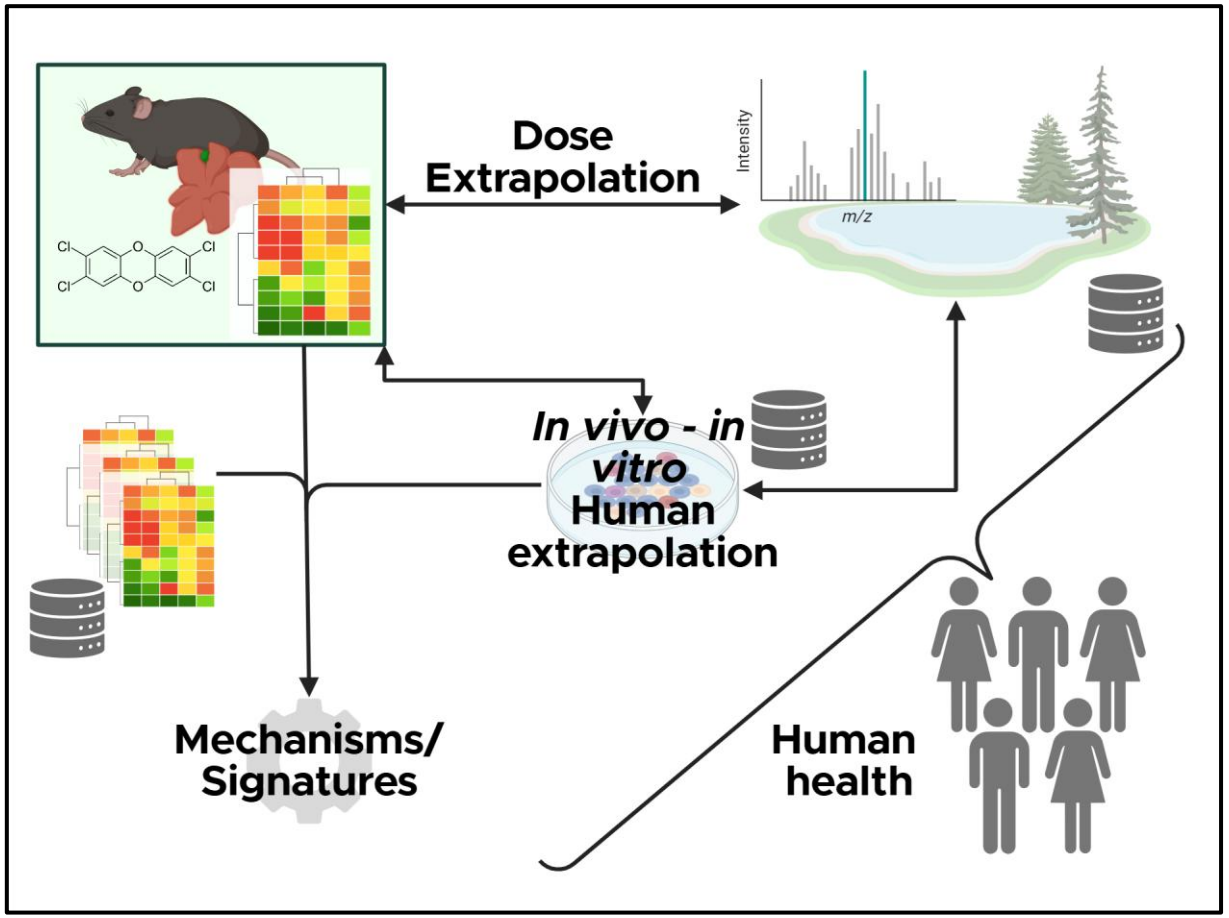
ToxDataCommons: Driving toxicology research forward through (meta)data sharing

Rance Nault
Department of Pharmacology & Toxicology
Institute for Integrative Toxicology
Superfund Research Center Data Management and Analysis Core
Michigan State University
naultran@msu.edu

WHY CARE ABOUT DATA SHARING



We **generate** huge amounts of data.

We also **NEED** more high-quality data to complete the picture.

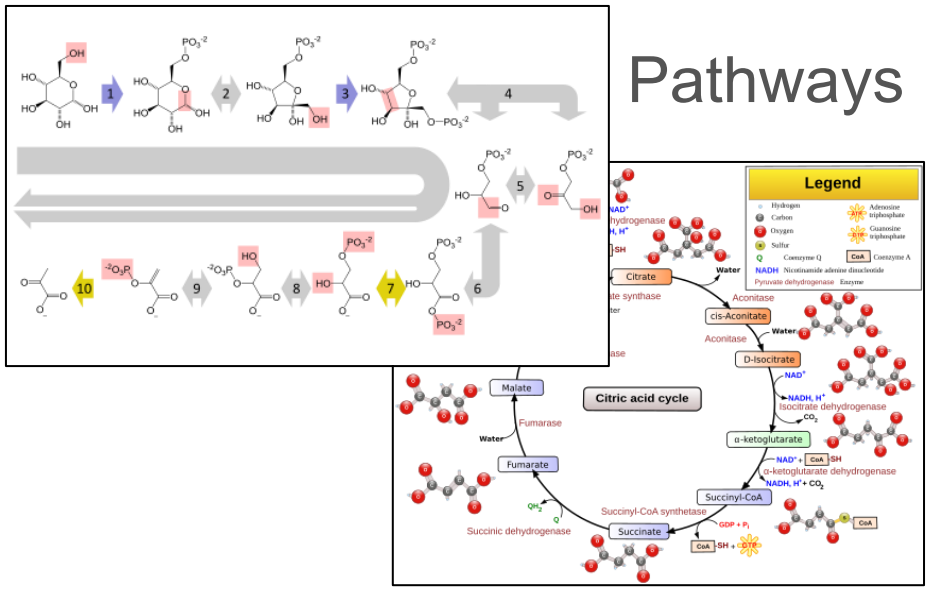
Image created with ChatGPT (DALL·E), 2025.

WHY CARE ABOUT DATA SHARING

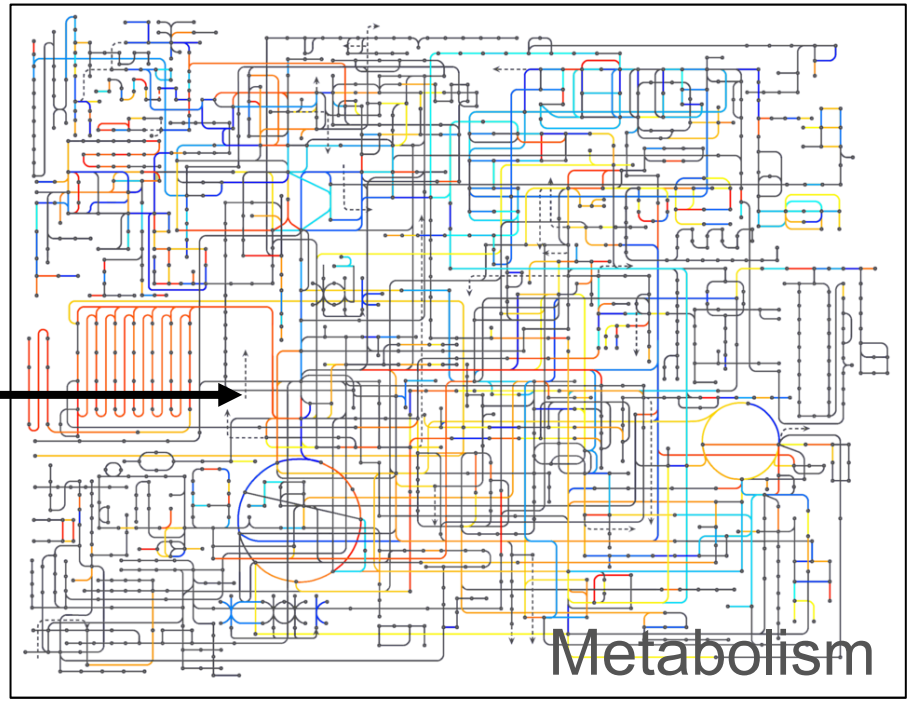
Reproducibility	Validation (*e.g.*, NAMs)	Discovery
• Can we get to the same answer using the same dataset? • Do the same study designs produce the same outcomes? • Can we explain differences in outcomes when metadata is richer?	• Benchmarking *in vitro* results to *in vivo* results (reducing animal use) • Can we build statistical/mathematical models to capture address differences between human, rodents, cells, …	• Leverage data from all studies of chemical X to refine safety limits, mechanisms, … • Using data hungry methods such as artificial intelligence to gain new knowledge

ROLE OF DATA IN AI APPLICATIONS

Research Question: Can we better characterize metabolic pathway changes by environmental toxicants considering metabolism as a network of reactions.



Pathways

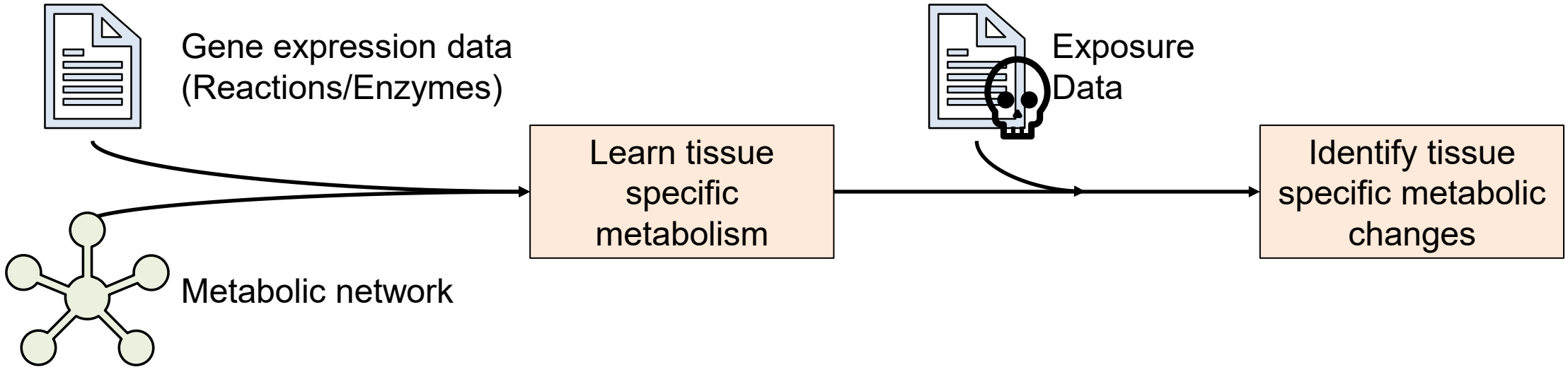


Metabolism

Yuan, K. and R. Nault (2025). Toxicol Sci 206(1): 19-29.

