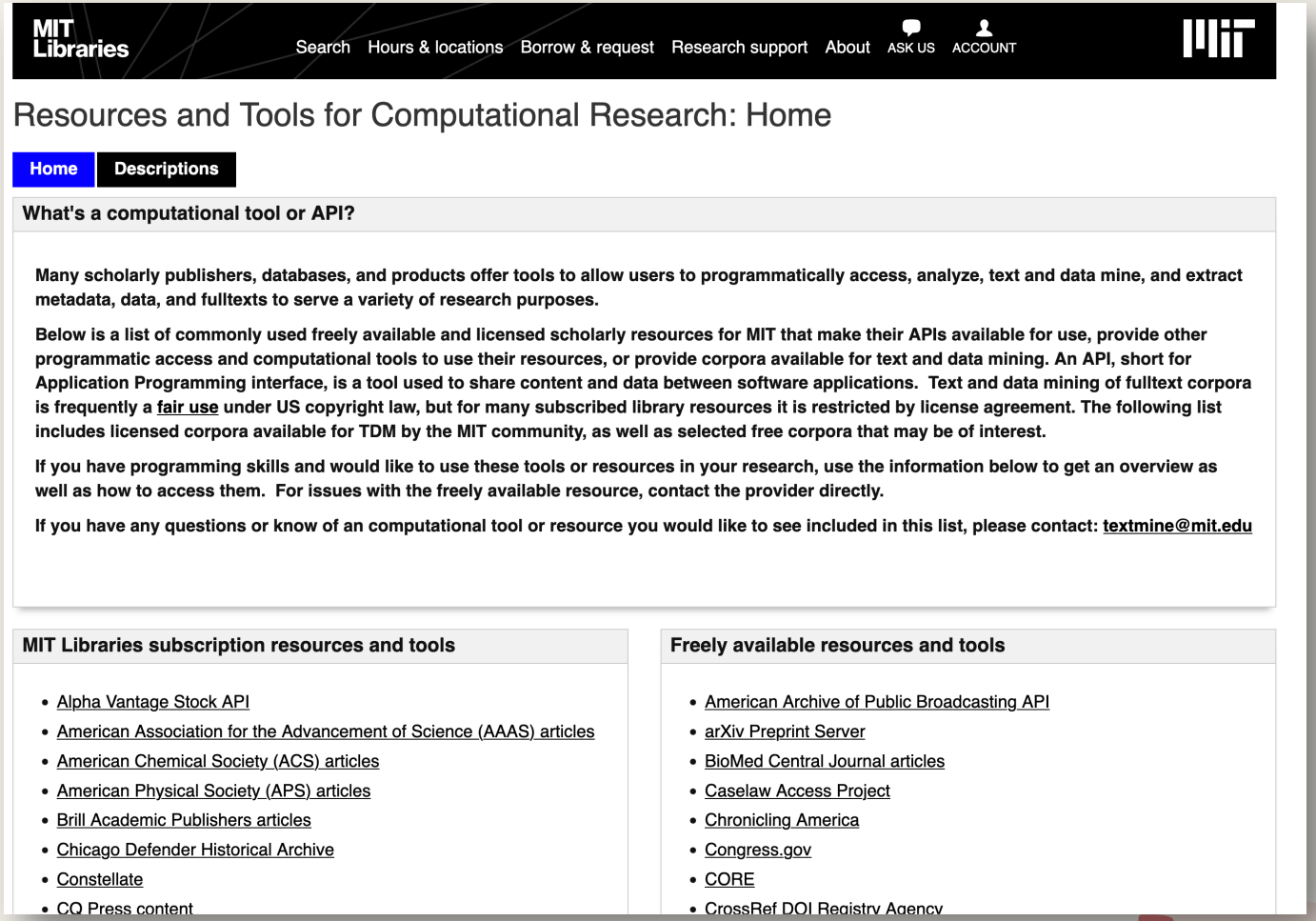
The core principles of an MIT Framework for publisher contracts are:

- No author will be required to waive any institutional or funder open access policy to publish in any of the publisher's journals.
- No author will be required to relinquish copyright, but instead will be provided with options that enable publication while also providing authors with generous reuse rights.
- Publishers will directly deposit scholarly articles in institutional repositories immediately upon publication or will provide tools/mechanisms that facilitate immediate deposit.
- Publishers will provide computational access to subscribed content as a standard part of all contracts, with no restrictions on non-consumptive, computational analysis of the corpus of subscribed content.
- Publishers will ensure the long-term digital preservation and accessibility of their content through participation in trusted digital archives.
- Institutions will pay a fair and sustainable price to publishers for value-added services, based on transparent and cost-based pricing models.

Terms of license agreements enable text and data mining (TDM) and subsequent machine learning and AI studies

Challenges for chemists:

- Terms are nuanced
- Terms and retrieval mechanisms vary among major publishers in chemistry
- Recent trend of some publishers in chemistry adopting a more conservative stance on TDM terms due to concerns on training public-facing AI models



MIT Libraries | Search | Hours & locations | Borrow & request | Research support | About | ASK US | ACCOUNT

Resources and Tools for Computational Research: Home

Home | Descriptions

What's a computational tool or API?

Many scholarly publishers, databases, and products offer tools to allow users to programmatically access, analyze, text and data mine, and extract metadata, data, and fulltexts to serve a variety of research purposes.

Below is a list of commonly used freely available and licensed scholarly resources for MIT that make their APIs available for use, provide other programmatic access and computational tools to use their resources, or provide corpora available for text and data mining. An API, short for Application Programming interface, is a tool used to share content and data between software applications. Text and data mining of fulltext corpora is frequently a fair use under US copyright law, but for many subscribed library resources it is restricted by license agreement. The following list includes licensed corpora available for TDM by the MIT community, as well as selected free corpora that may be of interest.

If you have programming skills and would like to use these tools or resources in your research, use the information below to get an overview as well as how to access them. For issues with the freely available resource, contact the provider directly.

If you have any questions or know of an computational tool or resource you would like to see included in this list, please contact: textmine@mit.edu

MIT Libraries subscription resources and tools

- Alpha Vantage Stock API
- American Association for the Advancement of Science (AAAS) articles
- American Chemical Society (ACS) articles
- American Physical Society (APS) articles
- Brill Academic Publishers articles
- Chicago Defender Historical Archive
- Constellate
- CQ Press content

