Tech Talk

Using Data Systems to Increase Accessibility in Disease Specific Research

July 26, 2023

Megan Carnes, PhD
Alex Harding, MS
Ravi Mathur, PhD

RTI INTERNATIONAL

Presenting Today

Megan Ulmer Carnes, PhD
Role: ME/CFS Network mPI
Technical Expertise: Genomics, Genetic Epidemiology, Microbiome, Bioinformatics

Alex Harding, MS
Role: map-systems Project Manager and Developer
Technical Expertise: Full-stack Web Applications, Data Visualization Dashboards, Cloud Infrastructure Management, Data Processing, and Software Architecture

Ravi Mathur, PhD
Role: map-systems Lead Bioinformaticist
Technical Expertise: Bioinformatics, Genetics, Metabolomics, Proteomics, Data Integration, Multi-Omics Analysis

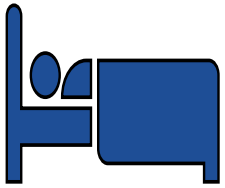
Presentation Overview

- Introduction to ME/CFS: Megan Carnes
  - mapMECFS tool
- Open system framework: Alex Harding
- New portal development: Ravi Mathur
  - E.g., mapOA

ME/CFS – A complex, multi-factorial disease

ME/CFS

- Serious, long-term illness
- Affects many body systems
- Often limits people from doing their usual activities



Symptoms

- Severe fatigue
- Sleep problems
- Confined to bed
- Pain, dizziness, and difficulty with memory and cognition

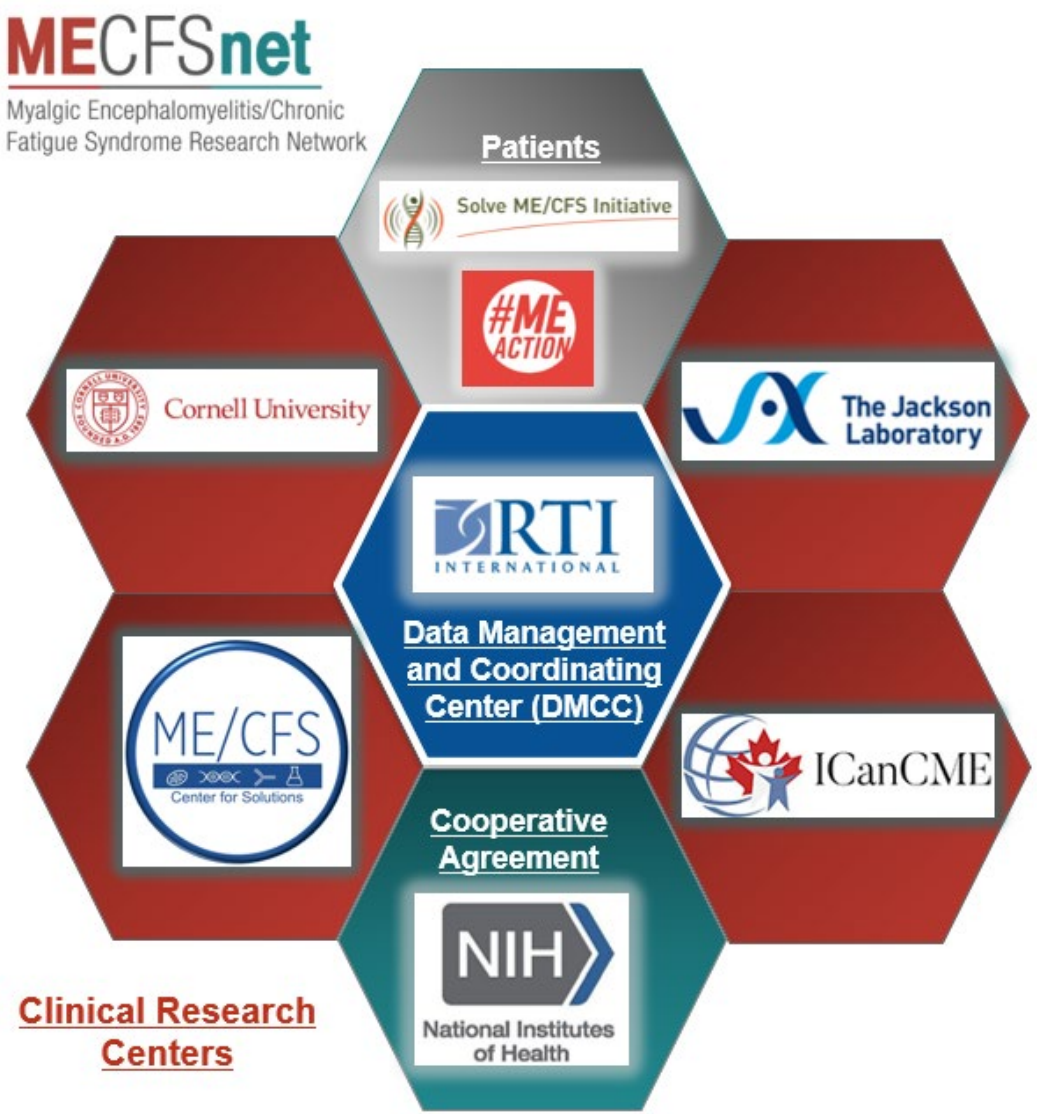
Post-Exertional Malaise (PEM)

- Worsening of symptoms following mental or physical activity

How many?

- Institute of Medicine estimates 836,000 to 2.5 million Americans live with ME/CFS
- Most have not been diagnosed

Network Research



One of the Data Management and Coordinating Center's (DMCC) Goals: Build infrastructure to support secure sharing of data across a wide range of biological and clinical experiments.

mapMECFS

ME/CFS-focused data repository

https://www.mapmecfs.org/

The Need for Data Integration

What's needed in order to gain a better perspective

- Inventory of the range of data being collected
- An understanding of where other investigators are looking

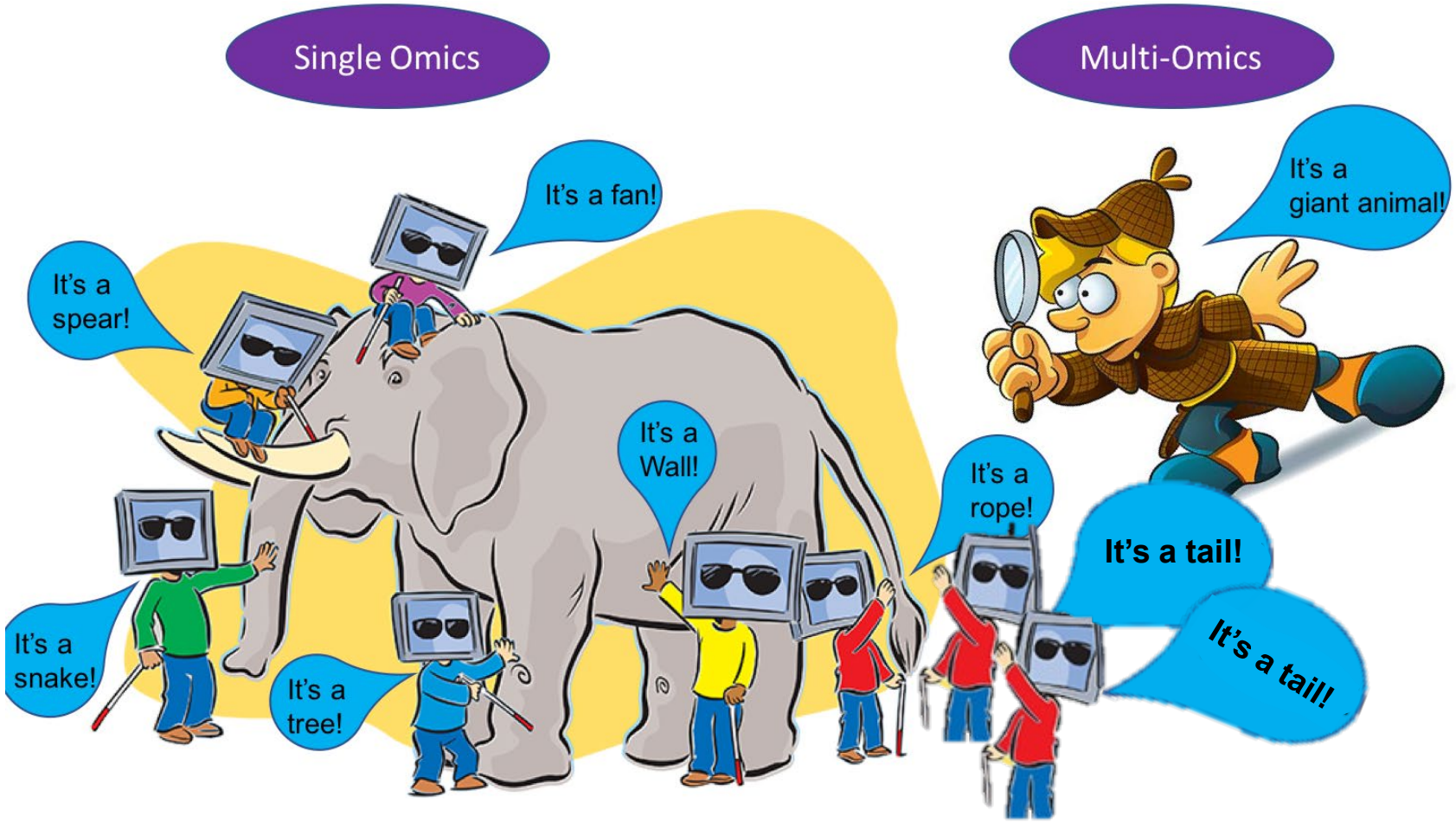
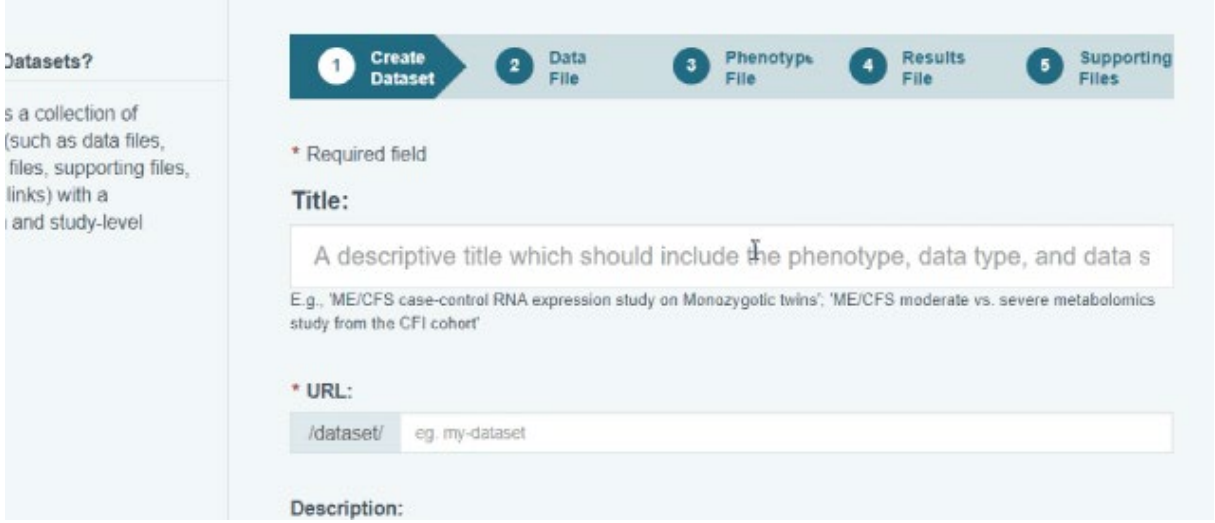