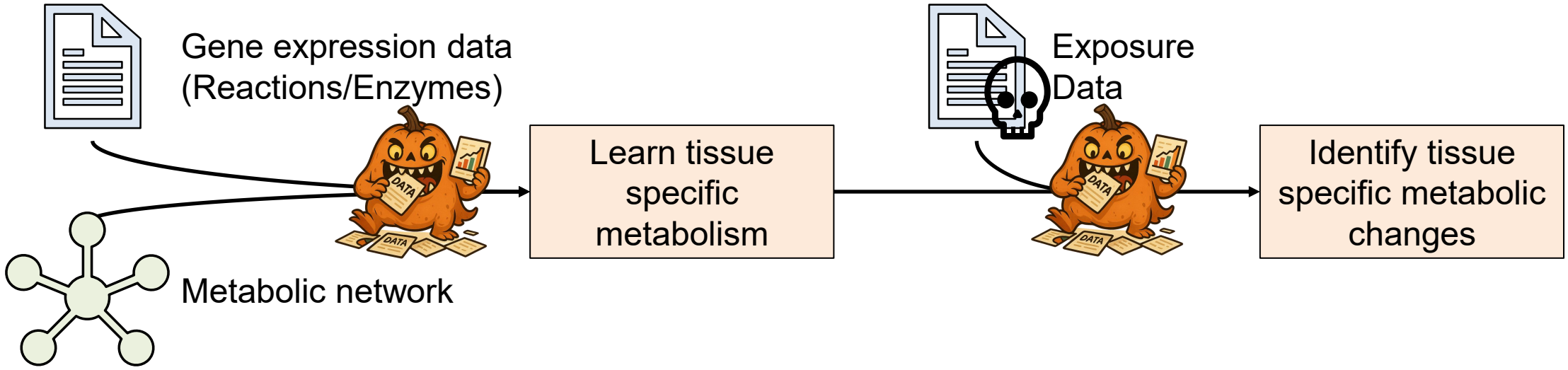
TRAINING A GRAPH NEURAL NETWORK

A Graph Neural Network (GNN) is a type of artificial intelligence applied to data that can be represented as a graph (*e.g.,* social networks, chemical structures)

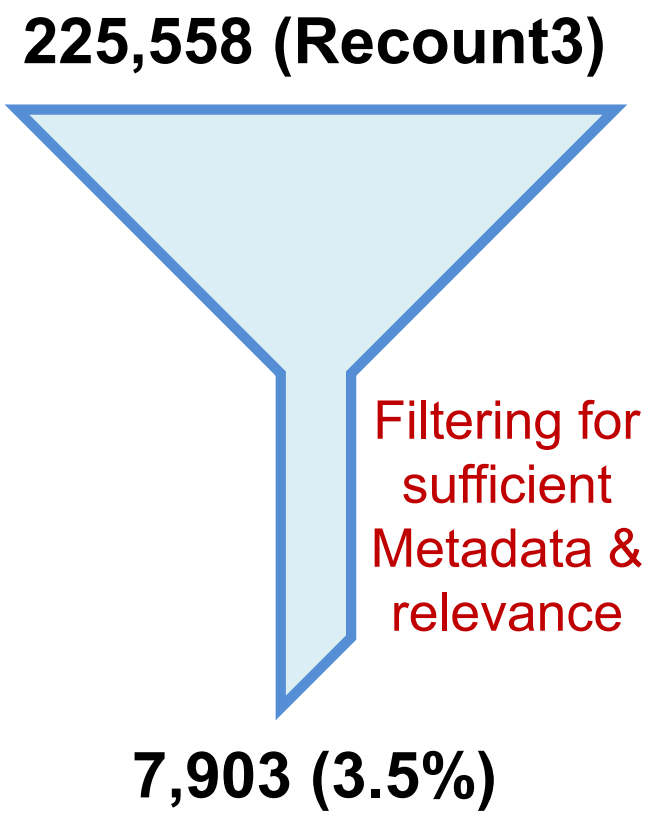


TRAINING A GRAPH NEURAL NETWORK

A Graph Neural Network (GNN) is a type of artificial intelligence applied to data that can be represented as a graph (*e.g.,* social networks, chemical structures)



TRAINING AI - THE METADATA PROBLEM

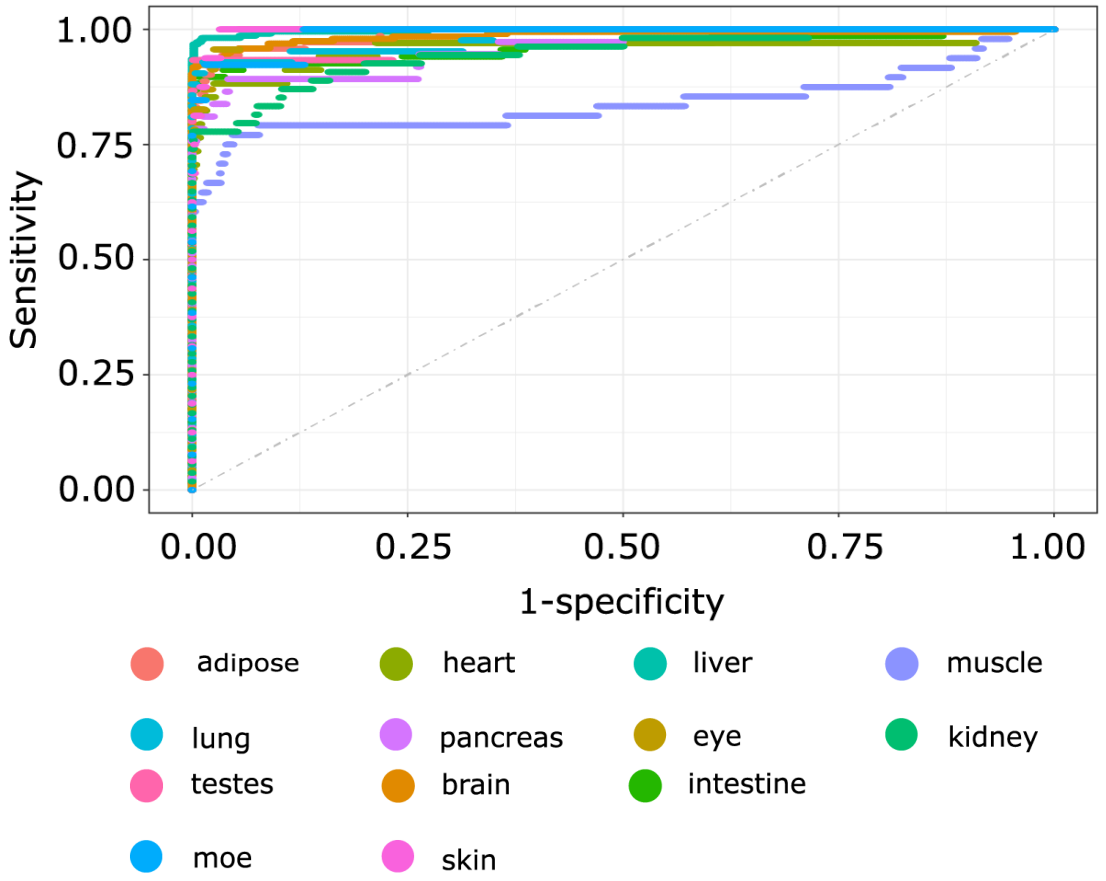


We started by downloading from GEO all metadata for 225,558 mouse samples.

- Many were excluded for relevance *(e.g., in vitro, genetically modified organisms, developmental (E0-E19), …).*
- A significant portion we couldn't determine the relevance *(missing strain, sex, treatment details, …).*

Toxicant perturbations samples were even more challenging to keep!

CLASSIFICATION OF TISSUE METABOLISM



Despite keeping only 3.5% of samples, there was sufficient data to accurately classify tissues based on expression of genes in the metabolic network.