Freely available resources and tools

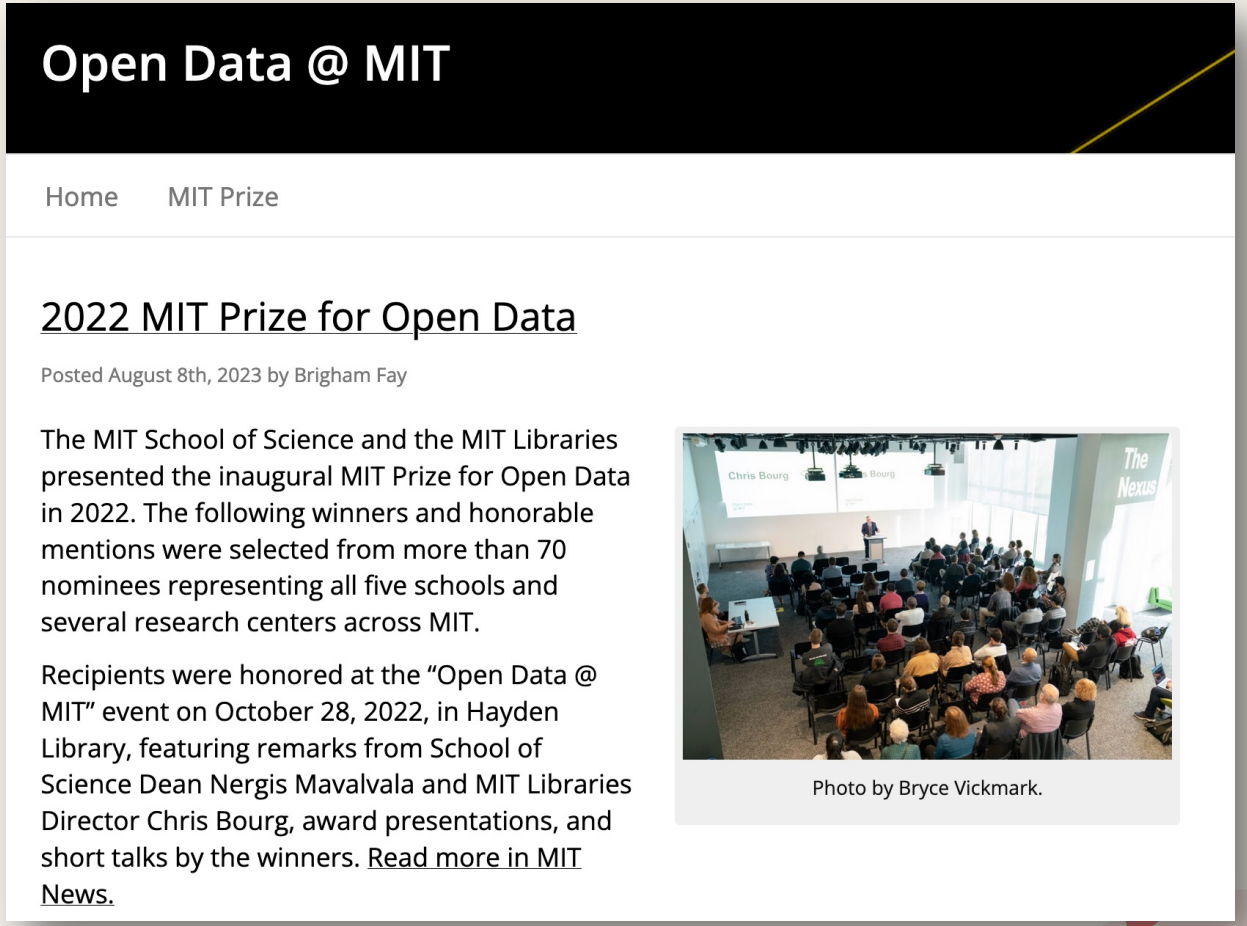
- American Archive of Public Broadcasting API
- arXiv Preprint Server
- BioMed Central Journal articles
- Caselaw Access Project
- Chronicling America
- Congress.gov
- CORE
- CrossRef DOI Registry Agency

See comparison grid at bit.ly/api-tdm

https://libguides.mit.edu/comptools, Accessed Feb 14, 2024

MIT Prize for Open Data recognizes impactful data sharing effort

- Two winner in Chemical Engineering for Inaugural year
- 10 + nominations in chemistry, chemical engineering, materials science and engineering for both 2022 and 2023



Open Data @ MIT

Home MIT Prize

2022 MIT Prize for Open Data

Posted August 8th, 2023 by Brigham Fay

The MIT School of Science and the MIT Libraries presented the inaugural MIT Prize for Open Data in 2022. The following winners and honorable mentions were selected from more than 70 nominees representing all five schools and several research centers across MIT.

Recipients were honored at the "Open Data @ MIT" event on October 28, 2022, in Hayden Library, featuring remarks from School of Science Dean Nergis Mavalvala and MIT Libraries Director Chris Bourg, award presentations, and short talks by the winners. Read more in MIT News.

Photo by Bryce Vickmark.

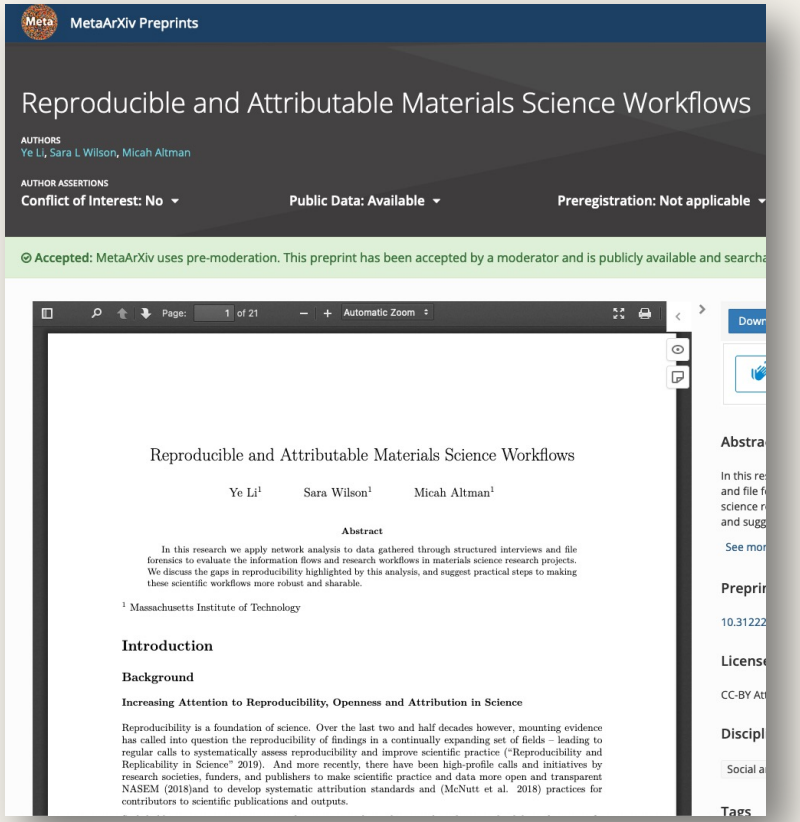
https://libraries.mit.edu/opendata/open-data-mit-home/mit-prize/2022-mit-prize-for-open-data/ , Accessed Feb 14, 2024

Extrinsic incentives (fund, promotion, recognition, and career advancement etc.) may not be sufficient to justify the additional time and effort to curate chemical data

"We spend time to take care of our data for **reproducibility** reasons... ...

It is **difficult** to **justify** the extra time spent **in further curating the data** so that others can use our data to publish a predictive model that **only benefits their own academic career**... ..."

-- Paraphrased from researchers' comments



Ye Li, Sara L. Wilson, and Micah Altman. Reproducible and Attributable Materials Science Workflows. MetaArXiv Preprints, **2022**, DOI:10.31222/osf.io/93a4e [Dataset] DOI:10.5281/zenodo.7158642 *(Accepted to present at IDCC24 and publish in the International Journal of Digital Curation)*

Besides extrinsic incentives, what could be the intrinsic motivations?

- Advancing science
- Accelerated discovery

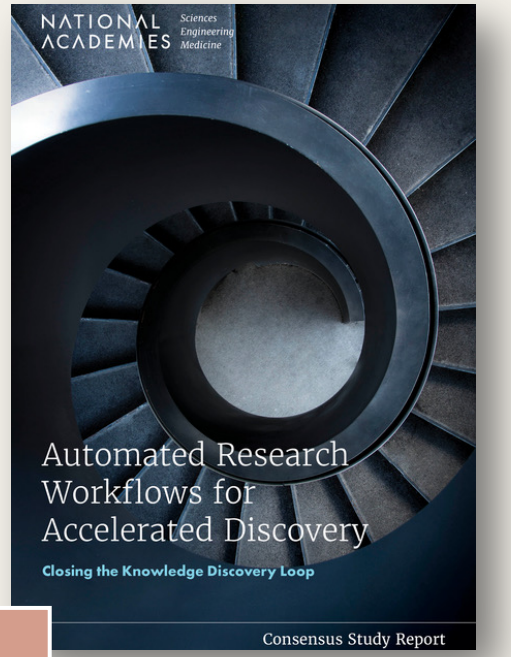
… especially what could make direct impact to their own lab!