Image from: http://melgen.org/multi-omics-approach/

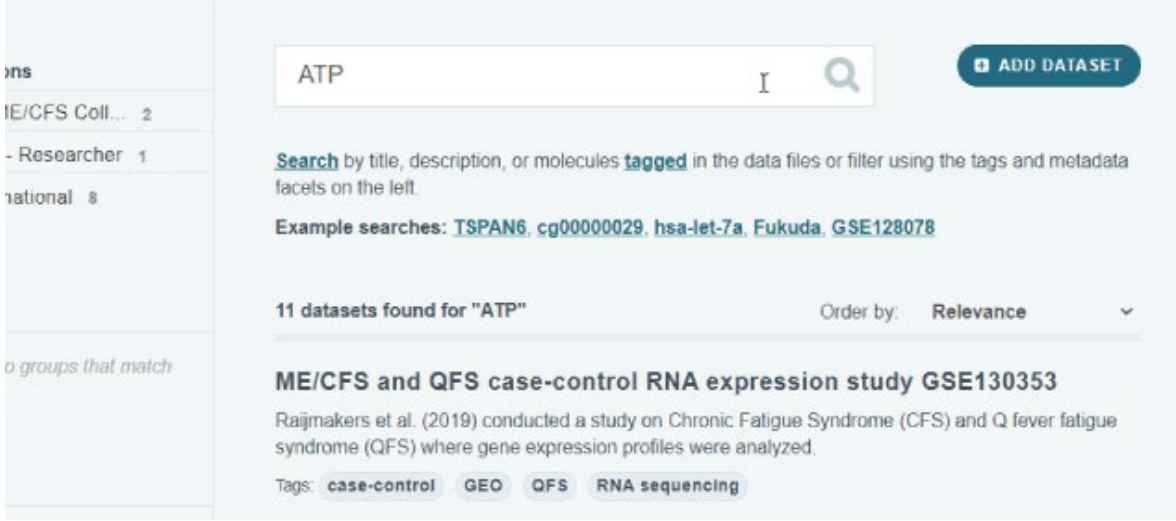
Researcher's Challenge	The mapMECFS Solution
• I do not know where to go for ME/CFS-specific data.	• mapMECFS is a comprehensive repository of ME/CFS-related research data
• Sharing data is time consuming because my data files are complex and variable.	• Upload data using an easy step-by-step form with pre-populated fields • Allows flexible file structures and data types
• I would like to compare my results other studies to draw biological conclusions.	• Contains custom search tools and curated ME/CFS literature to enable quick, cross-study comparisons