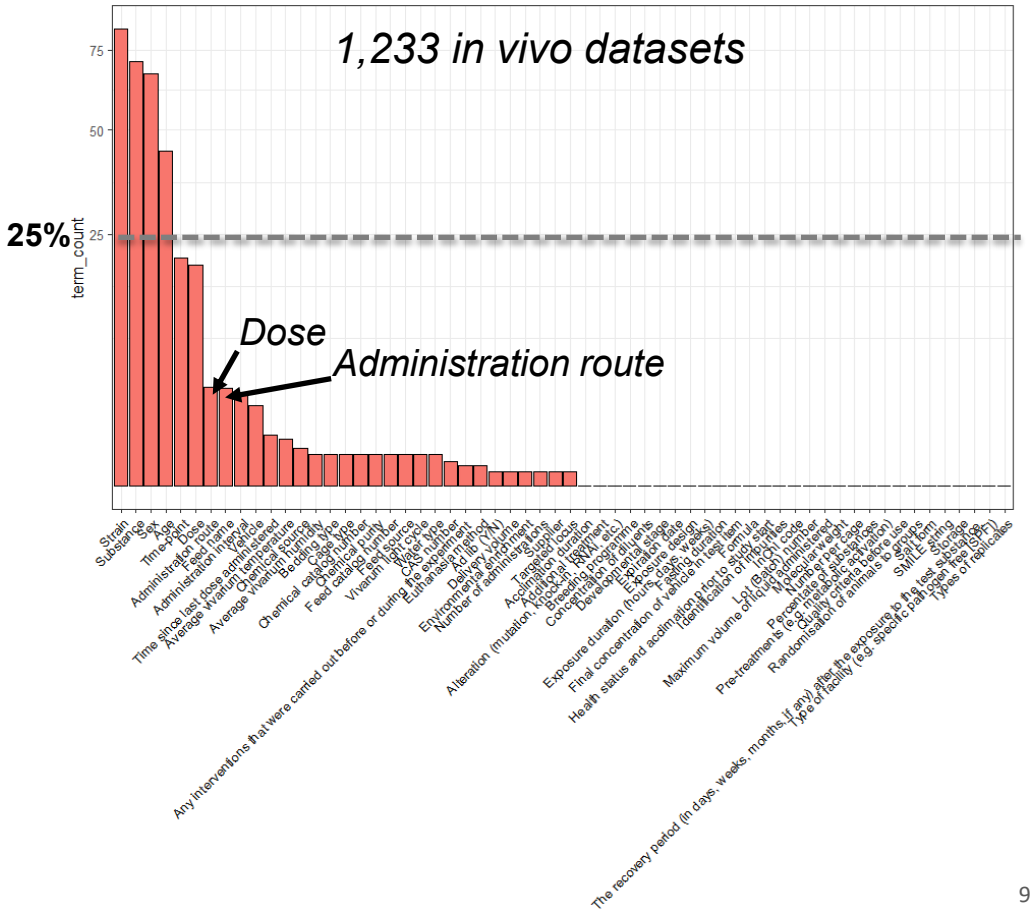
This was much more challenging for toxicant-driven gene expression changes.

UNDER-REPORTING OF METADATA

Several toxicant exposure gene expression studies could be identified.

Significant gaps in metadata reporting making it more challenging to provide AI enough information.

These challenges are also important in other data integration and reuse activities.



Nault, R., et al. (2023). Environ Health Perspect 131(6): 65001.

UNRAVELING DATA MANAGEMENT

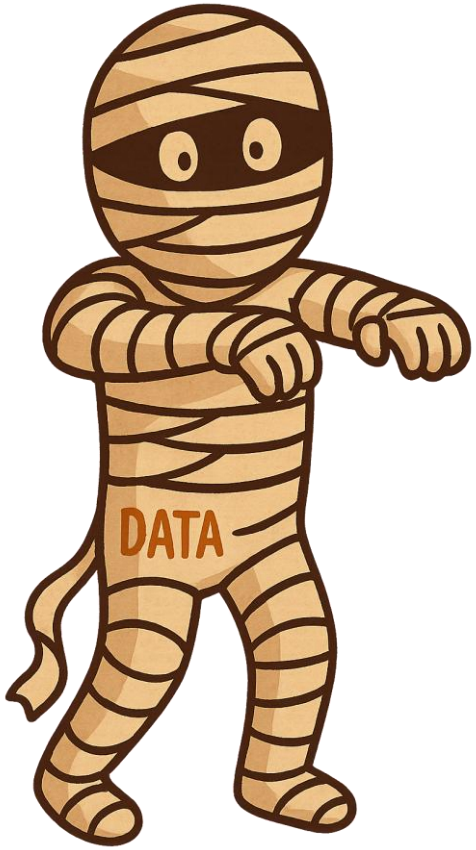
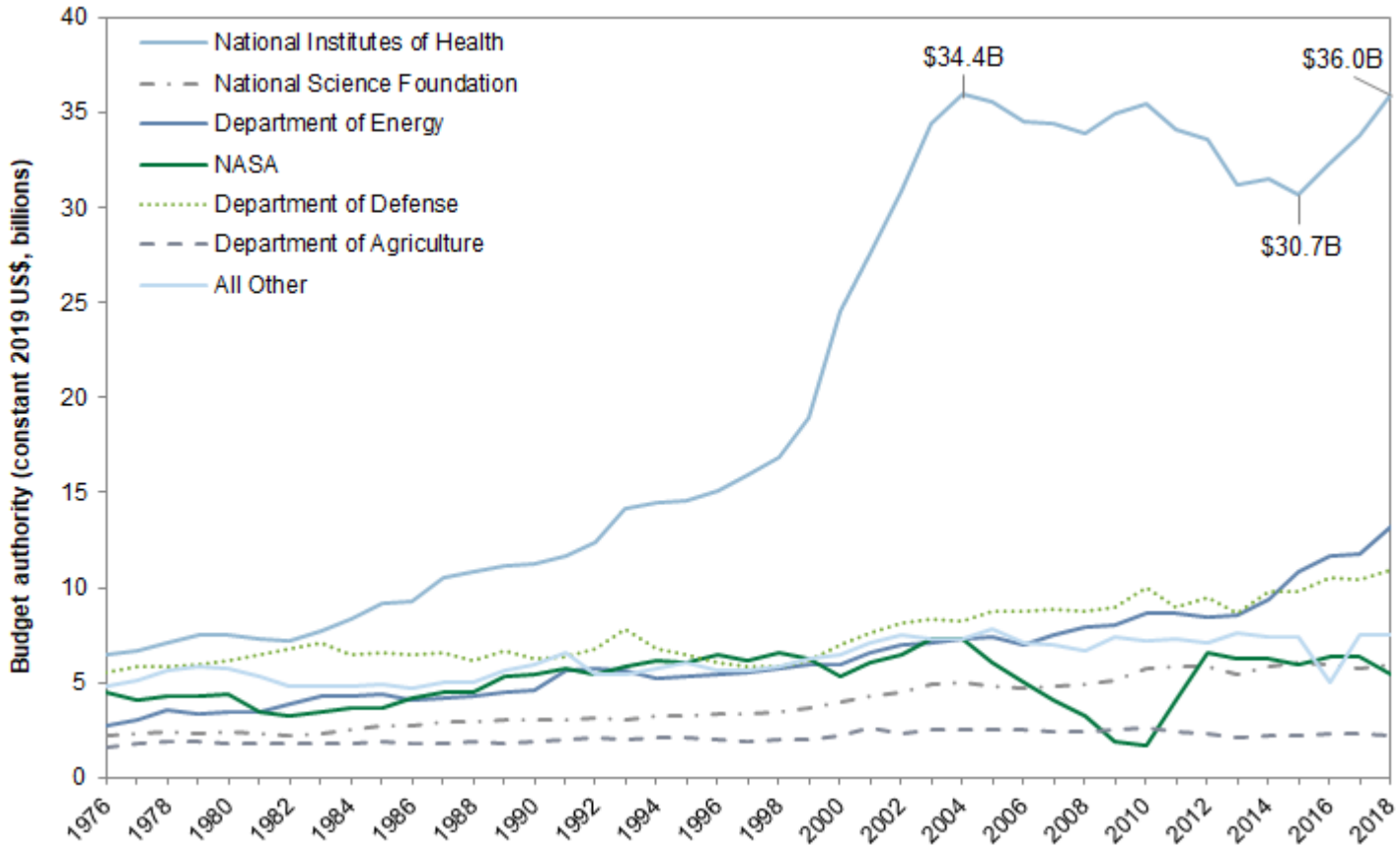


Image created with ChatGPT (DALL·E), 2025.

DATA HAS VALUE – BUT WE CAN ADD MORE



Billions of dollars are given to support the generation of new research data annually.

How can we maximize the return on investment and minimize unnecessary repetition.

Reproducibility is still important!

Source: American Association for the Advancement of Science, Goldman Sachs Global Investment Research

WIDESPREAD INITIATIVES

FAIR Principles have been central to new NIH data management policies.

"NIH encourages data management and data sharing practices consistent with the FAIR data principle"

NIH aims to require data sharing of all research data.

"… regardless of whether the data are used to support scholarly publications"

SRP … DMAC to support the management and integration of data assets.

NIH

NSF National Science Foundation

NIH National Institute of Environmental Health Sciences

Superfund Research Program

FAIR DATA PRINCIPLES

Image: https://www.linkedin.com/pulse/why-your-data-must-fair-david-crosswell/

FAIR principles represents a set of broad concepts, not a specific set of rules.

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons — Show fewer authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

487k Accesses | **4537** Citations | **2021** Altmetric | Metrics

FAIR DATA PRINCIPLES: ROLE OF METADATA

Image: https://www.lir

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

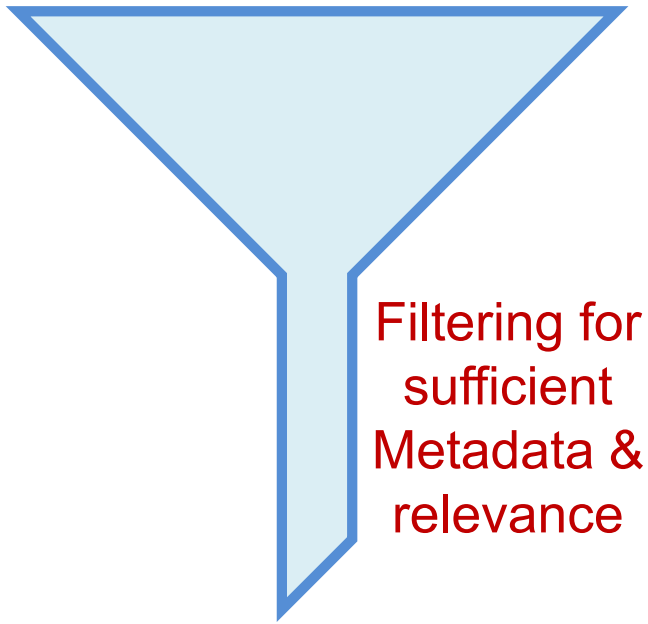
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