Automated Research Workflow (ARW) may provide such an intrinsic motivation

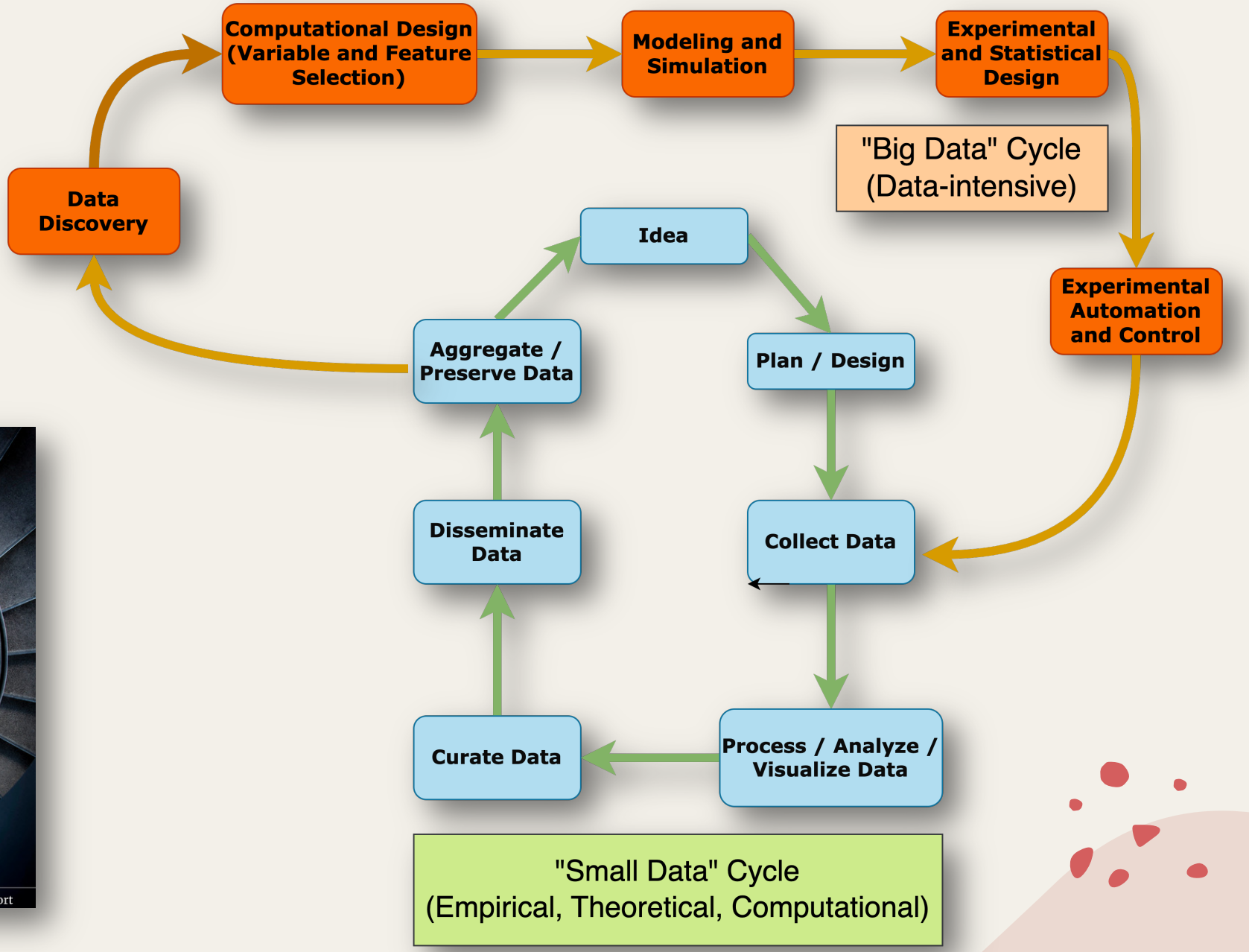
"… It could be a different story if the data we shared can be **immediately turned into automated design** in our lab through one of these ML/AI models **integrated in our workflow**."

-- Paraphrased from researchers' comments

Automated Research Workflow (ARW) closes the loop between “Big Data” and ”Small Data”