One Stop Shop for ME/CFS Research

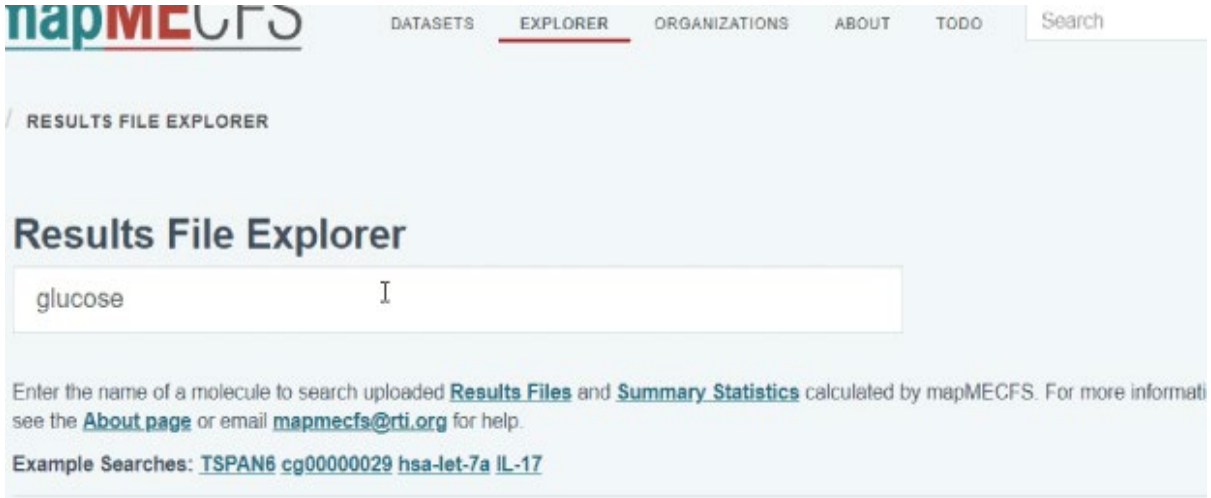
Share New Data



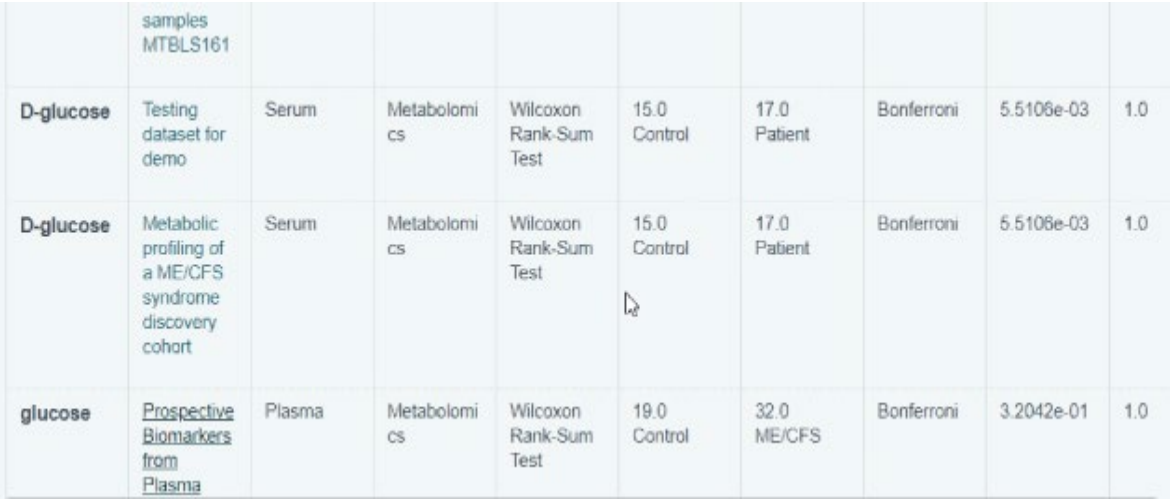
Search Existing Datasets



Compare Study Results

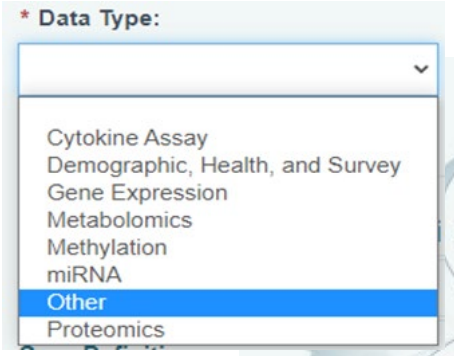


Download Relevant Files



	samples MTBLS161								
D-glucose	Testing dataset for demo	Serum	Metabolomics	Wilcoxon Rank-Sum Test	15.0 Control	17.0 Patient	Bonferroni	5.5106e-03	1.0
D-glucose	Metabolic profiling of a ME/CFS syndrome discovery cohort	Serum	Metabolomics	Wilcoxon Rank-Sum Test	15.0 Control	17.0 Patient	Bonferroni	5.5106e-03	1.0
glucose	Prospective Biomarkers from Plasma	Plasma	Metabolomics	Wilcoxon Rank-Sum Test	19.0 Control	32.0 ME/CFS	Bonferroni	3.2042e-01	1.0

Site Contents



61 PUBLIC DATASETS

>300 RESULT FILES

119 SITE USERS

https://www.mapmecfs.org/

mapMECFS Key Features: Search/Dataset

IL6

Search by title, description, or molecules tagged in the data files or filter using the tags and metadata facets on the left.

Example searches: TSPAN6, cg00000029, hsa-let-7a, Fukuda, GSE128078

7 datasets found for "IL6" Order by: **Relevance**

ME/CFS case-control RNA expression study (GPL96) GSE14577

Gow et al. (2009) conducted a case-control study with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) on samples (8 ME/CFS and 7 healthy controls) on two...

Tags: GEO has data file has phenotype file

ME/CFS case-control RNA expression study (GPL97) GSE14577

Gow et al. (2009) conducted a case control study on myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) on a of 15 samples (8 ME/CFS and 7 health controls) on two...

Tags: GEO has data file has phenotype file

ME/CFS and IFS case-control RNA expression study on Monozygotic twins GSE16059

Byrnes et al. (2009) conducted a study on myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) and Idiopathic Chronic Fatigue (ICF) where RNA expression profiles on

Files

Data File	**miRNA profiling data table** TXT gse70371_matrix_data_wmirbase_uploaded.txt The normalized signal intensity of miRNA generated from the Ambion Bioarray V1. *111 KiB*	Preview Download Edit
Phenotype File	**Phenotypes and covariates for Petty, et al.** TXT gse70371_phenotype_uploaded.txt Information on phenotypes and covariates for the Petty, et al. (2016) study. *2.1 KiB*	Preview Download Edit
Supporting File	**GEO Public Data Use Agreement** HTML Data use agreement for this public dataset which has been extracted from the...	More information Go to resource Edit
Supporting File	**Data Generation Summary** PDF gse70371_sop_v2.0.pdf The data generation summary for this study includes array information, study... *178 KiB*	More information Download Edit
Results File	**Example_Result_File** TSV me-cfs-case-control-mirna-expression-study-gse70371-summary-stats-2020-04-22.tsv Calculated summary statistics from mapMECFS as an example result file. N... *12.5 KiB*	Preview Download Edit

Result File

- Results from experiment-specific analysis generally containing p-values or adjusted p-values

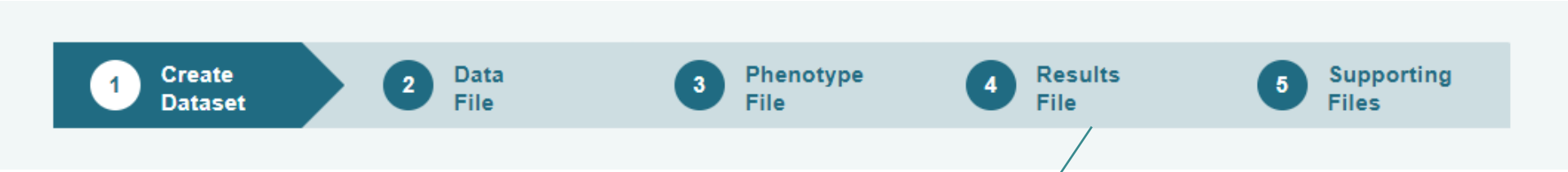


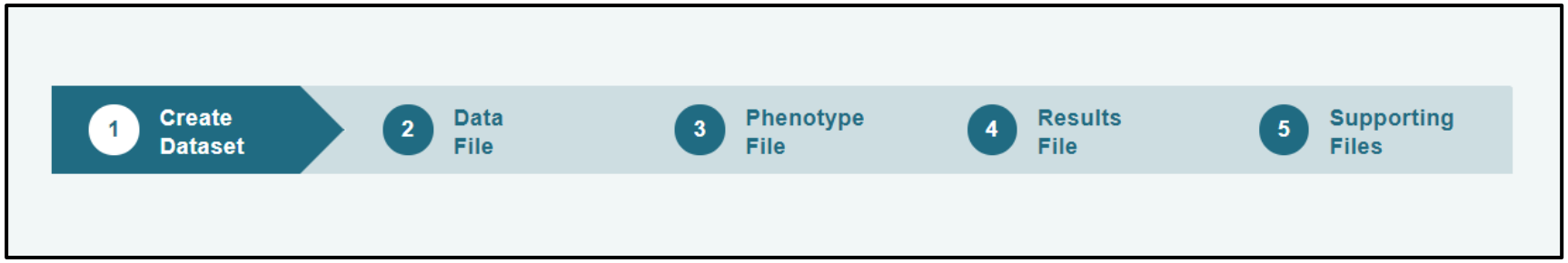
Table 2 from manuscript

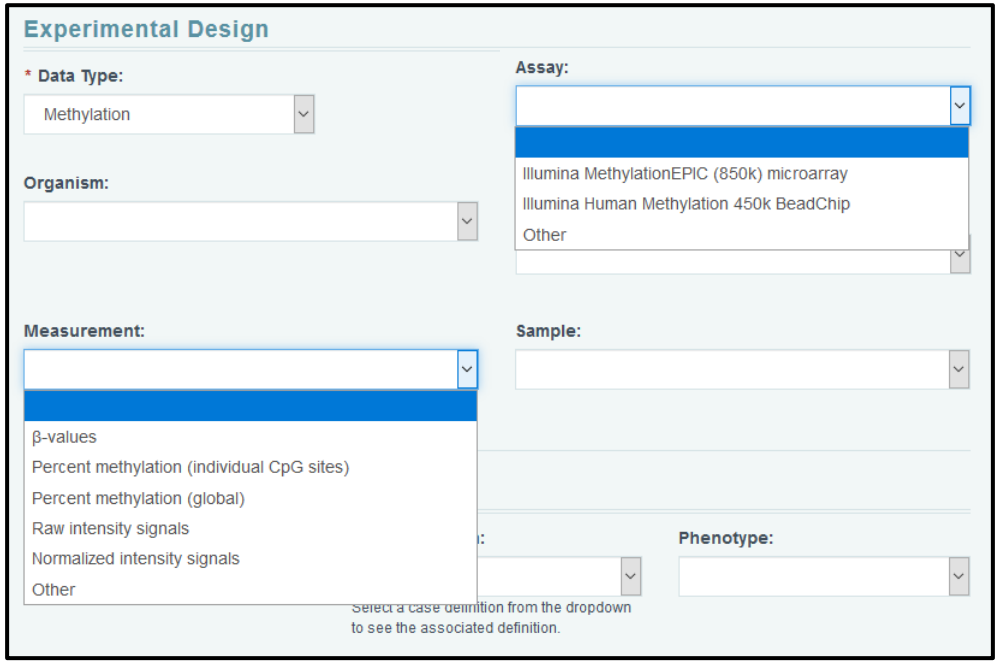
Molecule	mean (CS)	sd (CS)	mean (CL)	sd (CL)	pvalue
IL6	0.36	0.46	0.53	0.51	0.245
IL7	0.85	0.29	0.89	0.2	0.657
CXCL8	5.4	2.07	4.56	1.72	0.217
IL10	0.49	0.11	0.43	0.1	0.116
IL12p40	0.2	0.34	0.09	0.17	0.217
IL12p70	0.59	0.48	0.45	0.14	0.351
IL13	0.42	0.18	0.29	0.21	0.057
IL15	0.95	0.44	0.77	0.32	0.196

- Molecule column is searchable on mapMECFS
- Benefit is improved findability over a search in PubMed
- Custom tool (Result File Explorer) allows for comparisons across studies

Results of t-tests of cytokine/chemokine levels comparing classical, short-duration ME/CFS cases to classical, long-duration ME/CFS cases (Hornig M, et al. Immune Transl Psychiatry. 2017)

mapMECFS Key Features: Easy Upload / Metadata





Custom Tools and Features to be Demoed

Molecule tagging

Uploaded data are tagged with known synonyms to improve searchability.