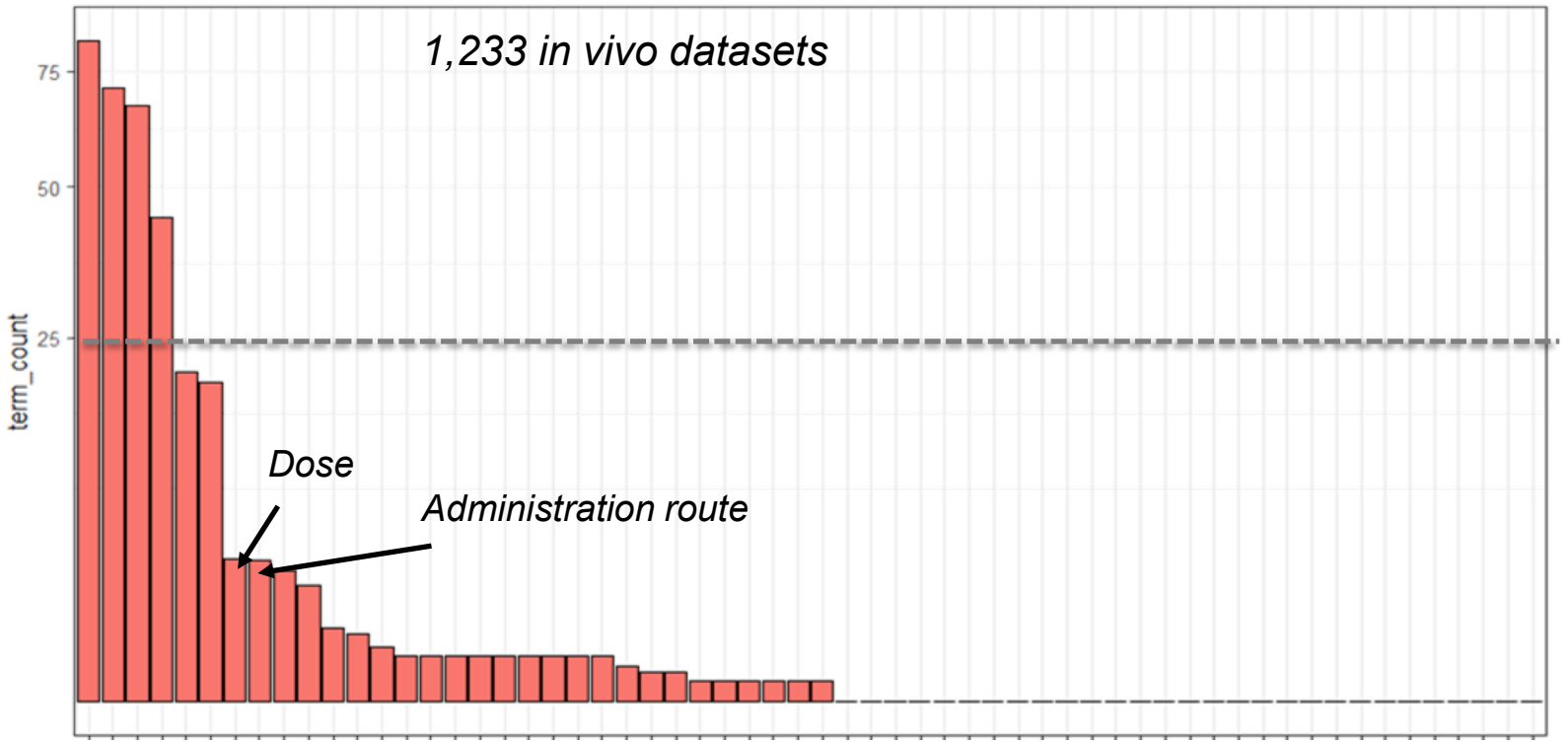
ARE WE ACHIEVING FAIR DATA?

225,558 (Recount3)



Filtering for sufficient Metadata & relevance

7,903 (3.5%)



WE *ARE* GOOD AT REPORTING METADATA

These studies had *poor reporting* in our analysis of GEO datasets

Study Name	Publication (manually) identified terms	GEO metadata terms
GSE116653: Analysis of the role of long non-coding RNAs involved in silica-induced pulmonary fibrosis of rat	**38**	**0**
GSE18858: Transcriptional Biomarkers to Predict Female Mouse Liver Tumors in Rodent Cancer Bioassays - A 26 Chemical Set	**45**	**0**
GSE70583: Response of rat liver to GW7647	**17**	**1**

Nault, R., et al. (2023). Environ Health Perspect 131(6): 65001.

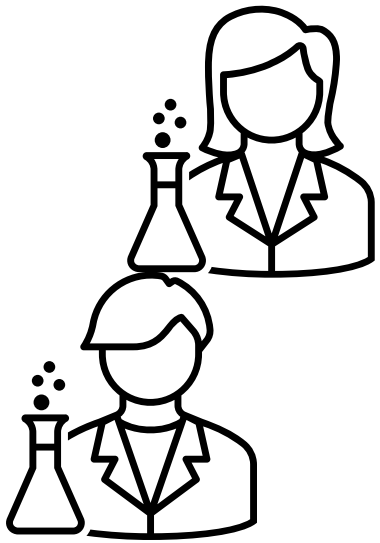
HOW DID WE GET HERE …

HOW DID WE GET HERE …

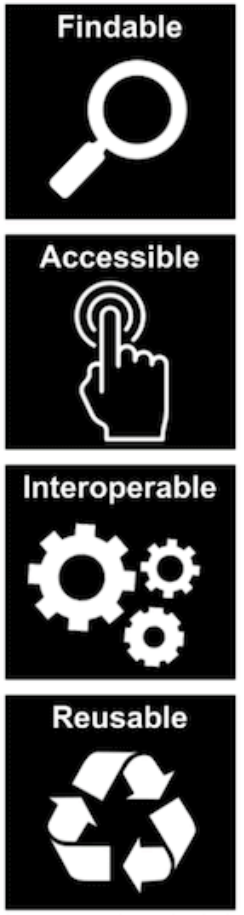
It's the tools, infrastructure, and their accessibility …

1) *Where are the reporting standards?*
2) *When/If you find them, how do you apply them?*
3) *How do you make sure they were applied correctly?*
4)*You have your standardized metadata … now what?*