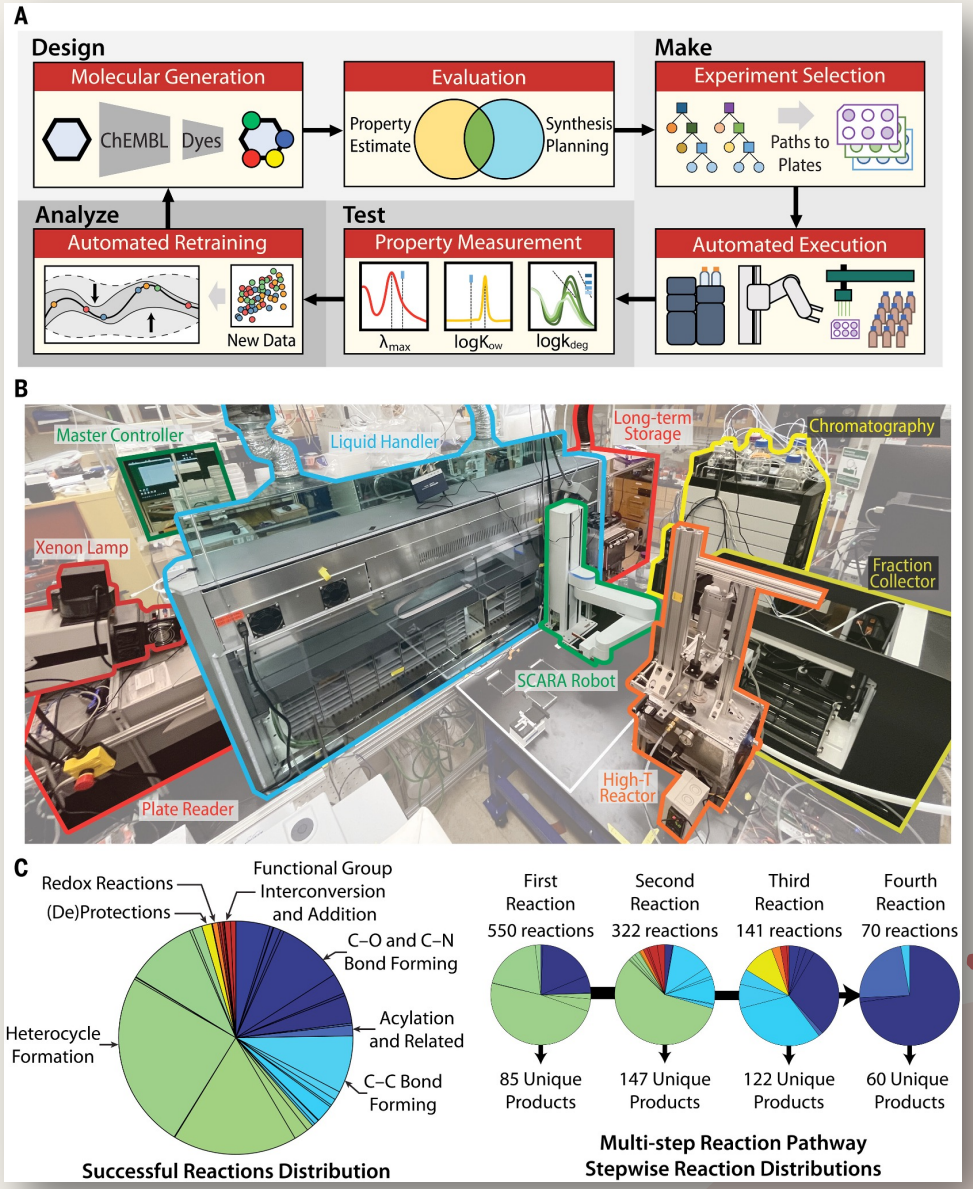


2022



ARW has become reality in chemistry, gradually but surely

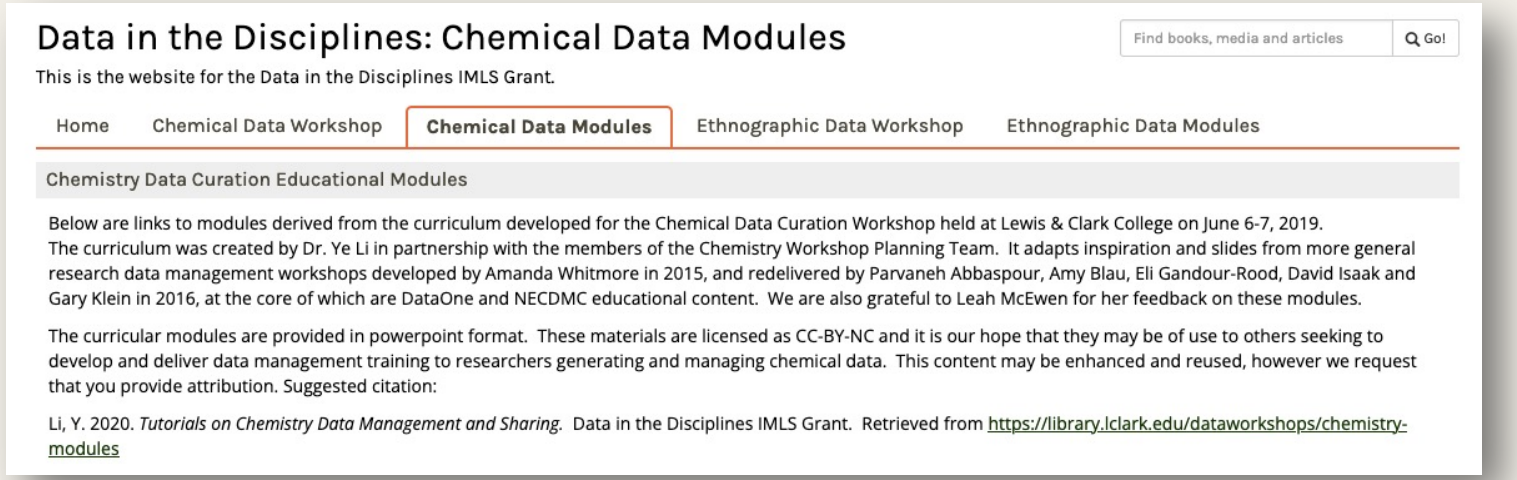
Further progress requires more **quality experimental data**, curated OR generated from FAIR data practices



Koscher, B. A, *et al.* (2023). **Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back.** *Science*, 382(6677), eadi1407. https://doi.org/10.1126/science.adi1407

Adding chemistry flavor to trainings on data management and sharing skills

- Leveraging grant-funded projects to amplify impact



Data in the Disciplines: Chemical Data Modules

Find books, media and articles | Go!

This is the website for the Data in the Disciplines IMLS Grant.

Home | Chemical Data Workshop | **Chemical Data Modules** | Ethnographic Data Workshop | Ethnographic Data Modules

Chemistry Data Curation Educational Modules

Below are links to modules derived from the curriculum developed for the Chemical Data Curation Workshop held at Lewis & Clark College on June 6-7, 2019. The curriculum was created by Dr. Ye Li in partnership with the members of the Chemistry Workshop Planning Team. It adapts inspiration and slides from more general research data management workshops developed by Amanda Whitmore in 2015, and redelivered by Parvaneh Abbaspour, Amy Blau, Eli Gandour-Rood, David Isaak and Gary Klein in 2016, at the core of which are DataOne and NECDMC educational content. We are also grateful to Leah McEwen for her feedback on these modules.

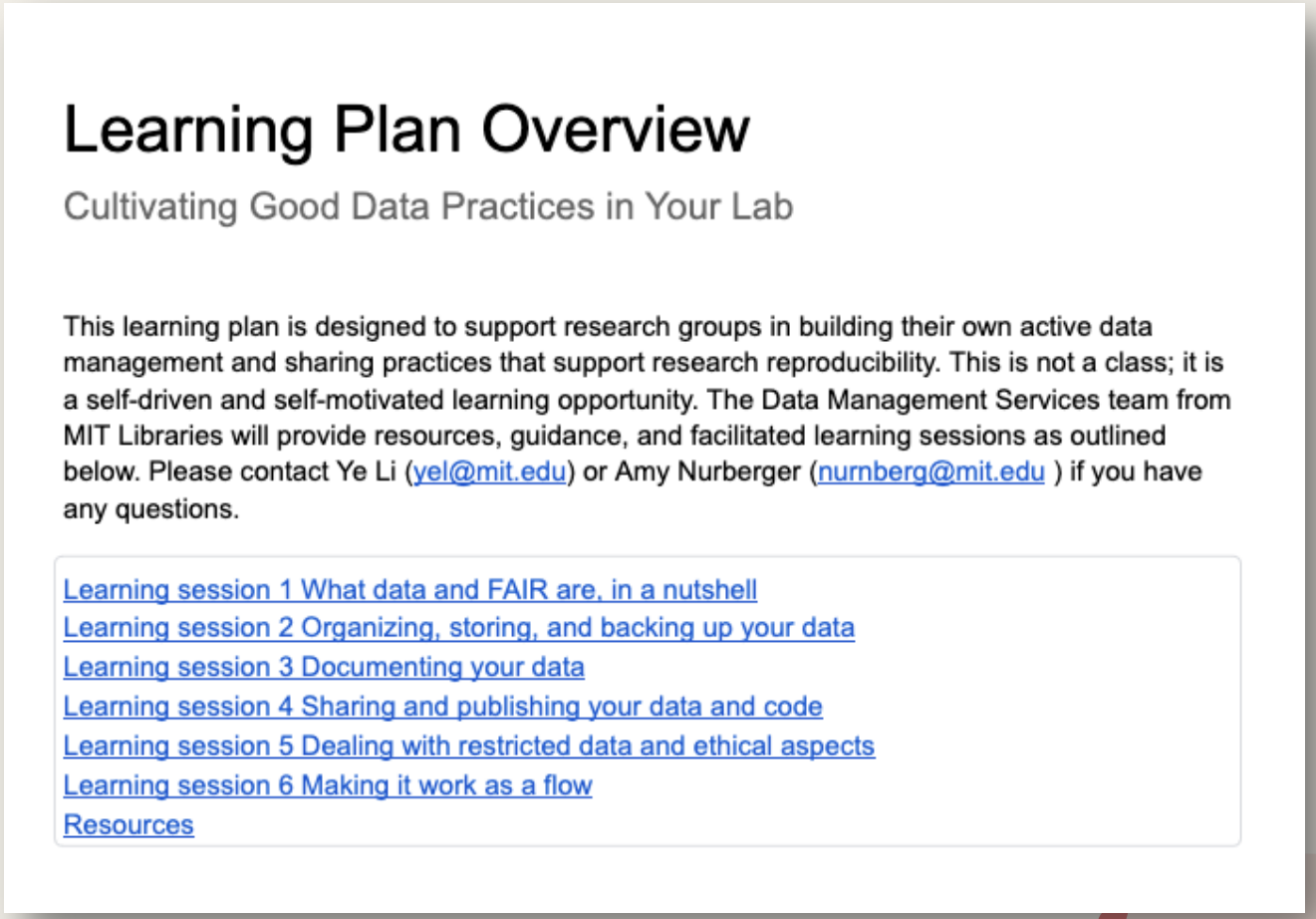
The curricular modules are provided in powerpoint format. These materials are licensed as CC-BY-NC and it is our hope that they may be of use to others seeking to develop and deliver data management training to researchers generating and managing chemical data. This content may be enhanced and reused, however we request that you provide attribution. Suggested citation:

Li, Y. 2020. *Tutorials on Chemistry Data Management and Sharing.* Data in the Disciplines IMLS Grant. Retrieved from https://library.lclark.edu/dataworkshops/chemistry-modules

Li, Y. 2020. *Tutorials on Chemistry Data Management and Sharing.* Data in the Disciplines IMLS Grant. Retrieved from https://library.lclark.edu/dataworkshops/chemistry-modules

Providing framework, strategies, and resources for research groups to improve their practices

- Workshop series customized for individual research groups can be most effective but difficult to scale
- Identifying data champions in each group is critical



Learning Plan Overview

Cultivating Good Data Practices in Your Lab

This learning plan is designed to support research groups in building their own active data management and sharing practices that support research reproducibility. This is not a class; it is a self-driven and self-motivated learning opportunity. The Data Management Services team from MIT Libraries will provide resources, guidance, and facilitated learning sessions as outlined below. Please contact Ye Li (yel@mit.edu) or Amy Nurberger (nurnberg@mit.edu) if you have any questions.