2 records

IL17 Go »

« 1 – 100 »

_id	Molecule	Synonym 1	Synonym 2	Synonym 3	Synonym 4	Synonym 5	Synony...	Synony...	Synony...	Synony...	Synony...	Synon
2236	IL17A	CTLA-8	CTLA8	IL-17	IL-17A	IL17						
12510	IL17RC	CANDF9	IL17-RL	IL17RL								

Supporting File

Data Generation Summary PDF

A description of the standard operating procedures for this study including...

119.6 KiB

More information
Download
Edit

Supporting File

Raw Data Availability

The link to access the raw ata fom this RNA sequencing experiment. Th aw data...

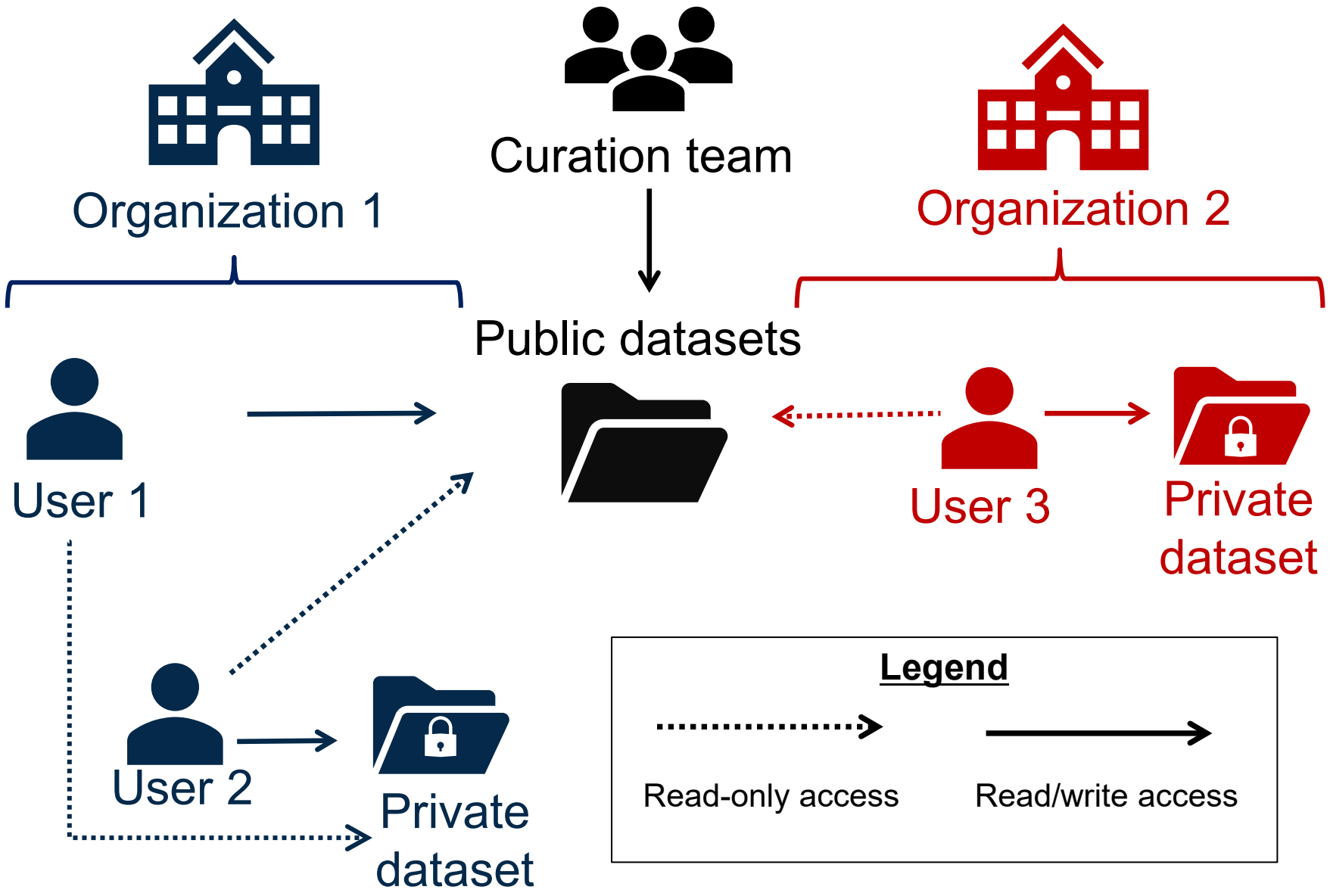
More information
Go to resource
Edit

Add new file

View Additional Search Terms

View Summary Statistics

Website Structure



- New users must be approved by NIH and agree to the DUA
- Uploaded data defaults to Private
- Public requests are reviewed for personally identifiable information before release

mapMECFS

Demo

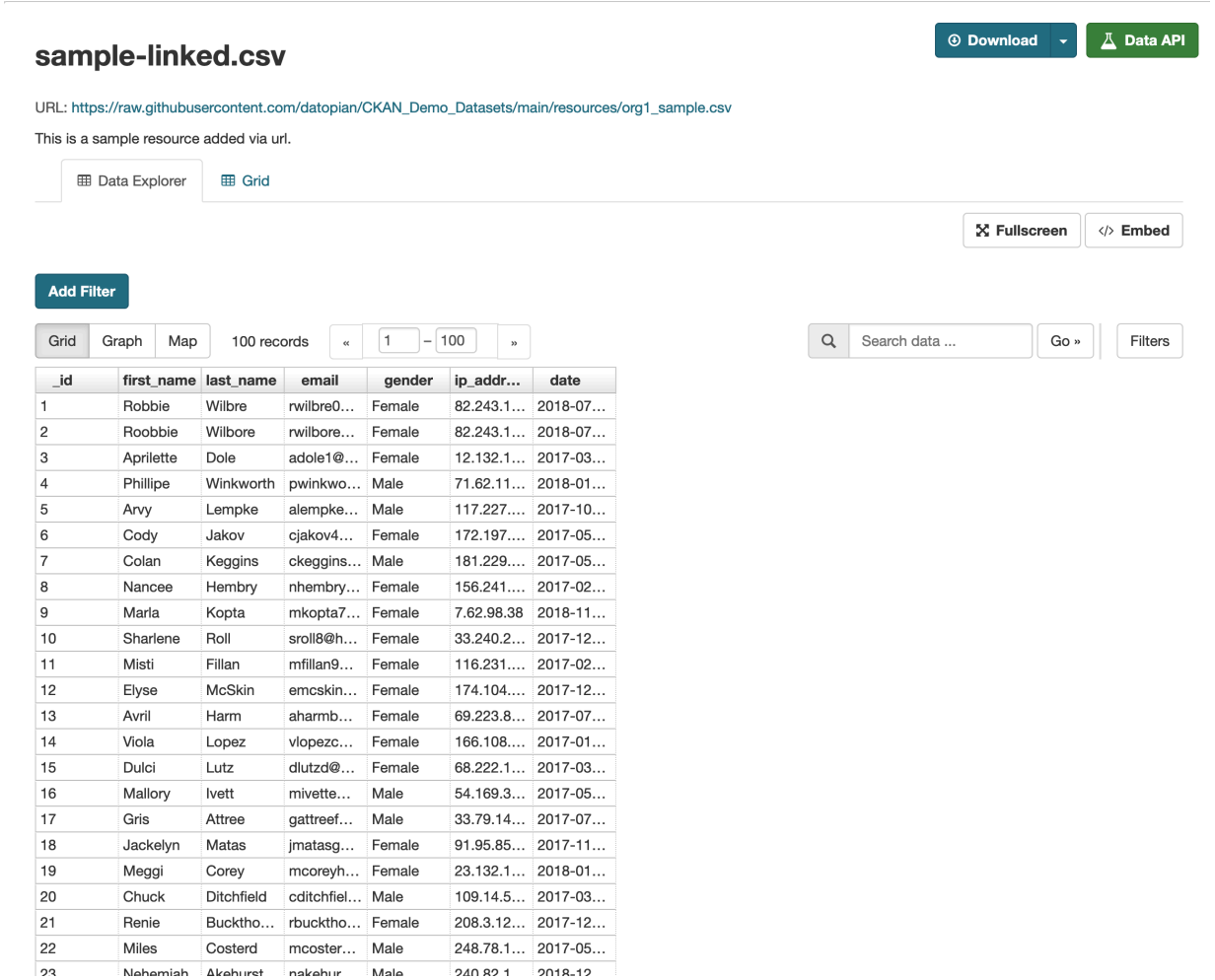
CKAN Framework

- **C**omprehensive **K**nowledge **A**rchive **N**etwork
- Open-source data portal framework built in Python
  - Opinionated data storage in PostgreSQL
  - Robust CRUD (**C**reate, **R**ead, **U**pdate, **D**elete) structure with role-based user controls
  - Data REST API
  - Full-Stack application with HTML/Javascript frontend
  - Extension architecture and community of extensions
- Authored by Open Knowledge Foundation, but recently moved to bilateral community stewardship
- Used for open data portals built by the US, Canadian, Australian governments to name a few
- Including data.gov!