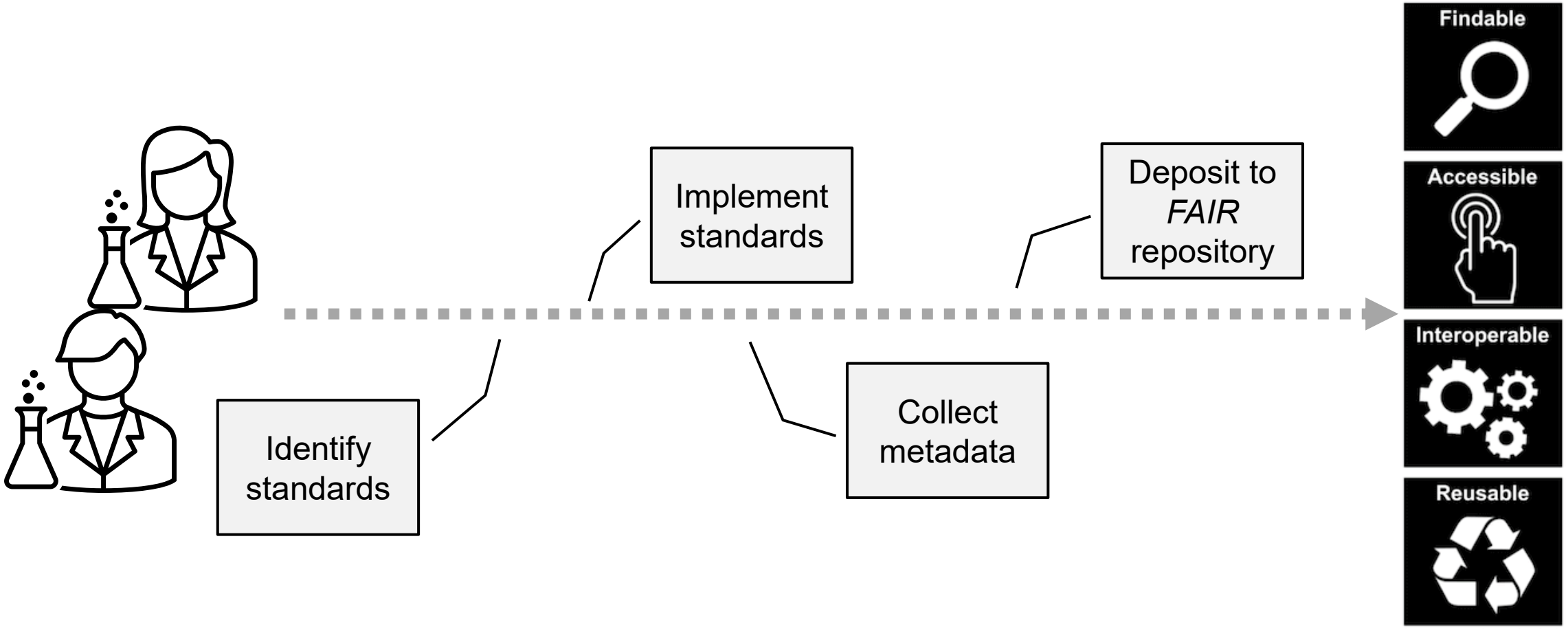
FROM THE LAB TO *FAIR*



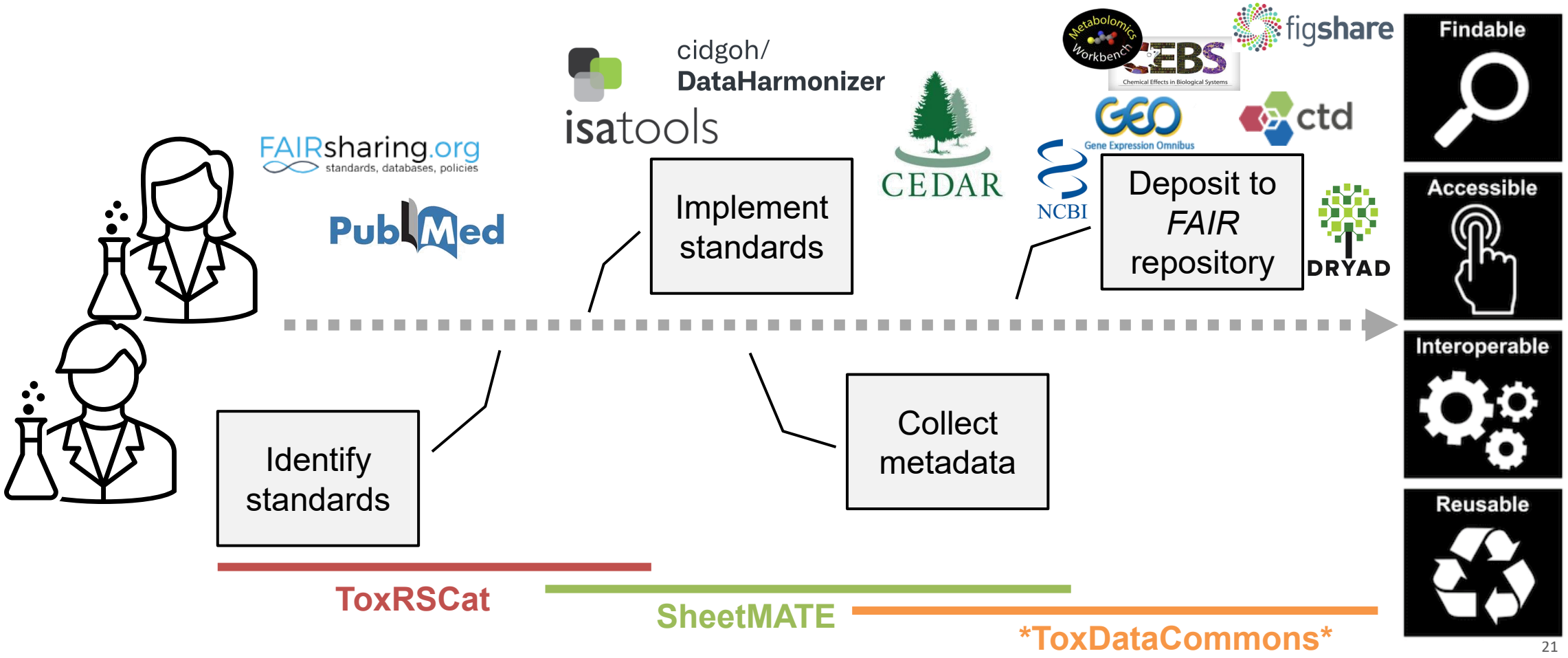
?



FROM THE LAB TO *FAIR*

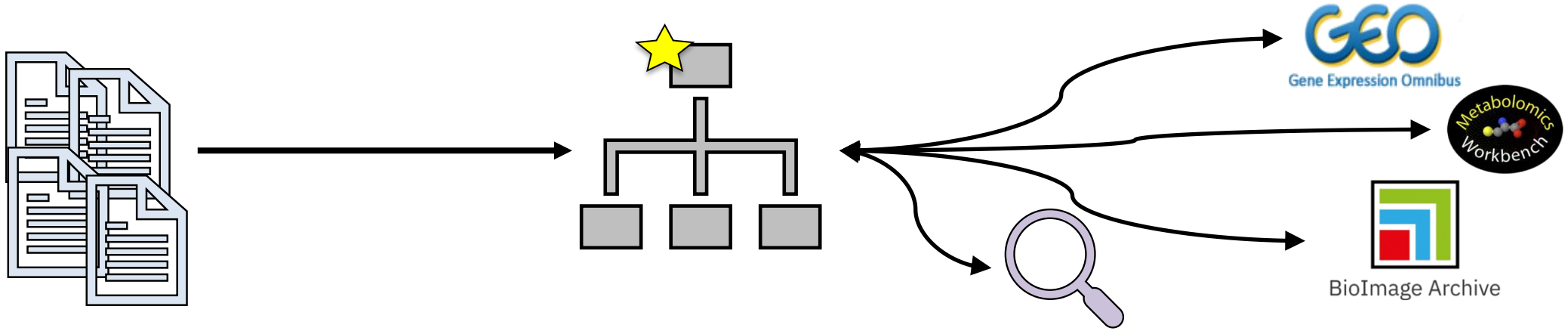


FROM THE LAB TO *FAIR*



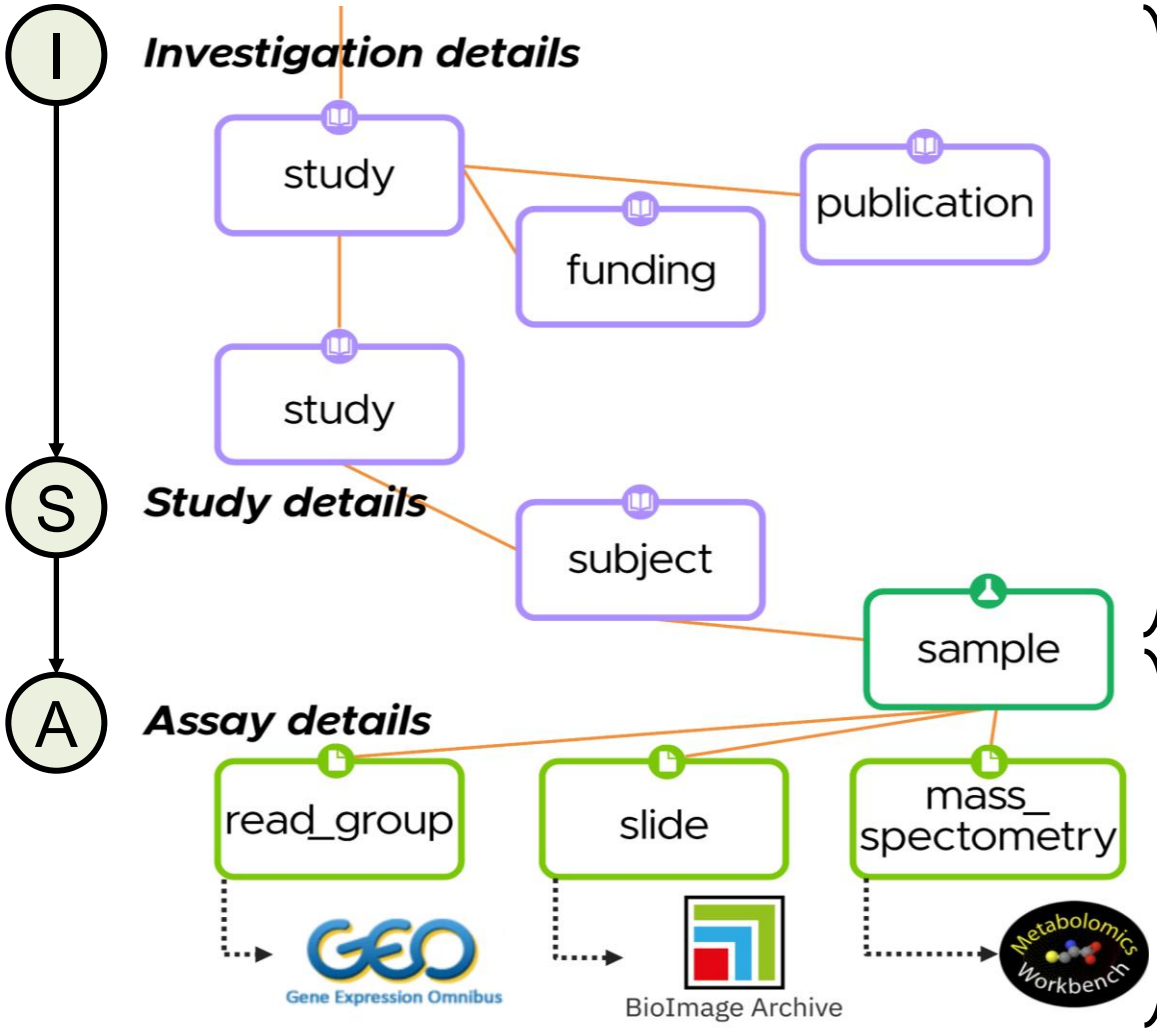
WHAT IS TOXDATACOMMONS?

ToxDataCommons is a FAIR resource aimed at helping the environmental health research community collect, share, and discover relevant AI-ready datasets.



We're note replacing existing repositories, but allowing researchers to add value to their public datasets

UNDERLYING DATA MODEL



Metadata about the overall study, experimental design, and individual subjects (*e.g., mice, humans, geographical location, …*)

Based on MIATE right now

Extendable: Based on endpoints

MINSEQE, REMBI, CIMR, …

UNDERLYING DATA MODEL

administrative JSON TSV Close

subject description of subject.

sex	• male • female • Not applicable • Missing • Not collected • Not provided • Restricted access	No	No Description
start_date	• string	No	The date the subject entered the study, which may differ from experiment_start_date (e.g., for acclimation).
start_date_age	• number	No	Age of the subject at the start of the study, in days. Used to normalize for developmental stage.
strain	• Not applicable • Missing • Not collected • Not provided • Restricted access • C57BL/6NCrl • C57BL/6J • C57BL/6NCrl_PkmΔDRE	No	No Description