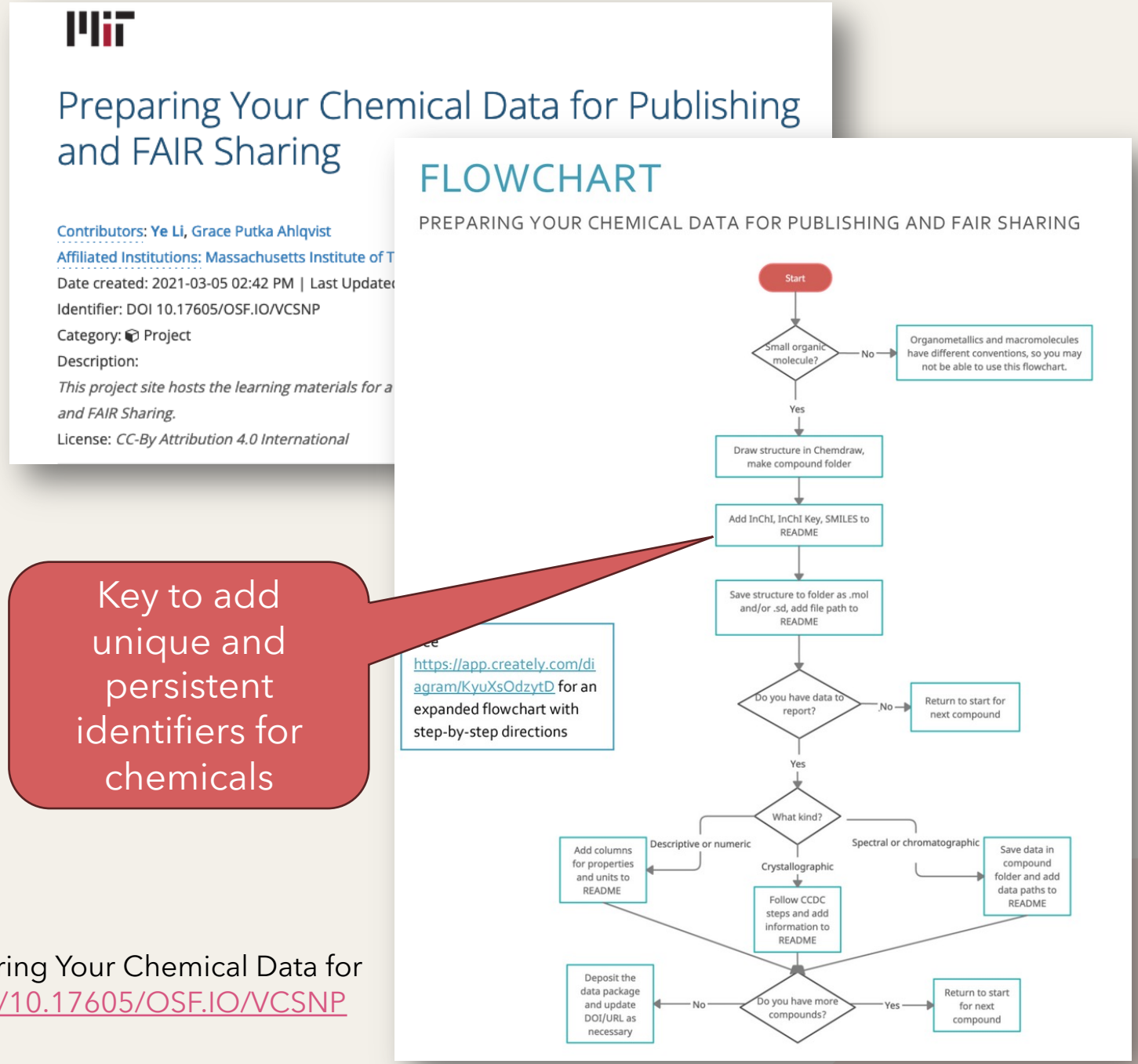
Learning session 1 What data and FAIR are, in a nutshell
Learning session 2 Organizing, storing, and backing up your data
Learning session 3 Documenting your data
Learning session 4 Sharing and publishing your data and code
Learning session 5 Dealing with restricted data and ethical aspects
Learning session 6 Making it work as a flow
Resources

An example workshop series provided by Data Management Services at MIT Libraries

Engaging data champions among graduate students and research staff to scale up in-depth support

EX:

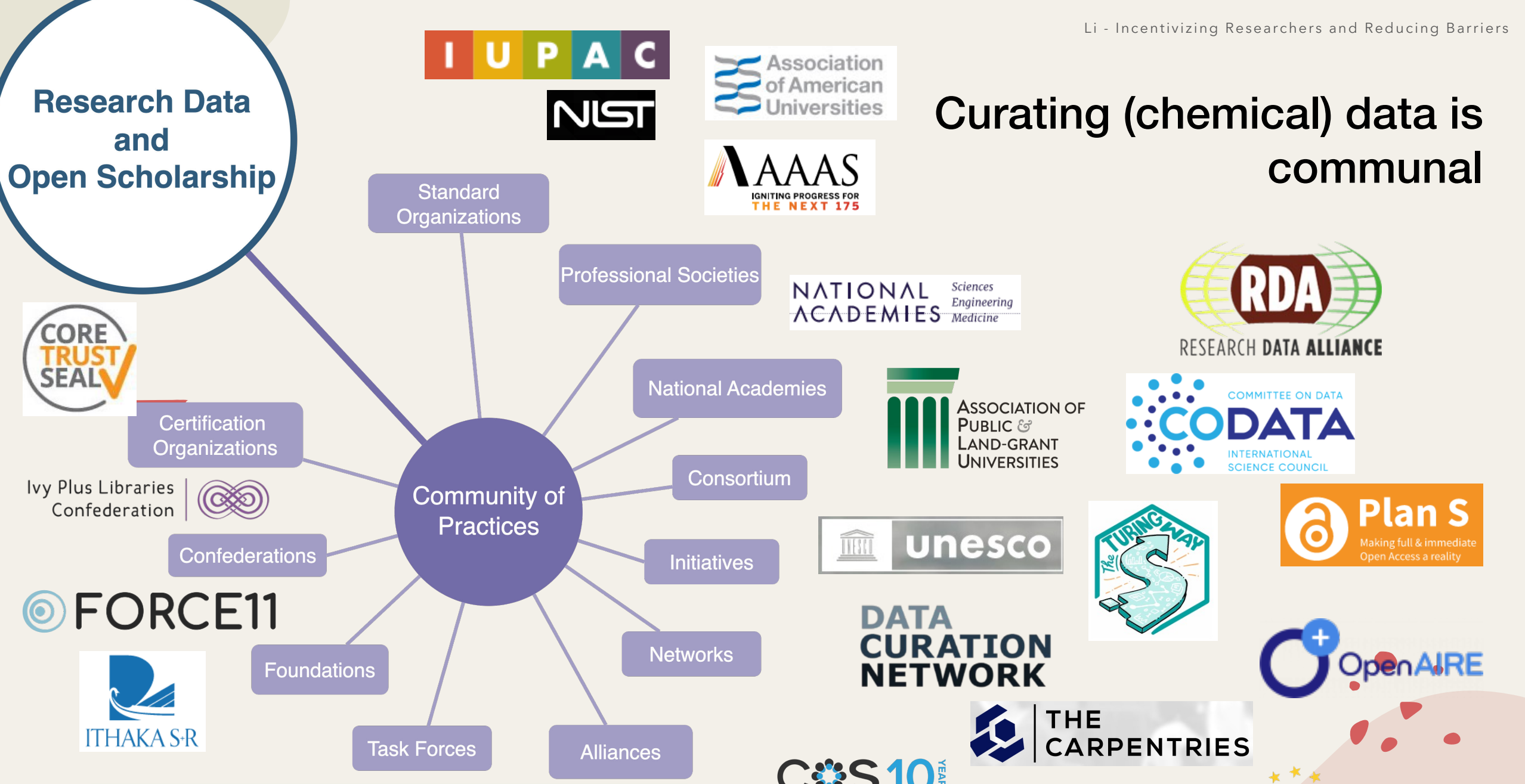
Minimalist approach towards FAIR practices before data standards become broadly adopted