CKAN Overview

CKAN allows users to upload any files or web URLs as **Resources. Resources** belong to **Datasets**, which are groups of **Resources** with shared metadata. **Datasets** belong to **Organizations**, which have **members** (users) with different levels of permissions.

CKAN Overview



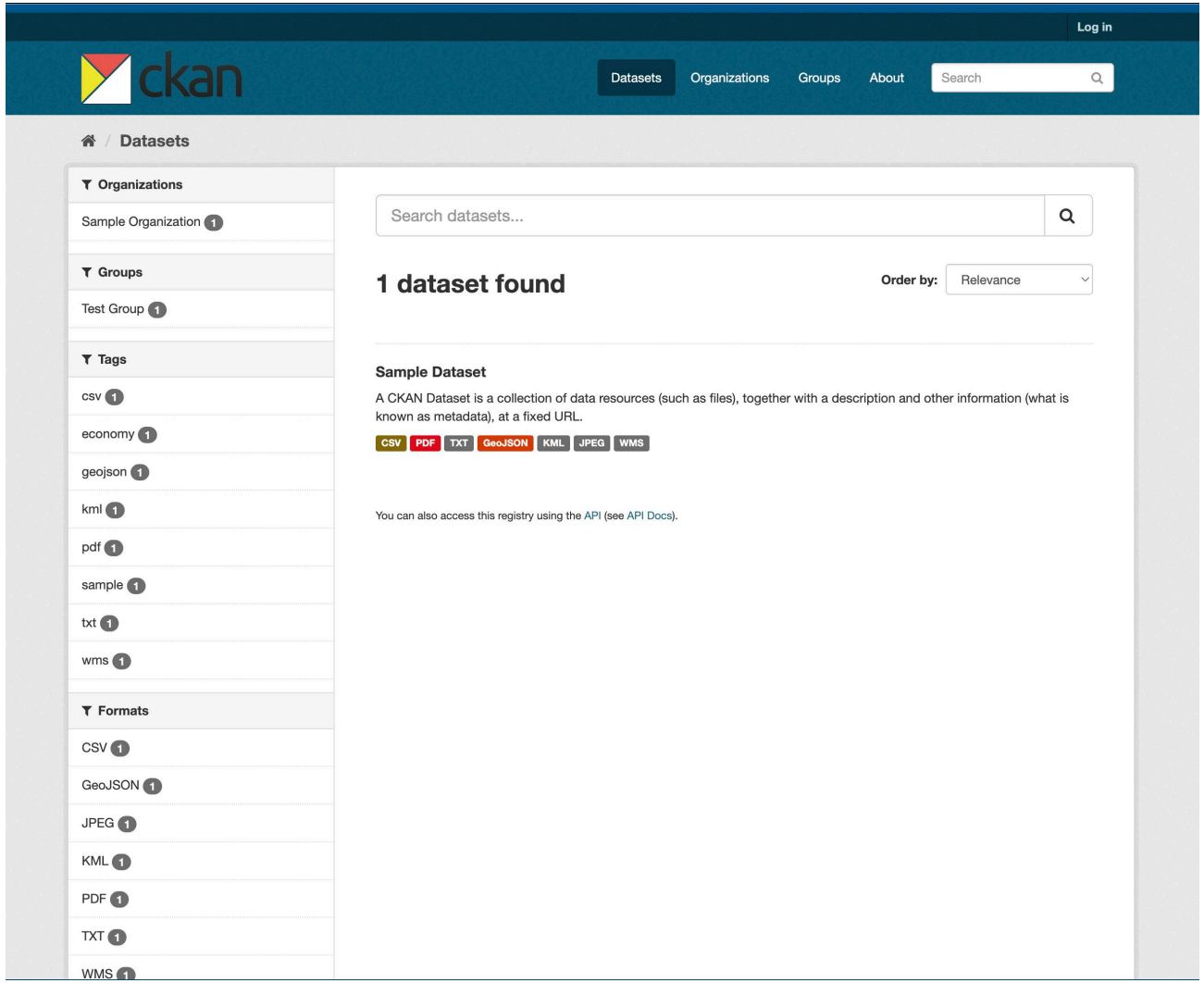
_id	first_name	last_name	email	gender	ip_addr...	date
1	Robbie	Wilbre	rwilbre0...	Female	82.243.1...	2018-07...
2	Roobbie	Wilbore	rwilbore...	Female	82.243.1...	2018-07...
3	Aprilette	Dole	adole1@...	Female	12.132.1...	2017-03...
4	Phillipe	Winkworth	pwinkwo...	Male	71.62.11...	2018-01...
5	Arvy	Lempke	alempke...	Male	117.227....	2017-10...
6	Cody	Jakov	cjakov4...	Female	172.197....	2017-05...
7	Colan	Keggins	ckeggins...	Male	181.229....	2017-05...
8	Nancee	Hembry	nhembry...	Female	156.241....	2017-02...
9	Marla	Kopta	mkopta7...	Female	7.62.98.38	2018-11...
10	Sharlene	Roll	sroll8@h...	Female	33.240.2...	2017-12...
11	Misti	Fillan	mfillan9...	Female	116.231....	2017-02...
12	Elyse	McSkin	emcskin...	Female	174.104....	2017-12...
13	Avril	Harm	aharmb...	Female	69.223.8...	2017-07...
14	Viola	Lopez	vlopezc...	Female	166.108....	2017-01...
15	Dulci	Lutz	dlutzd@...	Female	68.222.1...	2017-03...
16	Mallory	Ivett	mivette...	Male	54.169.3...	2017-05...
17	Gris	Attree	gattreef...	Male	33.79.14...	2017-07...
18	Jackelyn	Matas	jmatasg...	Female	91.95.85...	2017-11...
19	Meggi	Corey	mcoreyh...	Female	23.132.1...	2018-01...
20	Chuck	Ditchfield	cditchfiel...	Male	109.14.5...	2017-03...
21	Renie	Bucktho...	rbucktho...	Female	208.3.12...	2017-12...
22	Miles	Costerd	mcoster...	Male	248.78.1...	2017-05...
23	Nehemiah	Akehurst	nakehur...	Male	240.82.1...	2018-12...

Resources uploaded to CKAN can be processed for **preview**, including tables, charts, interactive visualization building tools, maps, and more.

CKAN Overview

CKAN provides a web-facing **search interface.**

These tools allow users to search **datasets** and **resources**, using a combination of text search, tags, and faceted search, similar to an Amazon shopping experience.



CKAN Overview

`ckan.logic.action.get.package_list`(*context, data_dict*)

Return a list of the names of the site's datasets (packages).

Parameters:

- **limit** (*int*) – if given, the list of datasets will be broken into pages of at most `limit` datasets per page and only one page will be returned at a time (optional)
- **offset** (*int*) – when `limit` is given, the offset to start returning packages from

Return type: list of strings

CKAN also provides an **HTTP REST API** for programmatic use of the data contained within, as well as for developers to interact with CKAN from advanced frontend tools for visualization and interfaces.

CKAN Overview

CKAN provides a robust **extension** API for software developers to add functionality to CKAN.