Lists all expected metadata for a subject

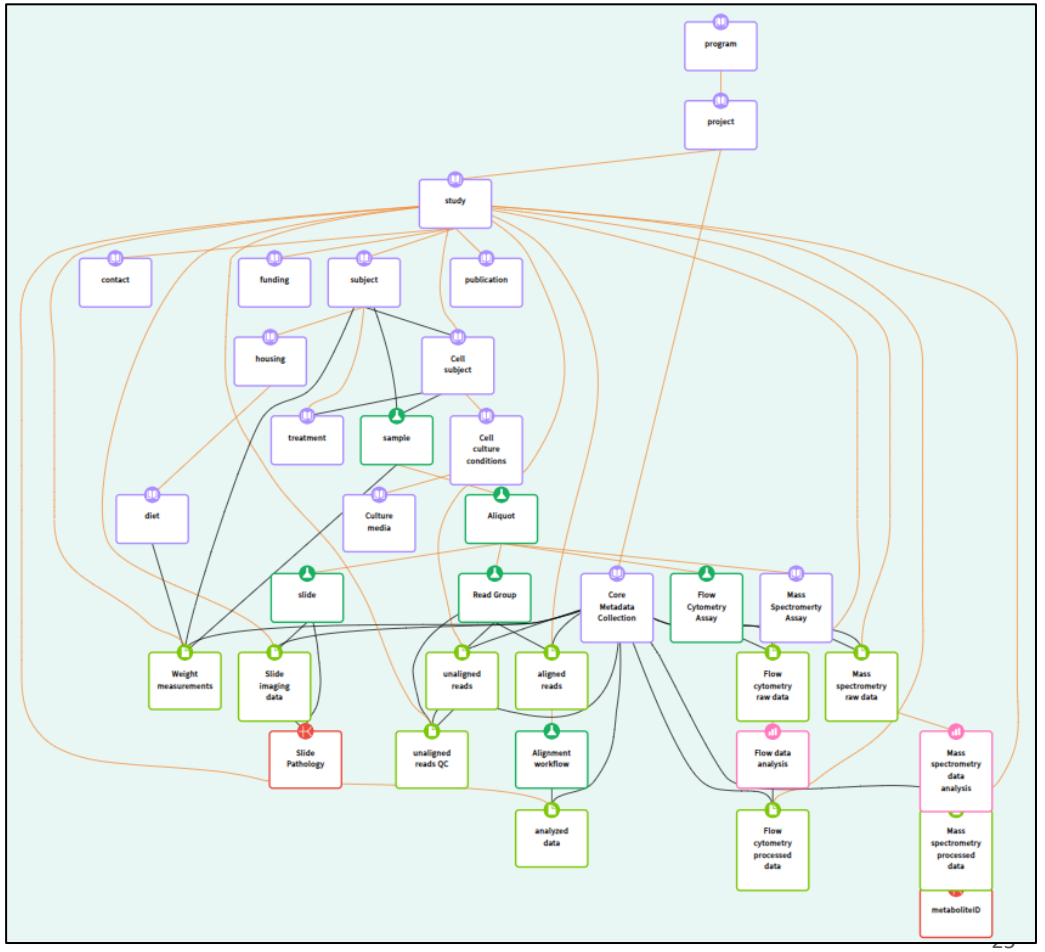
Prevents typos (*e.g.*, BABL/C vs. BALB/C)

UNDERLYING DATA MODEL

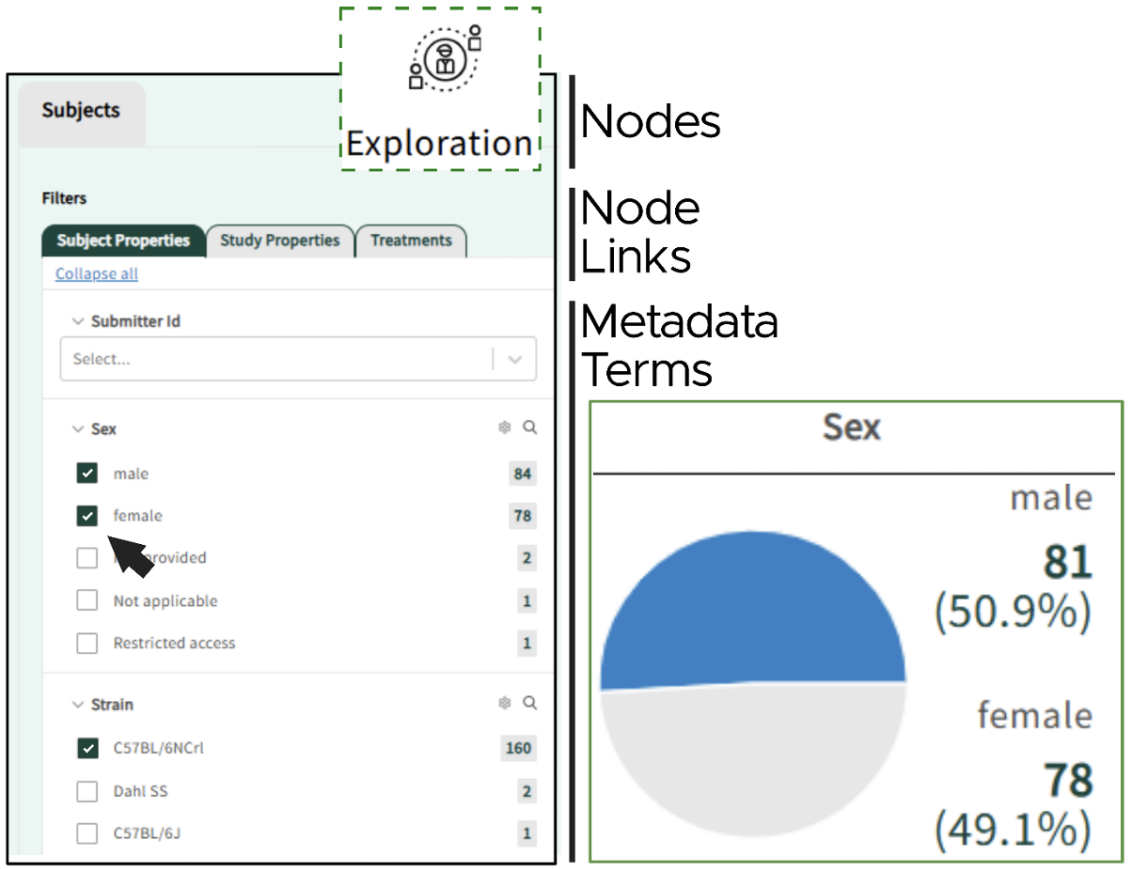
Our data model is evolving, we need user input for prioritization.

In progress:

- Addition of pup – dam/sire relationships
- Integrating a preliminary environmental sampling module
- Integrating a community survey module



IMPACT OF ROBUST METADATA



Data can now be explored in a very structured manner knowing exactly what can the *'rules'* are.

No need to search multiple terms:

("TCDD" OR "Dioxin" OR "2,3,7,8-tetrachlorodibenzo-p-dioxin") AND ("liver" OR "hepatic" OR "hepatocyte") …

Both a user interface and API can be used to find these data.

BUILDING COHORTS OF DATASETS

A citable persistent ID (PID) is assigned to each dataset

Project Id	Source Node	study_id	File Name	File Size	GUID
TRAINING-training001	slide_image	study_test001	PPGTEST_PPG001_MAP.txt	5.66 KB	dg.TDC/25d854ec-bfbd-4e85-be11-686f0d8d9d8b
TRAINING-training001	slide_image	study_test001	test.txt	9 B	dg.TDC/169cb934-9939-4157-b36a-e9b6a434af25
TRAINING-training001	slide_image	study_test001	PRJ129_bodyweights.txt	12.58 KB	dg.TDC/54ad8253-dfa7-4eb5-a549-f0fe7ce21337
TRAINING-training001	slide_image	study_test001	PRJ129_sampleweights.txt	5.15 KB	dg.TDC/98481fc5-d300-47c7-a0e3-ff86aaf979a6
TRAINING-training001	slide_image	study_test001	test.txt	9 B	dg.TDC/4f04343d-5bfa-46b6-b025-003284753d63
TRAINING-training001	slide_image	study_test001	PPGTEST_PPG001_SBP.txt	5.6 KB	dg.TDC/93277b6a-f6eb-4a71-8843-b743d8671e64

Previous Page 1 of 1 20 rows Next

WHY WE CREATED TOXDATACOMMONS

We found a gap between existing infrastructure and the FAIR sharing of *all* research data.

Existing resources are siloed. Mapping expression data to metabolomic data, imaging data, weight data… was extremely difficult.

Gen3 commons showed a lot of promise

PROBLEM SOLVED!

Image created with ChatGPT (DALL·E), 2025.

PROBLEM SOLVED!

Image created with ChatGPT (DALL·E), 2025.

MAKING METADATA COLLECTION ACCESSIBLE

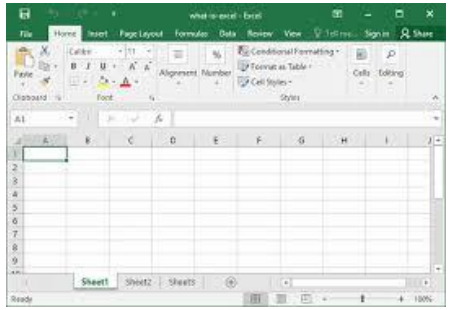
The research community has aimed to make metadata collection more accessible for years through ***Templatization***:

- Who creates the templates?
- Who has access to the templates?
- Do we use web forms or spreadsheets?
- …