Key to add unique and persistent identifiers for chemicals

Li, Y., & Ahlqvist, G. P. (2021, June 28). Preparing Your Chemical Data for Publishing and FAIR Sharing. https://doi.org/10.17605/OSF.IO/VCSNP

Curating (chemical) data is communal

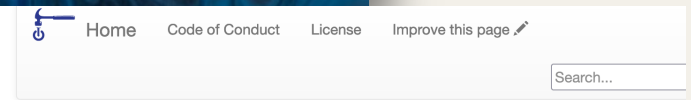


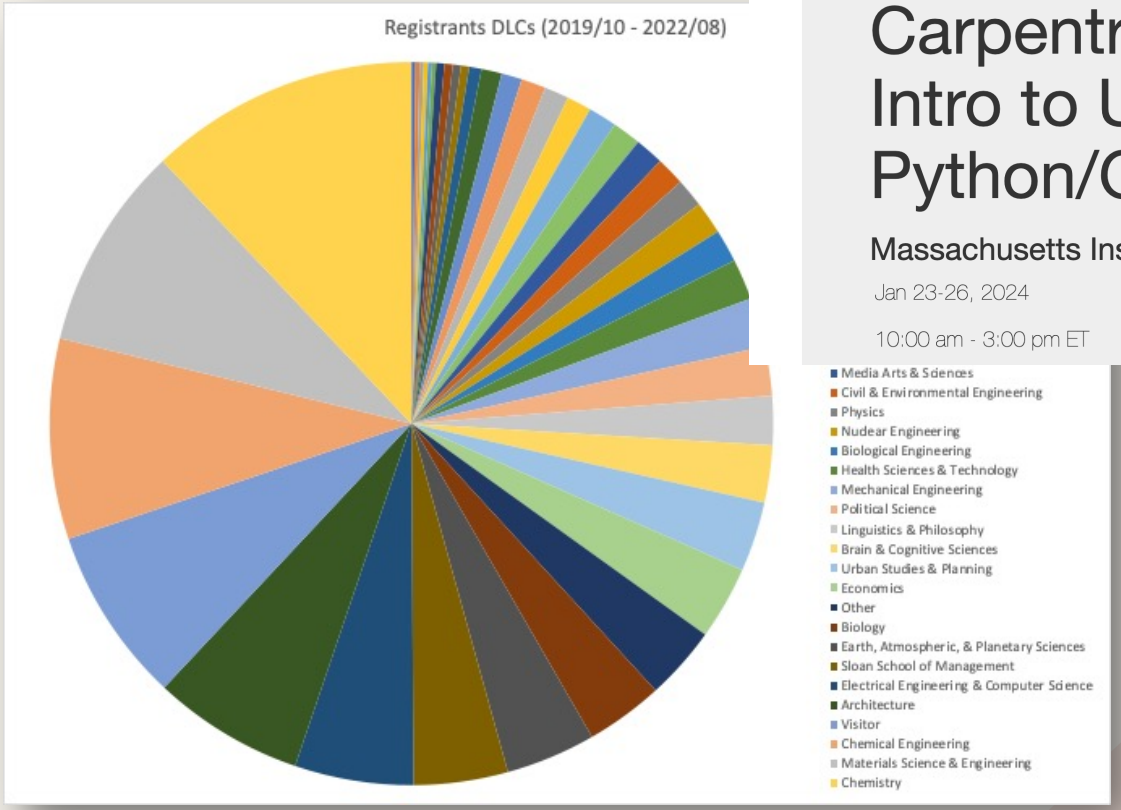
Roundtable on Aligning Incentives for Open Scholarship

Delivering open science and reproducible research skills through teaching foundational coding skills

- Leveraging the power of the global learning community
- Participation from the Departments of Chemistry, Materials Science and Engineering, and Chemical Engineering ranked top three

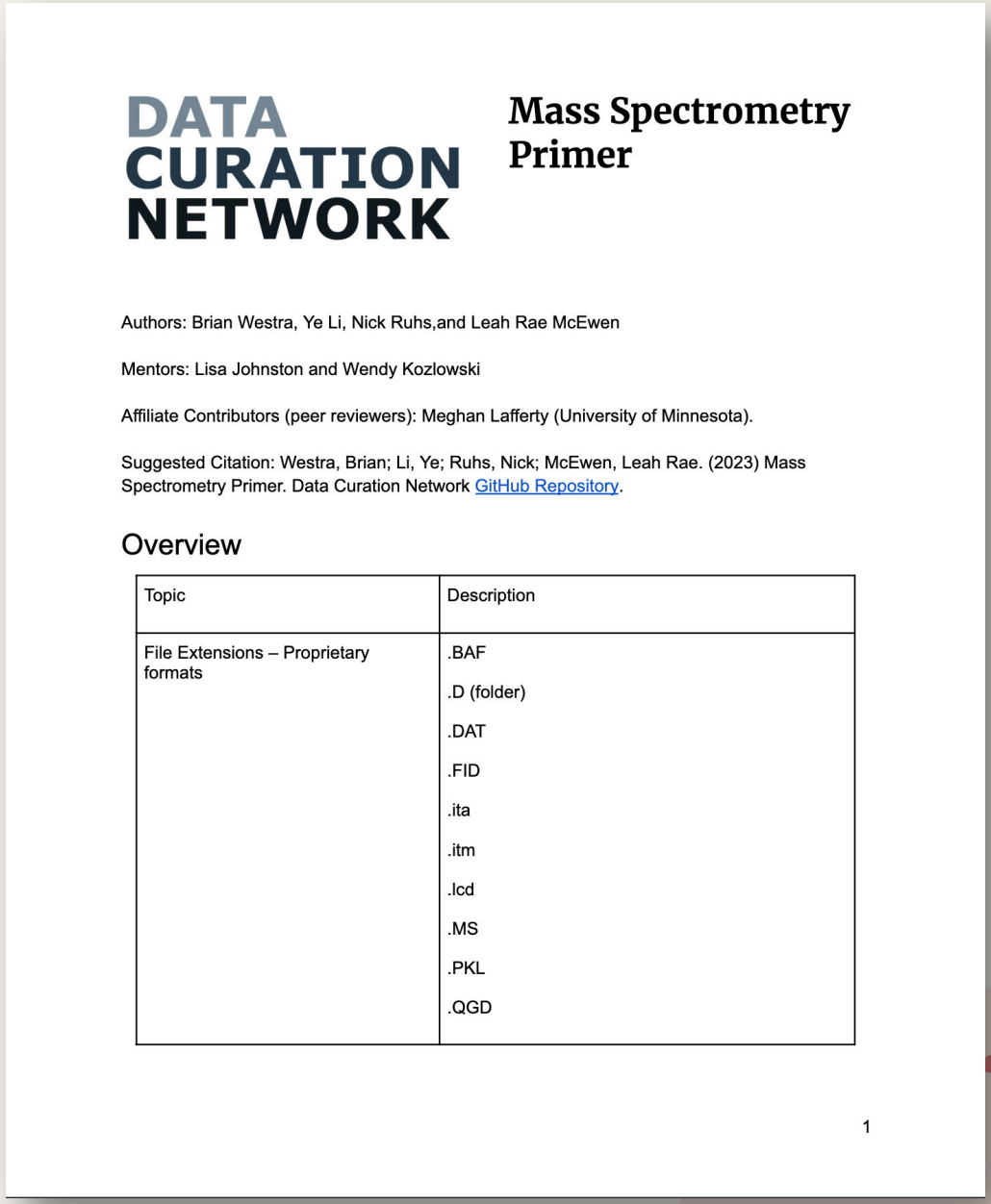
https://github.com/Carpentries-MIT





Providing guidelines on curating a specific type of chemical data through community of practices

- Supporting both researchers and data curators
- From formats to metadata and documentation
- Availability varies depending on data types



DATA CURATION NETWORK

Mass Spectrometry Primer

Authors: Brian Westra, Ye Li, Nick Ruhs,and Leah Rae McEwen

Mentors: Lisa Johnston and Wendy Kozlowski

Affiliate Contributors (peer reviewers): Meghan Lafferty (University of Minnesota).