In the spirit of open-source software, many developers (including RTI) have chosen to make their work open-source and available.

	ckanext-orgdashboards	CKAN extension for creating organization dashboards.
	ckanext-disqus	Extension that adds the Disqus commenting system to CKAN.
	ckanext-googleanalytics	CKAN extension to integrate Google Analytics data into CKAN. Gives download stats on package pages, list of most popular packages, etc.
	ckanext-harvest	This extension provides a common harvesting framework for ckan extensions and adds a CLI and a WUI to CKAN to manage harvesting sources and jobs.
	ckanext-spatial	This extension contains plugins that add geospatial capabilities to CKAN.
	ckanext-realtime	CKAN plugin which makes your CKAN site into a Realtime Data Portal.
	ckanext-dataspatial	Dataspatial is a Ckan extension to provide geospatial awareness of datastore data.
	ckanext-requestdata	This extension introduces a new type of dataset in which access to data is by request.
	ckanext-orgportals	CKAN extension for creating organization portals.
	Data Solr	Datasolr is a Ckan extension to use Solr for datastore queries.
	ckanext-cas	CAS (Central Authentication Service) client extension for CKAN.
	ckanext-s3filestore	Use Amazon S3 as a filestore for CKAN.
	ckanext-c3charts	c3js based charts for CKAN.
	ckanext-cloudstorage	Implements support for resource storage against multiple popular providers via apache-libcloud (S3, Azure Storage, etc...).
	ckanext-dcat	This extension provides plugins that allow CKAN to expose and consume metadata from other catalogs using RDF documents serialized using DCAT.
	ckanext-fluent	This extension provides a way to store and return multilingul fields in CKAN datasets, resources, organizations and groups.

RTI's CKAN Extensions - Custom Tools

User-Facing

SummaryStatistics

- Generates descriptive statistics based on user supplied data
  - Available on public GitHub

SearchTerms

- Data undergoes synonym matching to improve data findability
  - Available on public GitHub

ResultExplorer

- Compiles results across datasets to compare and visualize data from multiple studies

Backend

AdvancedAuth

- Provides enhanced security features; protects data from unregistered users, initiates new user workflow, sharing public data request workflow
  - Available on public GitHub

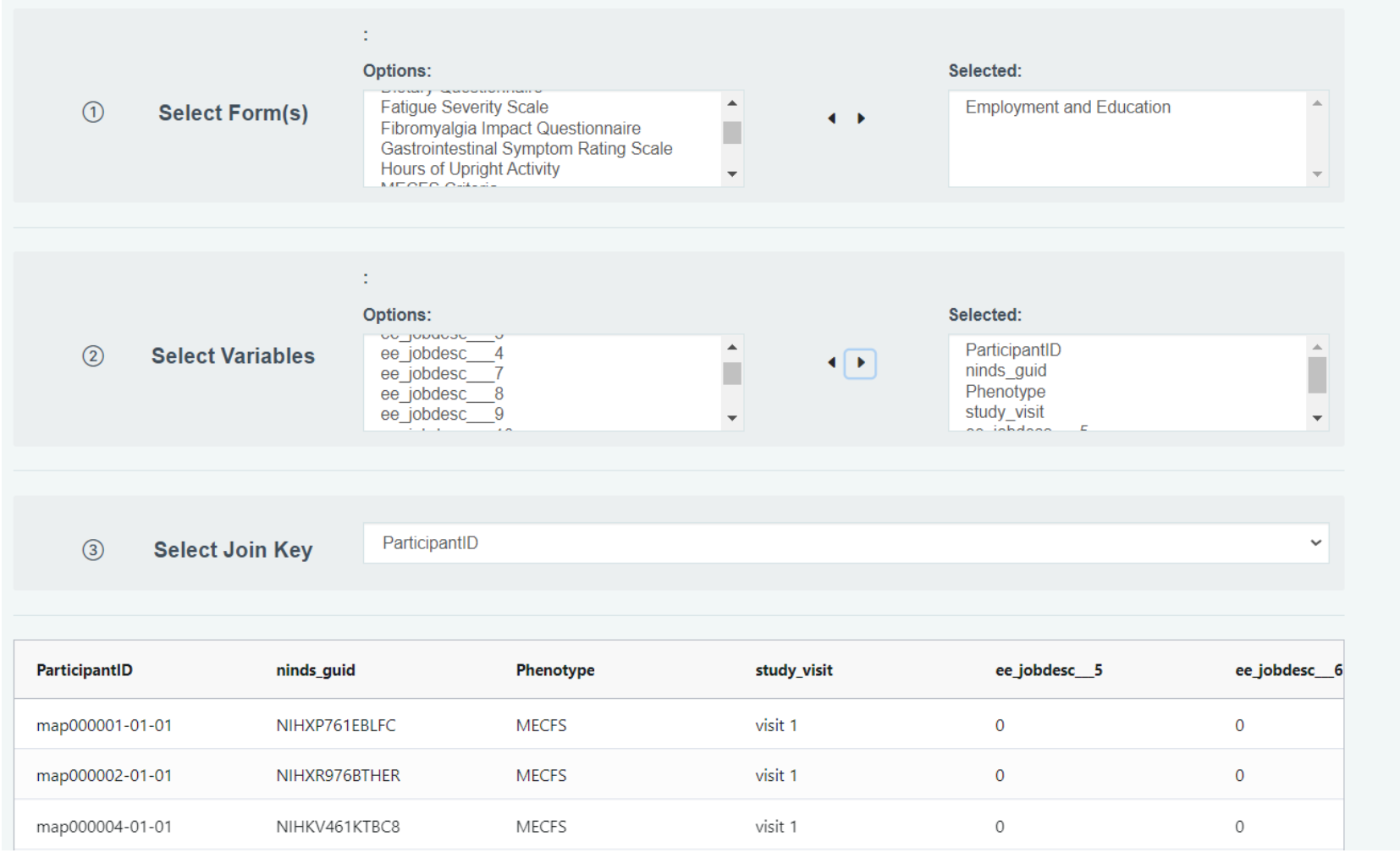
AuditExplorer

- Data access logs; enables quick response to data security and data quality issues

QAChecker

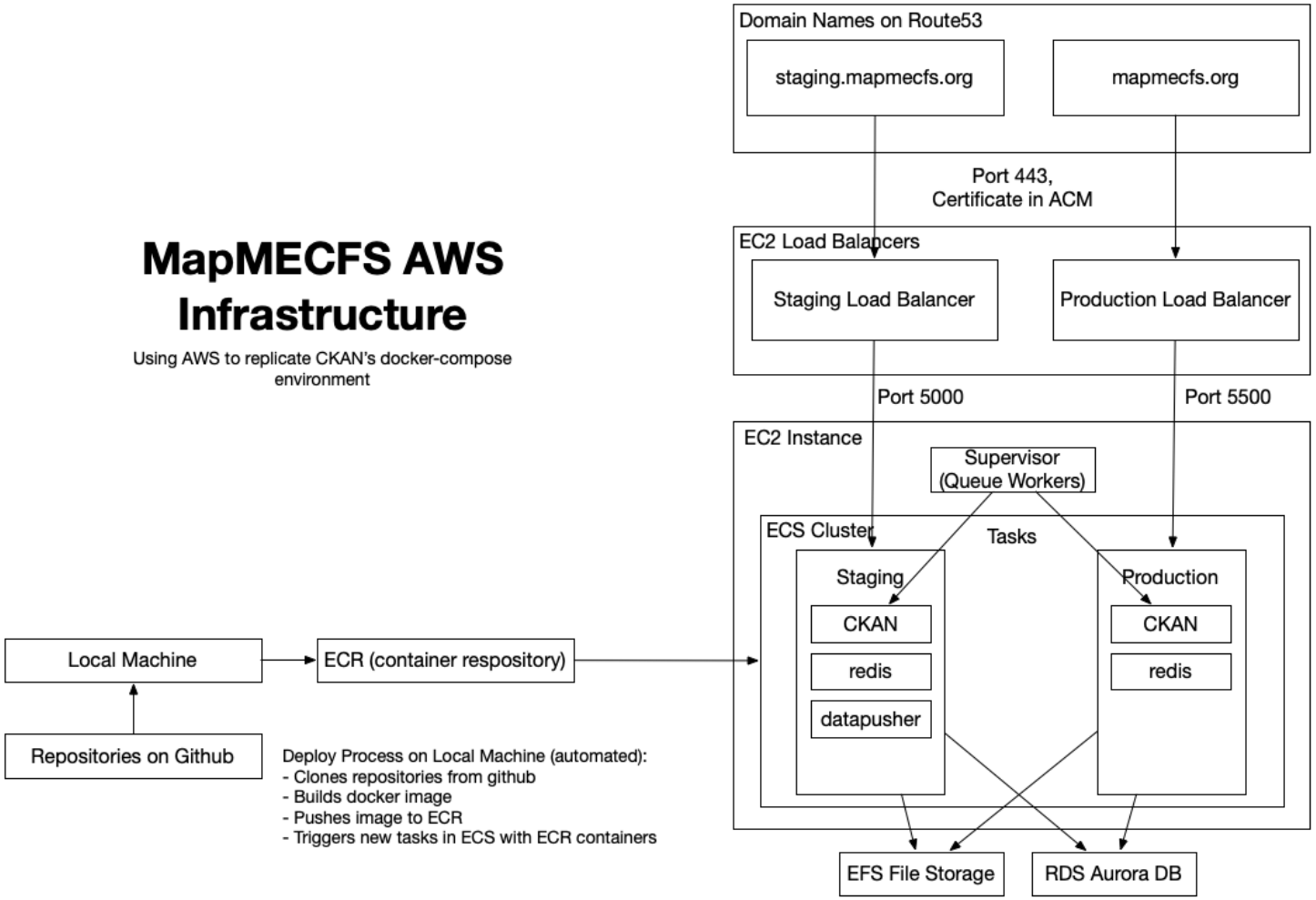
- Automated checking of datasets to easily identify datasets with processing errors