cidgoh/
DataHarmonizer

isatools

CEDAR



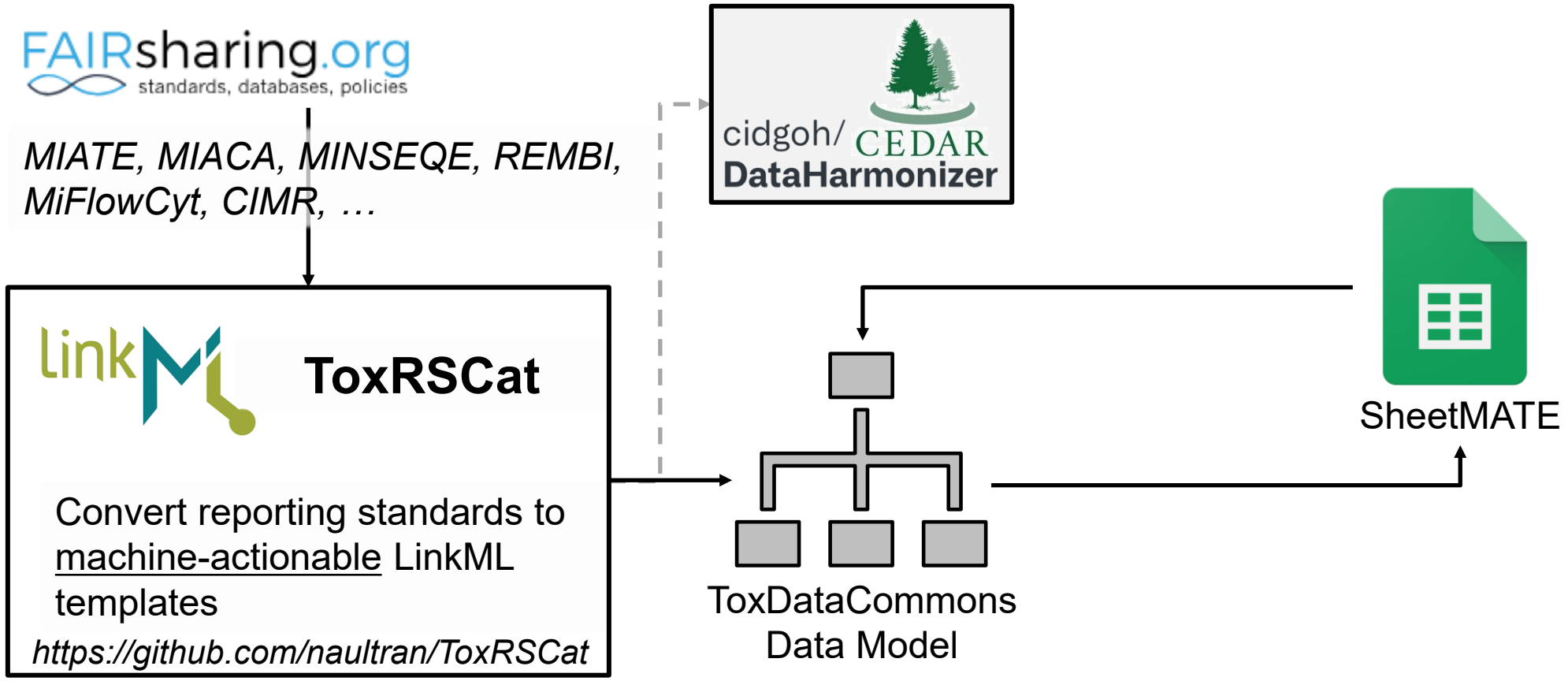
MAKING METADATA COLLECTION ACCESSIBLE

The research community has aimed to make metadata collection more accessible for years through ***Templatization***:

- Who creates the templates?
- Who has access to the templates?
- Do we use web forms or spreadsheets?
- …

- **Domain experts**
- **Everyone/Anyone**
- **Spreadsheets**
- **Machine-actionable**

BUILDING A COMMUNITY RESOURCE

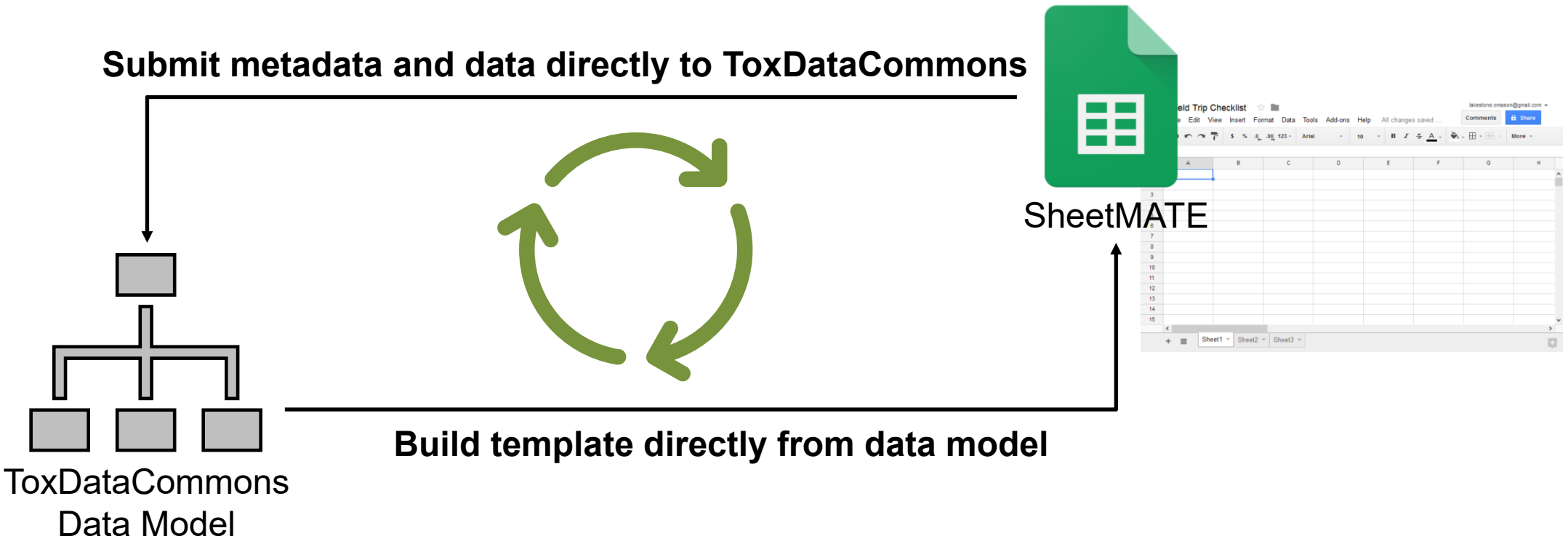


WHY SHEETMATE?

SheetMATE was primarily driven by user experience:

- **ISACreator:** Tool was retired and no longer maintained. Template format was machine-actionable but not widely used.
- **DataHarmonizer:** Builds templates from LinkML but the spreadsheet functionality often felt limiting compared to Excel/Sheets.
- **CEDAR Workbench:** Web form format. Templates built and shared by users (can vary a lot). Difficult to do batch collections.
- None of the tools could be directly submitted to ToxDataCommons

SHEETMATE: CLOSING THE LOOP



TOXDATACOMMONS IN ACTION: DATA UPLOAD

ToxDataCommons holds all the metadata. Can we use it to improve data submissions to established repositories?

Without using ToxDataCommons

Treatment ID:	TR003988
Treatment Summary:	TCDD elicits similar effec heightened sensitivity to is right-shifted). To obser day (PND) 25 C57BL/6N Innovive Innocages (San and a 12 h/12 light/dark c (Sigma-Aldrich, St. Louis total of 7 administered dc study duration as opposit lives of TCDD in humans

Treatment ID:	TR003613
Treatment Summary:	Beginning on post-na 58, mice were euthar
Treatment Compound:	2,3,7,8-tetrachlorodib
Treatment Route:	Oral gavage
Treatment Dose:	0, 1, 3, 10, or 30 micı
Treatment Dosevolume:	0.1 milliliters
Treatment Doseduration:	28 days
Treatment Vehicle:	sesame oil

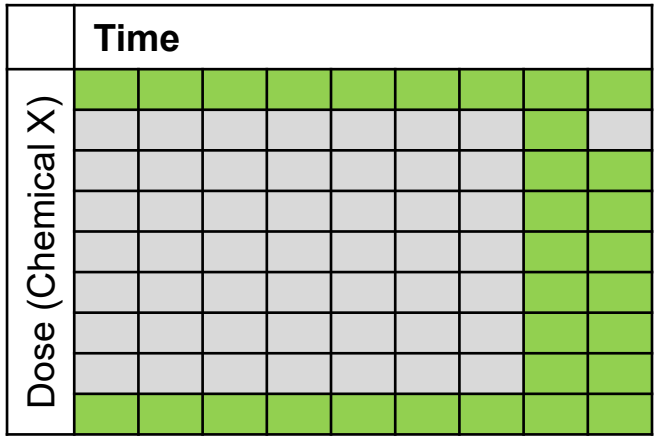
UsingToxDataCommons

Treatment ID:	TR003667
Treatment Summary:	Mice were orally ga administrations.
Treatment Compound:	DTXSID2021315:
Treatment Route:	Oral Gavage Route
Treatment Dose:	['0.0', '0.01', '0.03',
Treatment Dosevolume:	0.1 mL
Treatment Vehicle:	DTXSID9033971
Animal Fasting:	ad libitum
Animal Endp Euthanasia:	Carbon dioxide asp
Animal Endp Tissue Coll List:	Liver

Treatment ID:	TR002872
Treatment Summary:	Mice were orally ga administrations.
Treatment Compound:	DTXSID4051378:
Treatment Route:	Oral Gavage Route
Treatment Dose:	['0.0', '0.03', '0.1', '(
Treatment Dosevolume:	0.1 mL
Treatment Vehicle:	DTXSID9033971:
Animal Fasting:	ad libitum
Animal Endp Euthanasia:	Carbon dioxide asp
Animal Endp Tissue Coll List:	Liver

TOXDATACOMMONS IN ACTION: INTEGRATION

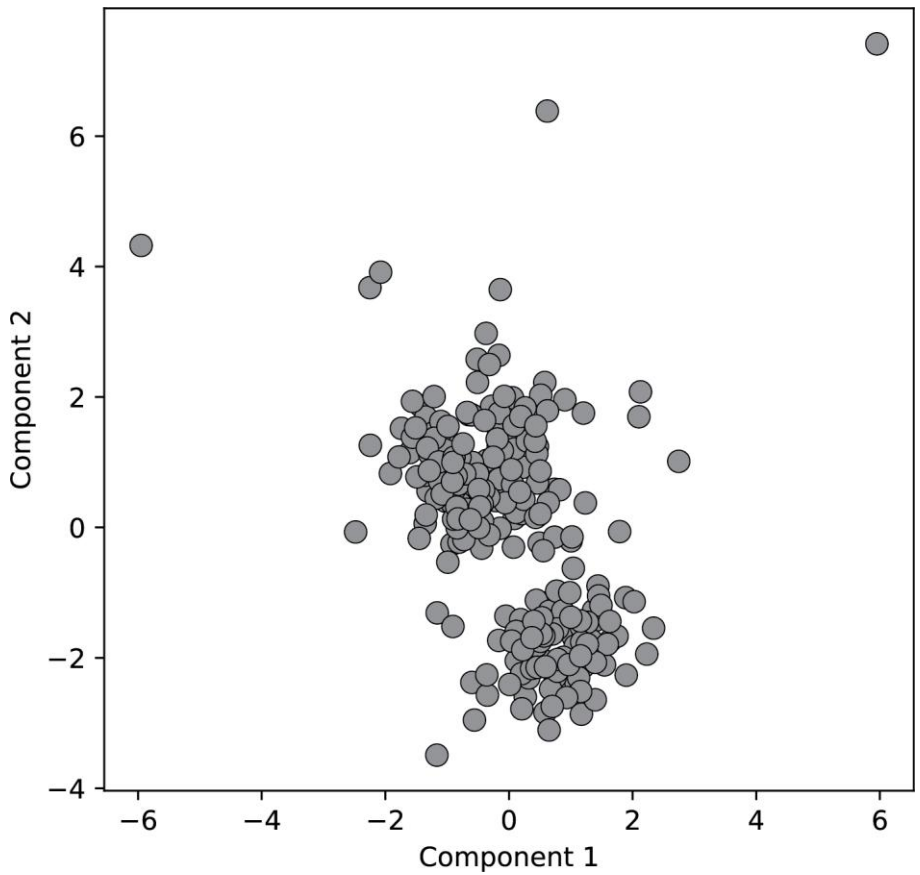
Datasets were generated in at least 3 independent studies with several measured endpoints (weights, pathology, gene expression, …).