Suggested Citation: Westra, Brian; Li, Ye; Ruhs, Nick; McEwen, Leah Rae. (2023) Mass Spectrometry Primer. Data Curation Network GitHub Repository.

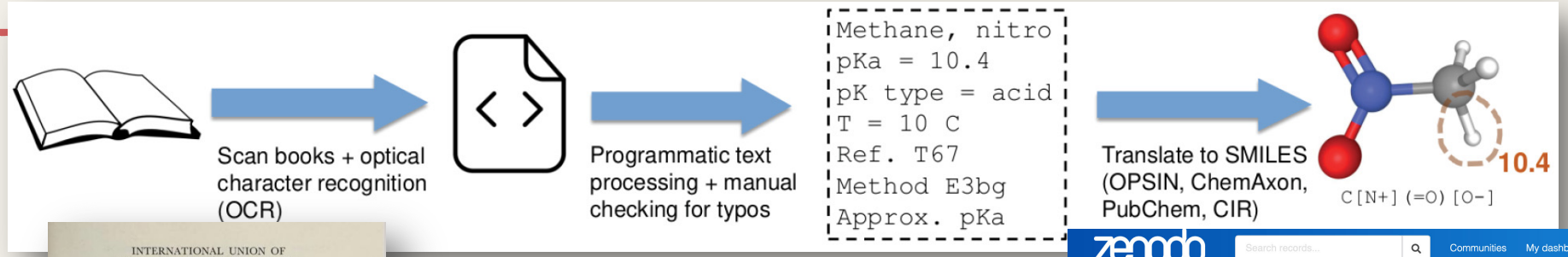
Overview

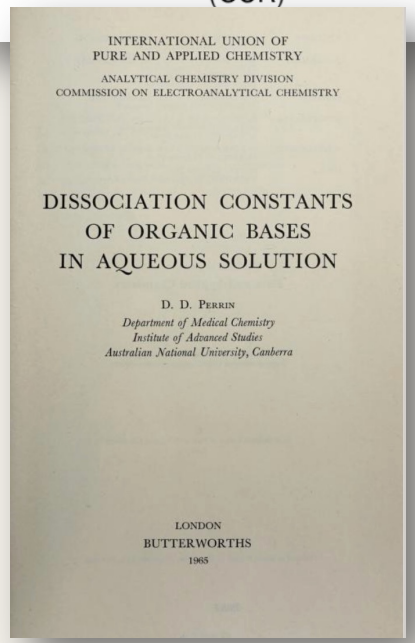
Topic	Description
File Extensions – Proprietary formats	.BAF .D (folder) .DAT .FID .ita .itm .lcd .MS .PKL .QGD

1

Westra, Brian; Li, Ye; Ruhs, Nick; McEwen, Leah Rae. (2023). Mass Spectrometry Primer. Data Curation Network. https://hdl.handle.net/11299/259173 .

Connecting researchers to community of practices to mine and curate chemical data collaboratively



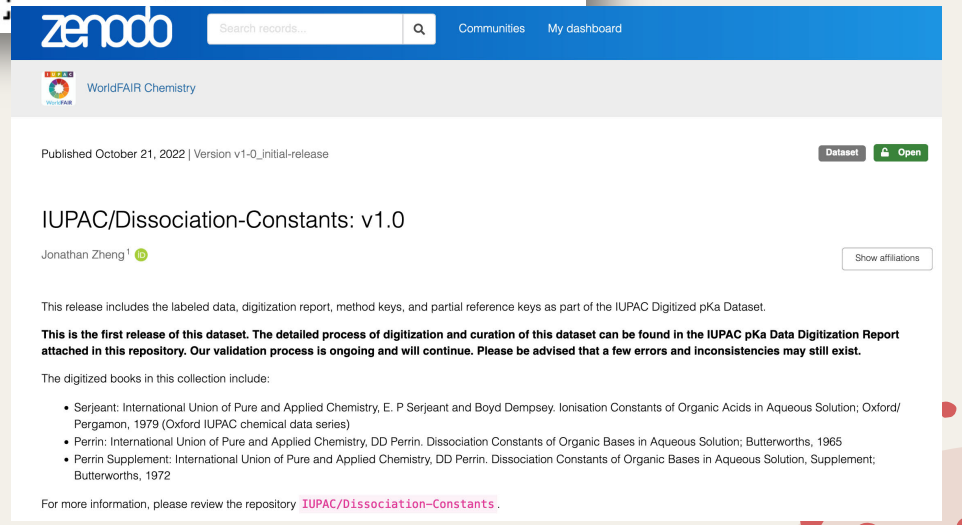

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

ANALYTICAL CHEMISTRY DIVISION
COMMISSION ON ELECTROANALYTICAL CHEMISTRY

DISSOCIATION CONSTANTS OF ORGANIC BASES IN AQUEOUS SOLUTION

D. D. PERRIN
Department of Medical Chemistry
Institute of Advanced Studies
Australian National University, Canberra

LONDON
BUTTERWORTHS
1965

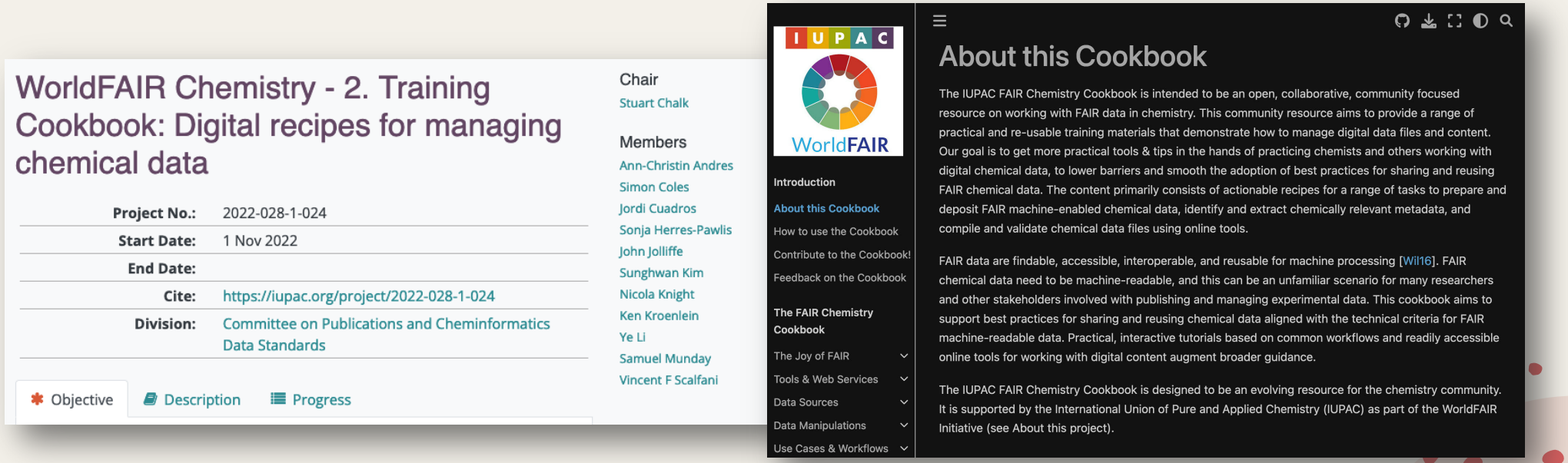


Zheng, Jonathan, Li, Ye and McEwen, Leah. "FAIR datasets for acid dissociation constants" *Chemistry International*, 45(3), 2023, 26-29. https://doi.org/10.1515/ci-2023-0310

Jonathan Zheng. (2022). IUPAC/Dissociation-Constants: v1.0 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7236453

Leveraging expertise from community of practices broadly to make global impact

https://iupac.org/project/2022-012-1-024/



https://iupac.github.io/WFChemCookbook

McEwen, Leah, & Bruno, Ian. (2023). WorldFAIR Project (D3.1) Digital recommendations for Chemistry FAIR data policy and practice (Version 1). Zenodo. https://doi.org/10.5281/zenodo.7887283

Data curation may continue long after the initial sharing and take communal coordination

- CCDC coordinated with publishers to retract crystallographic data related to paper-mill publications
- Retraction of datasets often ties to retraction of the related publications

CCDC

Deposit Structures | Access Structures

Home | Solutions | Community | Discover | Consultancy Services

Retractions in the Cambridge Structural Database

Solution

2023

Update: 1st November

In our last data release of 2023, scheduled for later this year, a further 152 entries implicated by the pre-print will be retracted. These retractions are already visible to users accessing the CSD via our web platforms and takes the total number of retractions related to the papermill pre-print in the CSD to 361.