RTI's New Data Integration Tool

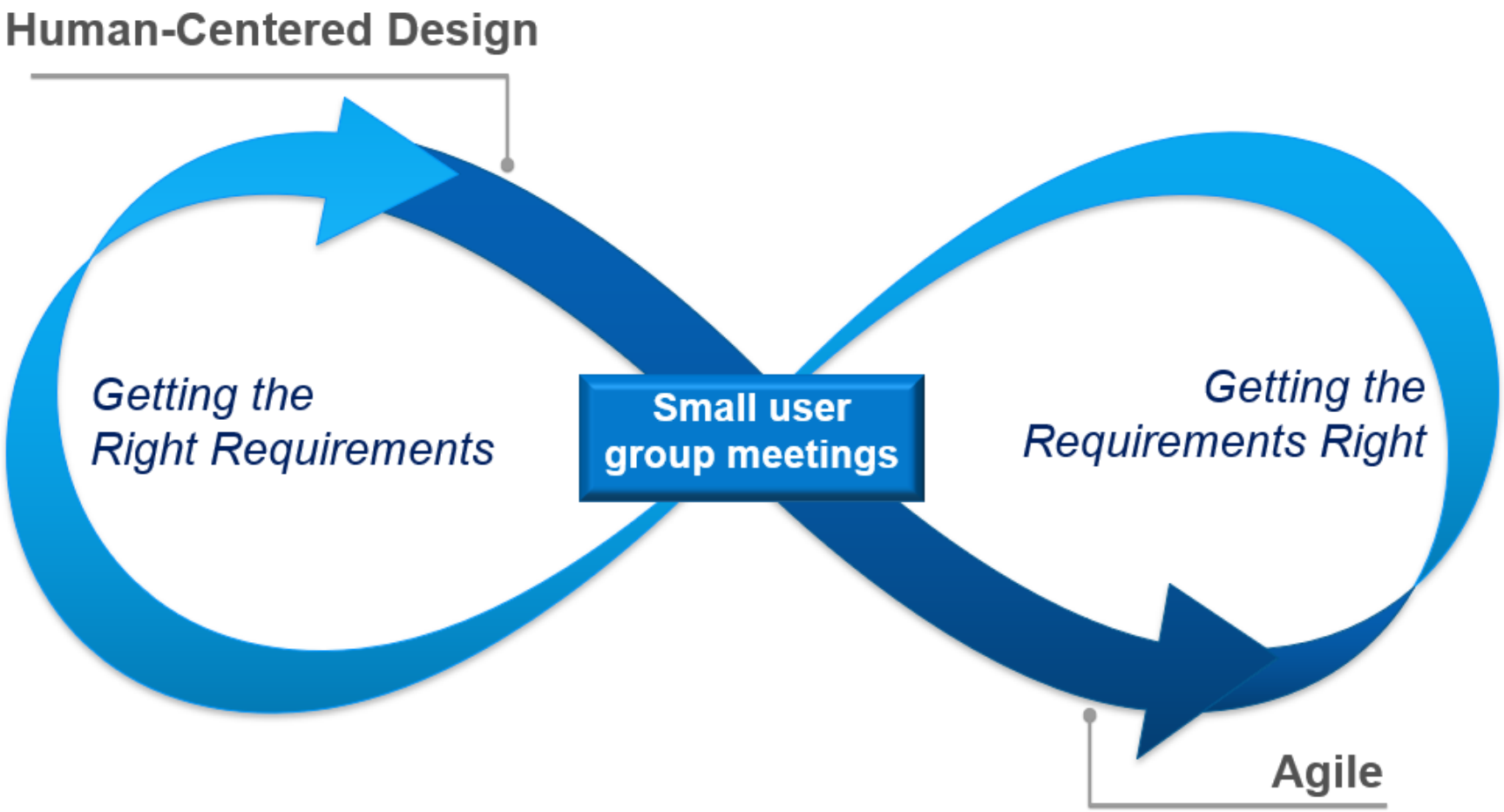


ParticipantID	ninds_guid	Phenotype	study_visit	ee_jobdesc__5	ee_jobdesc__6
map000001-01-01	NIHXP761EBLFC	MECFS	visit 1	0	0
map000002-01-01	NIHXR976BTHER	MECFS	visit 1	0	0
map000004-01-01	NIHKV461KTBC8	MECFS	visit 1	0	0

mapMECFS Cloud Infrastructure

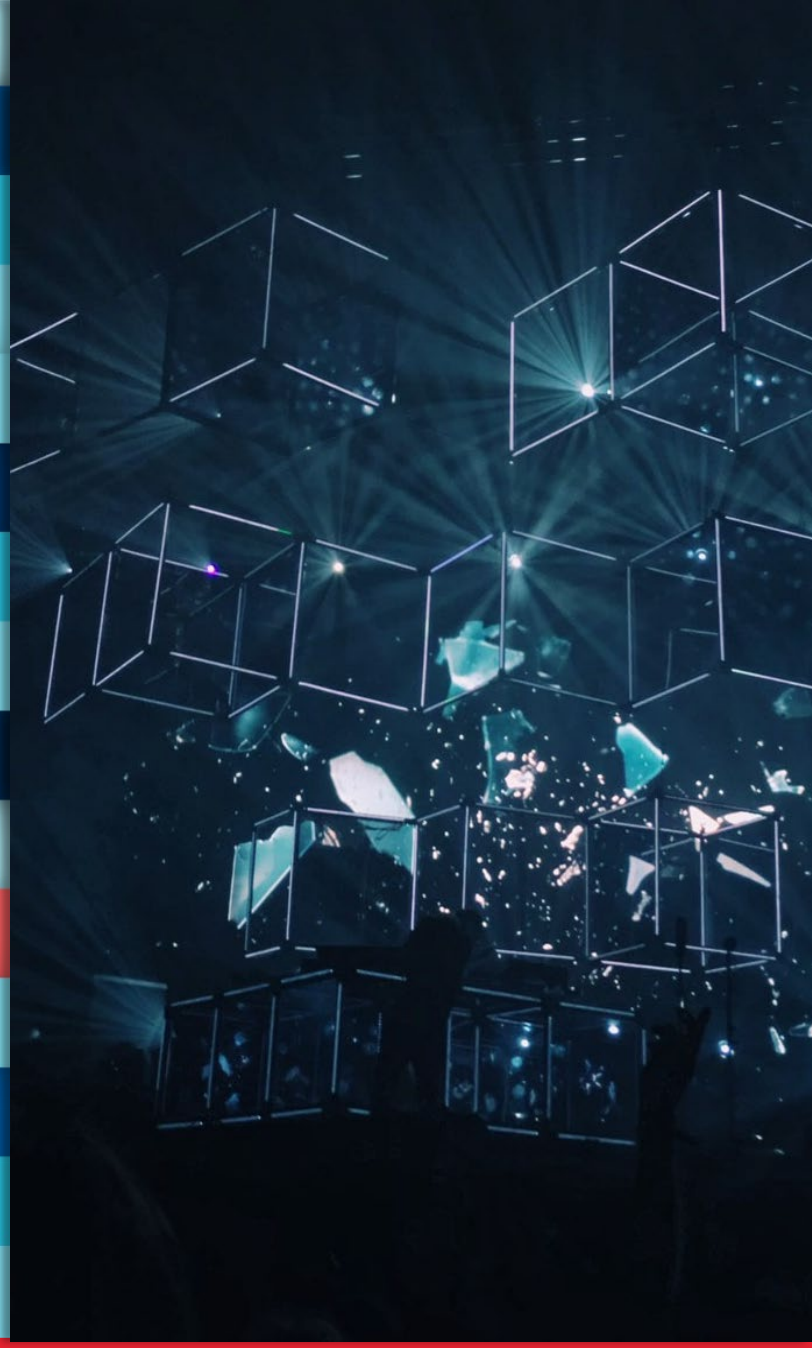


Going from CKAN to mapMECFS



The Map sCientific Universe (MCU)

With the extensions built and open-sourced and our infrastructure/ automation abstracted to be more reusable, we can now build **new data portals** in CKAN, inheriting all the great work done by the mapMECFS team with minimal effort.



Log in Register

IOmics mapOA

DATASETS ORGANIZATIONS GROUPS ABOUT Search

Welcome to **mapOA**, a data sharing portal created by the **Integrative Omics Center for Opioid Addiction Research (IOmics-OA)**. The overarching goal of IOmics-OA is to accelerate the neurobiological understanding of opioid addiction (OA) and to identify biologically actionable drivers. The mapOA portal is designed to share opioid addiction-related data and results from IOmics-OA and others.