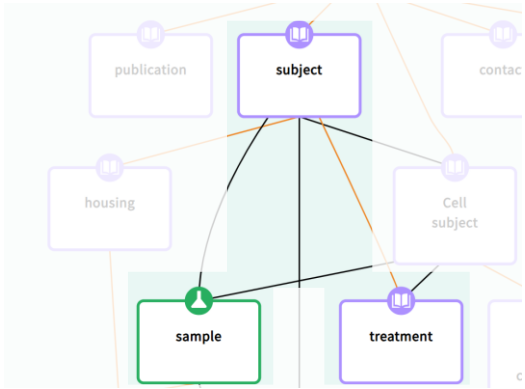


Can we identify informative patterns about chemical toxicity by integrating these studies?

TOXDATACOMMONS IN ACTION: INTEGRATION

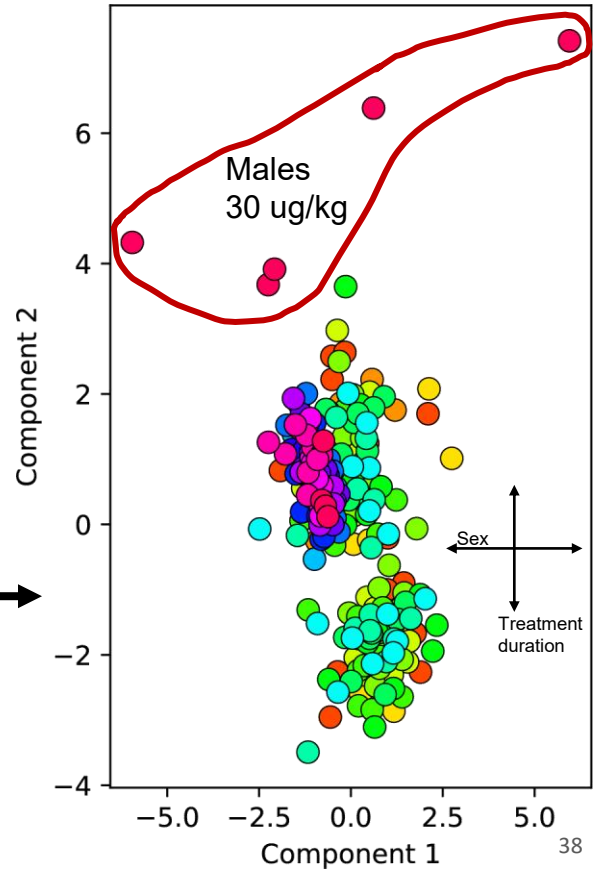
Dynamic time warping analysis of daily body weights of each *subject*.





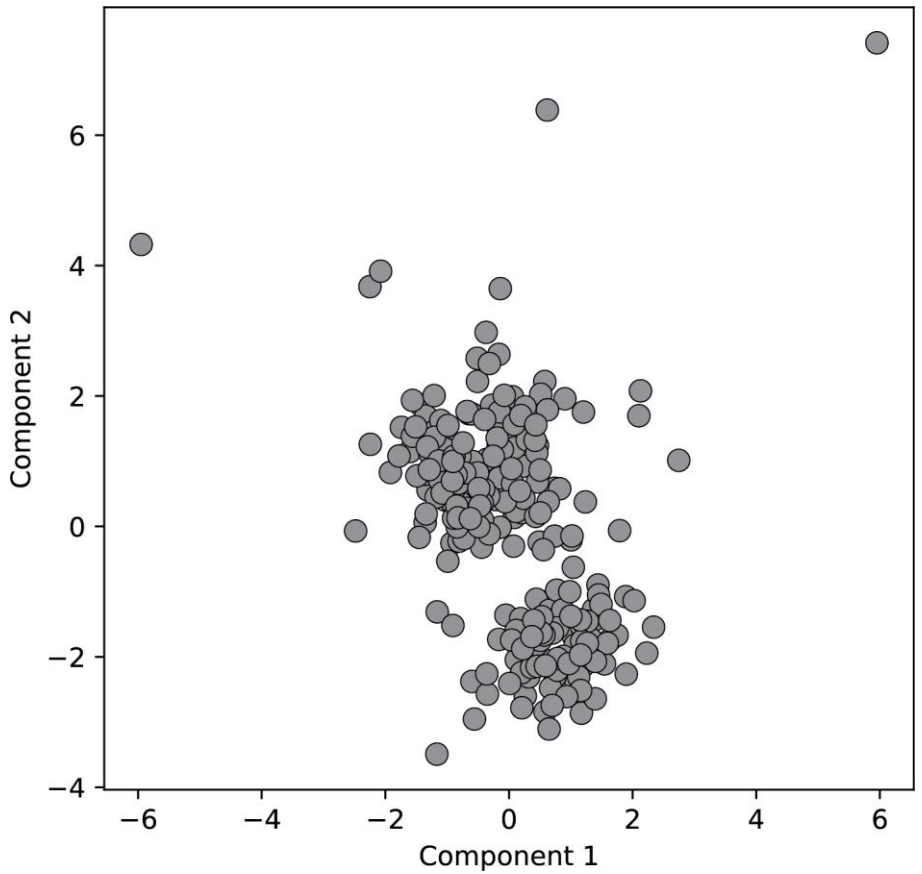
Sex
Dose
…

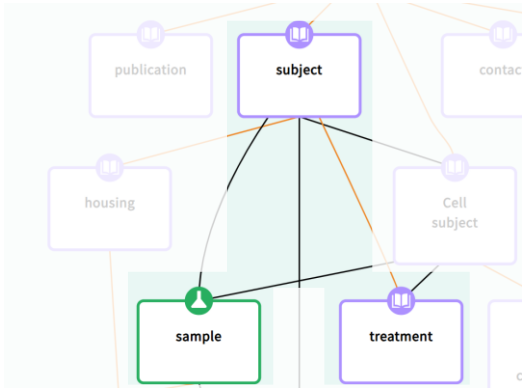
Metadata annotation reveals clustering



TOXDATACOMMONS IN ACTION: INTEGRATION

Dynamic Time Warping analysis of daily body weights of each *subject*.

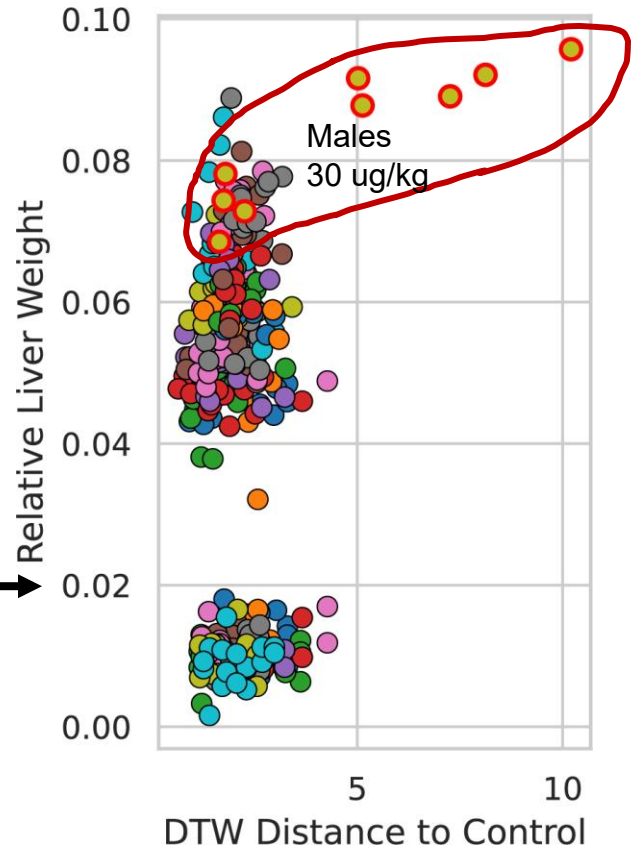




Sex
Dose
…

Relative liver weight (*sample*)

Metadata annotation reveals clustering



SUMMARY

- Improved data sharing has the potential to benefit the entire environmental health research community
- The MSU SRC has tackled this by building infrastructure and resources. *It is only a single piece of the puzzle!*
- **Collaboration**, feedback, ideas, and transparency regarding challenges will help us move forward more effectively.

Image created with ChatGPT (DALL·E), 2025.

ToxDataCommons VISION

MSU Superfund Research Program

30+ Investigators across many departments and colleges

2025

MSU EHS Researchers

500+ Investigators across many departments and and colleges

2026

External Data

30+ Superfund Research Centers and even more EHS researchers

2028

ACKNOWLEDGEMENTS

MSU Superfund Research Center
*Keji Yuan
Tim Zacharewski
Giovan Cholico
Eric Kasten
Jonathan Babbage
Todd Hall

CTDS:
Chris Meyer
Ed Malinowski
Jawad Qureshi

KRUMWARE

Funding:
NIEHS SRP P42 ES004911

NIH National Institute of Environmental Health Sciences
Superfund Research Program

QESTIONS?