We have updated the comments for 39 structures related to the papermill pre-print to say that following extensive review by the publishers and the CCDC there are currently no concerns on the data. A further 36 structures related to the papermill pre-print have been updated where our investigations have identified the entry has almost identical reflection data with another entry in the CSD with a different structural formula.

Investigations are continuing for the remaining 556 structures implicated by the pre-print. Cases where reflection data has not yet been made available can be especially challenging to assess. We will continue to work with publishers, reviewers and depositors and are grateful for how the community has come together to tackle the issues so far.

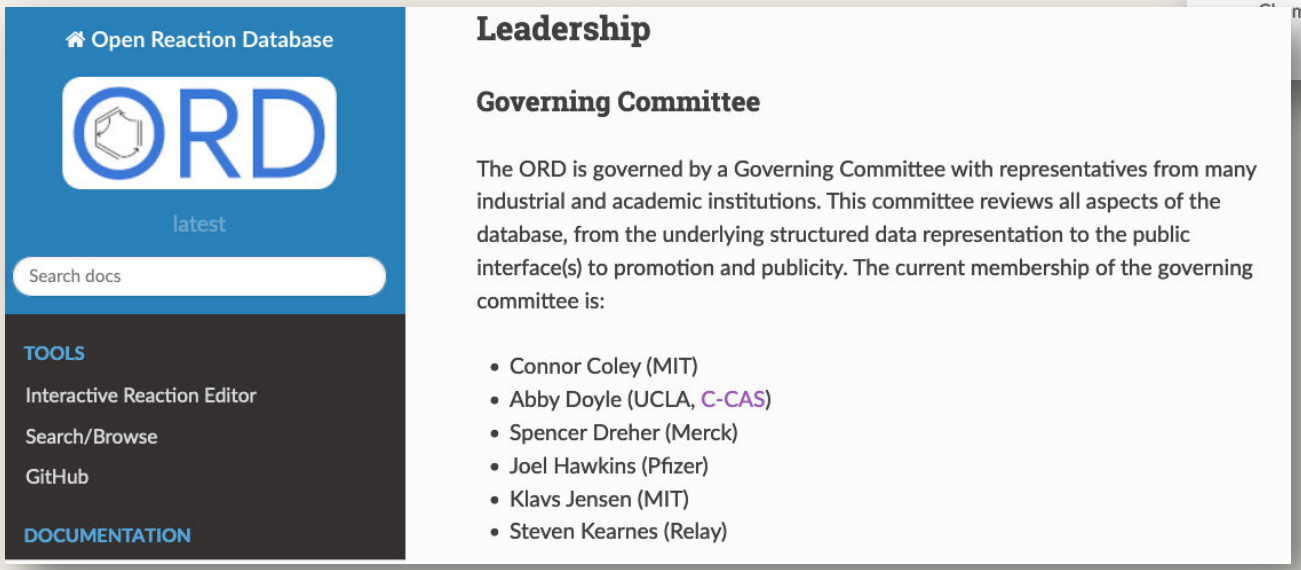
Work to broaden the automated data integrity checks and processes, conversations with publishers on how we can support their peer review

https://www.ccdc.cam.ac.uk/support-and-resources/support/case/?caseid=819cfd76-c25d-40a2-ac9b-b4cf20d775a7 , Accessed Feb 14, 2024

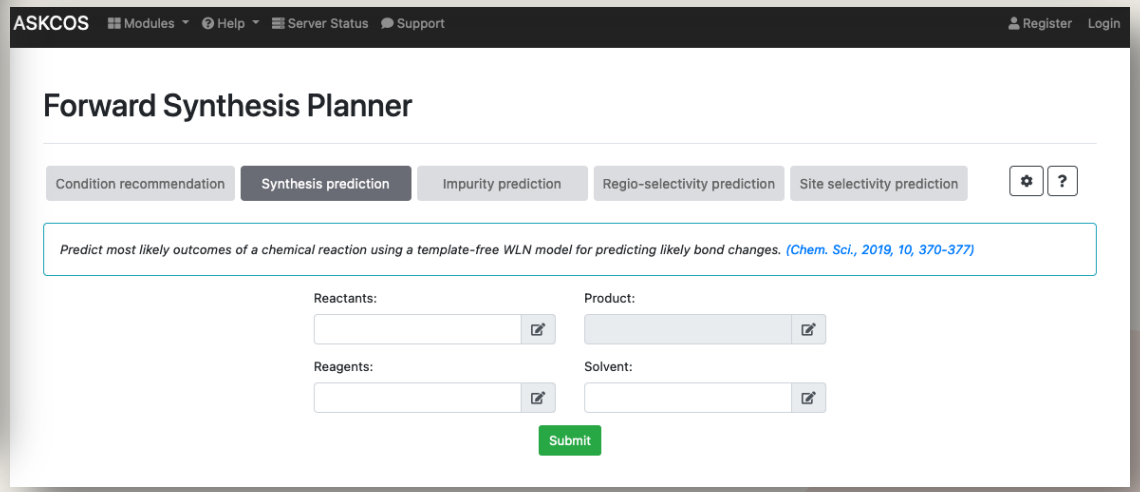
https://www.science.org/content/blog-post/faked-crystals-and-faked-data , Accessed Feb 14, 2024

Industry partnership can accelerate discovery while making progress in open and FAIR data

https://mlpds.mit.edu/

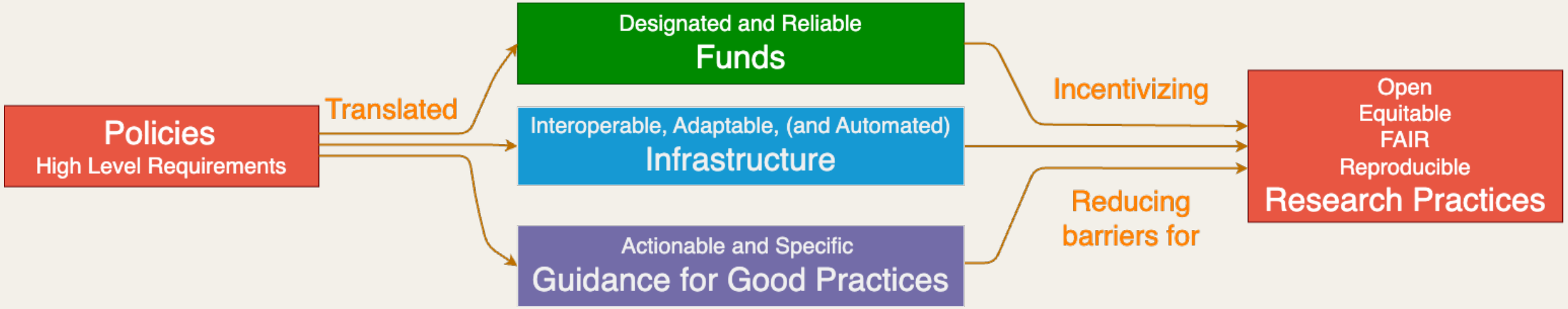


https://docs.open-reaction-database.org



https://askcos.mit.edu

All stakeholders collaborate to enable Open Science and FAIR data practices



Thank you!

Questions?

Contact Ye Li at
yel@mit.edu

NATIONAL ACADEMIES Sciences Engineering Medicine

Future Implications of Open-Access & FAIR Data Practices on Chemistry and Chemical Engineering Publications – A Workshop
Session III: Supporting the Research Community