The site was created by RTI International with funding form NIH's National Institute on Drug Abuse. Please visit the **IOmics-OA website** or email **iomics-oa@rti.org** for more information.

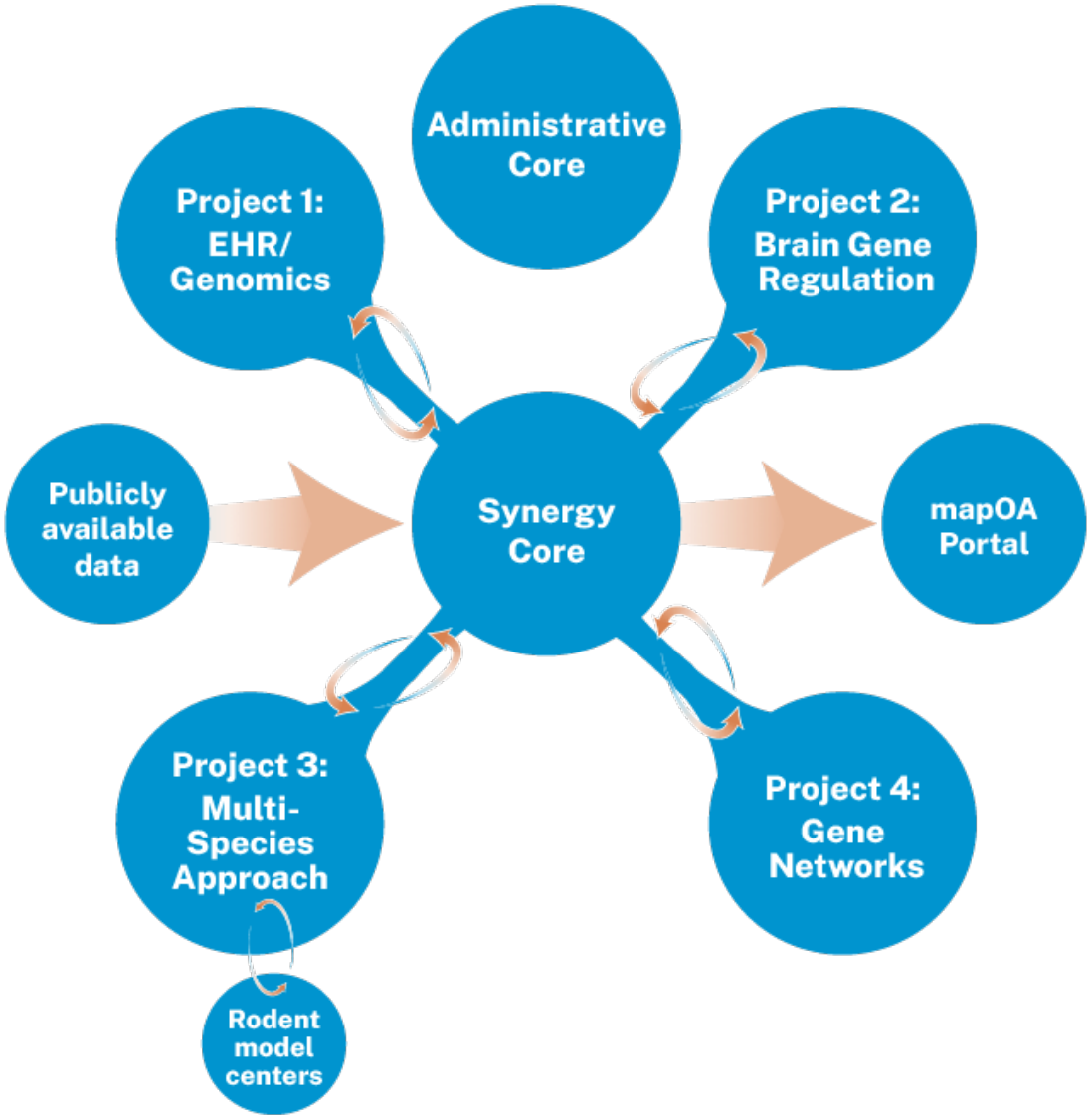
Explore Data

Upload Data

RTI INTERNATIONAL Geisinger Whitehead Institute JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

IOmics-OA: mapOA Portal

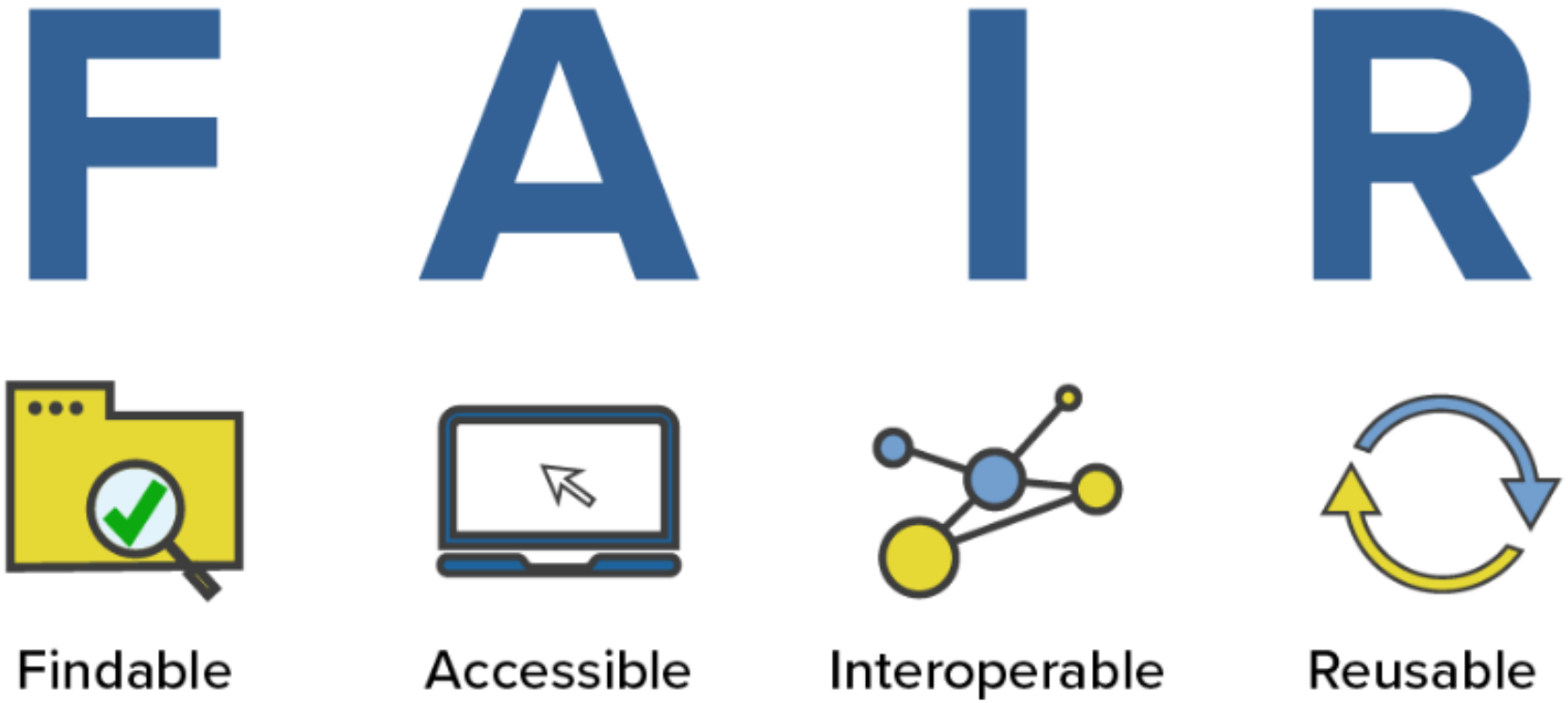
- Bringing together multi-omic data from a variety of sources
- With the variety of data, the map framework is being adapted to handle more data types (e.g., GWAS and integrated analyses), variety of analyses (e.g., meta-analysis and genomic SEM), and desired data security



Adapting map Framework

- Study design metadata is adapted to reflect Opioid Addiction, for example
  - Phenotype and Case definition
  - Comorbidities or OA characteristics
- Handling sample level and summary statistics data
- Analysis metadata is adapted to reflect new analyses including integrated analysis
  - Making relationships between datasets clear
- Data security is being adapted for eventual public consumption
- Development is coordinated with subject matter experts to design the desired system

FAIR Guiding Principles for scientific data management and stewardship



Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-200.html

Want to learn more about mapMECFS?

Email the mapMECFS support team
mapMECFS@rti.org

Visit our website
https://www.mapmecfs.org/

delivering **the promise of science**
for global good

RTI INTERNATIONAL