
The Dataset Catalog Beta
National Library of Medicine

Overview: Dataset Catalog

PubMed of Datasets!!

WHAT

- A catalog of biomedical datasets from selected publicly available repositories.

WHY

- Provide discovery systems for users, as stated in NIH Strategic Plan for Data Science and NLM 2017-2027 Strategic Plan.
- Allow researchers to discover biomedical datasets from many repositories.
- Drive adoption and acceptance of the NLM standard - DATaset Metadata Model (DATMM).

HOW

- Launch a beta version of Dataset Catalog using DATMM:
  - Assess Usage, Scalability and Sustainability of Dataset Catalog
  - Receiving user feedback to drive further development
  - Drive interest and adoption of DATMM

Dataset Catalog Team Members

NLM Team

- Peter Seibert
- Nancy Fallgren
- Alvin Stockdale

Presenting

- Jeff Beck
- David Hale
- Development and Programming Contractor Support (approximately 5 staff members)

DATaset Metadata Model (DATMM) Format

Resource Description Framework (RDF)

Language of the Semantic Web/Linked Data

Expressed in statements called *triples*

Subject <predicate> Object

Fritz <is> a cat

Anthony S. Fauci <has identifier> https://orcid.org/0000-0002-7865-7235

DATMM Overview

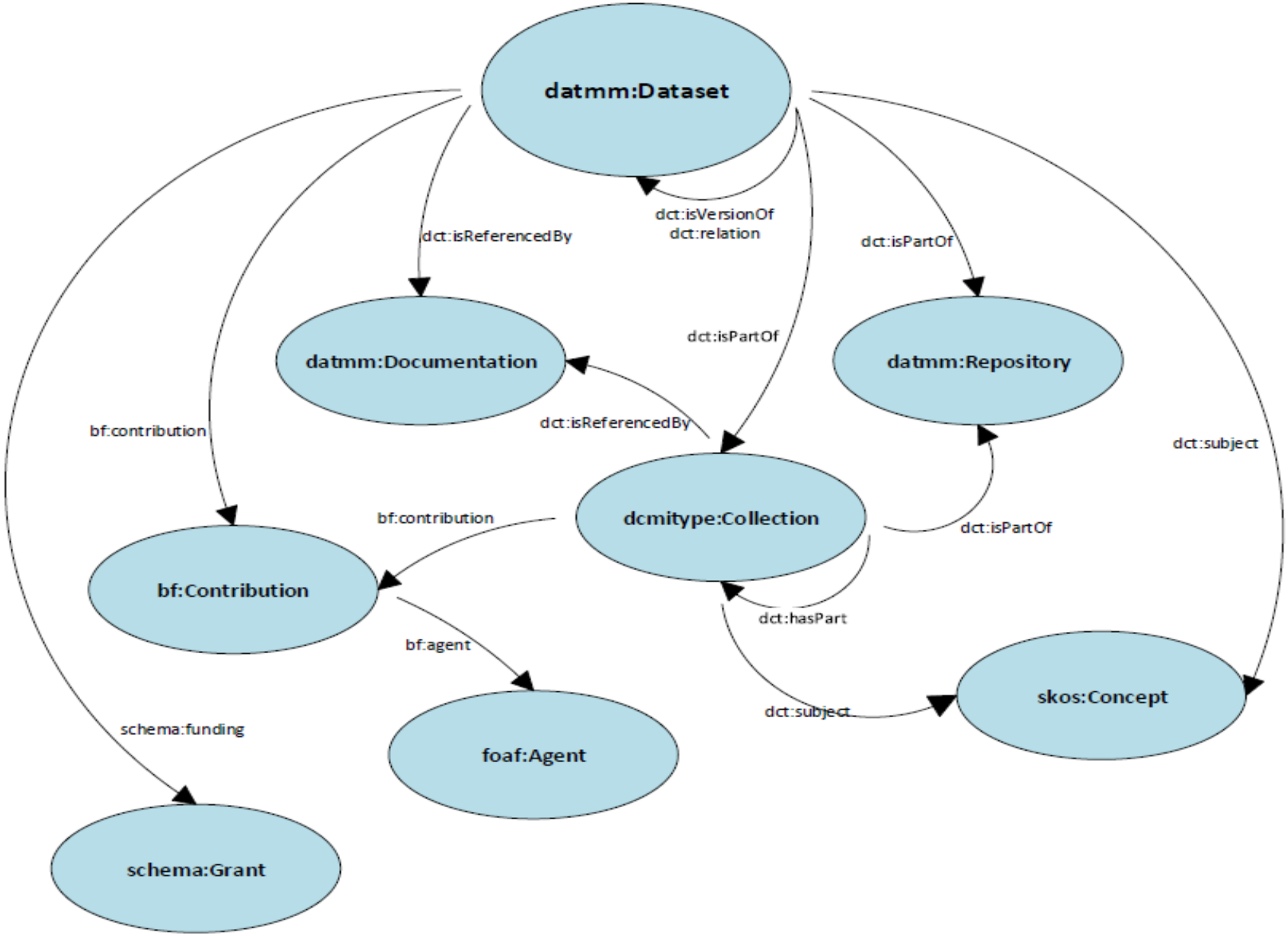


Figure 1. DATMM top-level Classes and their relationships to each other

DATMM Classes and Properties

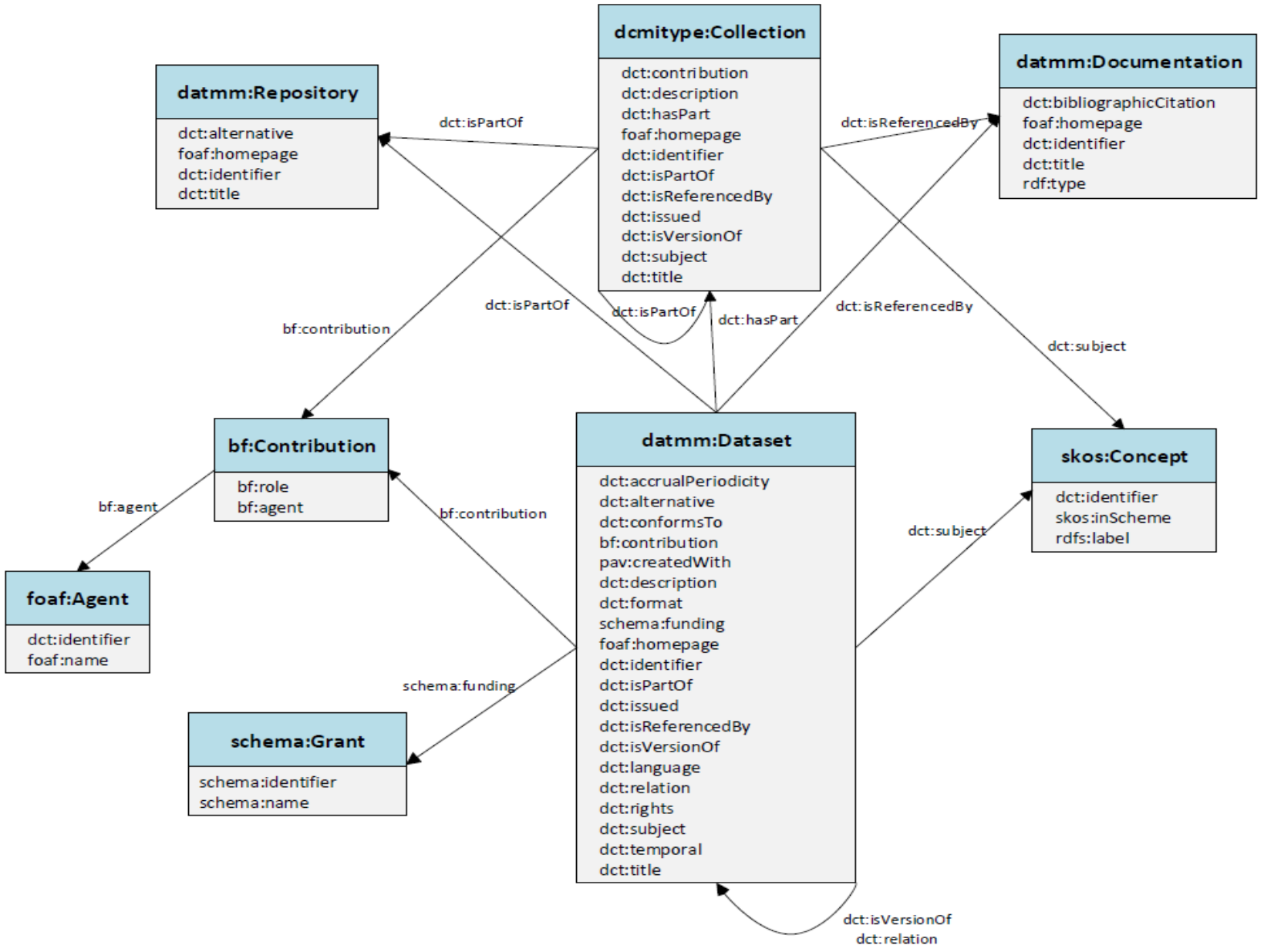


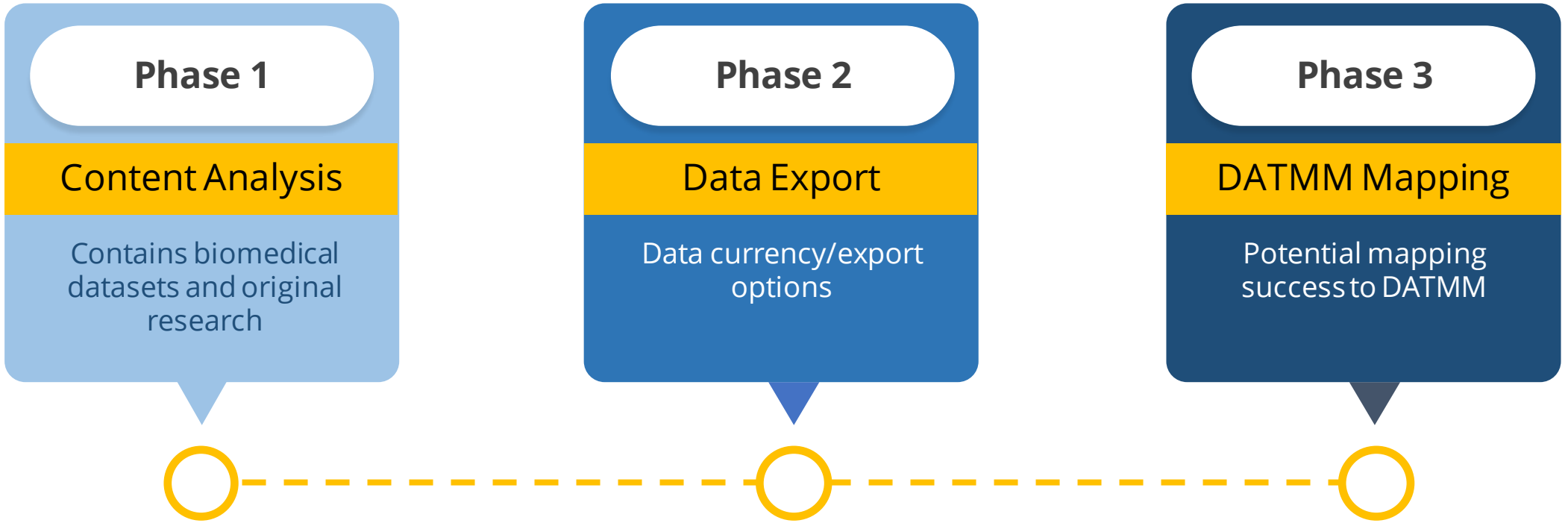
Figure 2. DATMM visualization showing Classes and properties

Finding Repositories

Resources:

- NIH BMIC repositories
- NIH GREI repositories
- PLOS repository list
- Re3data.org repository finder

Evaluating Repositories

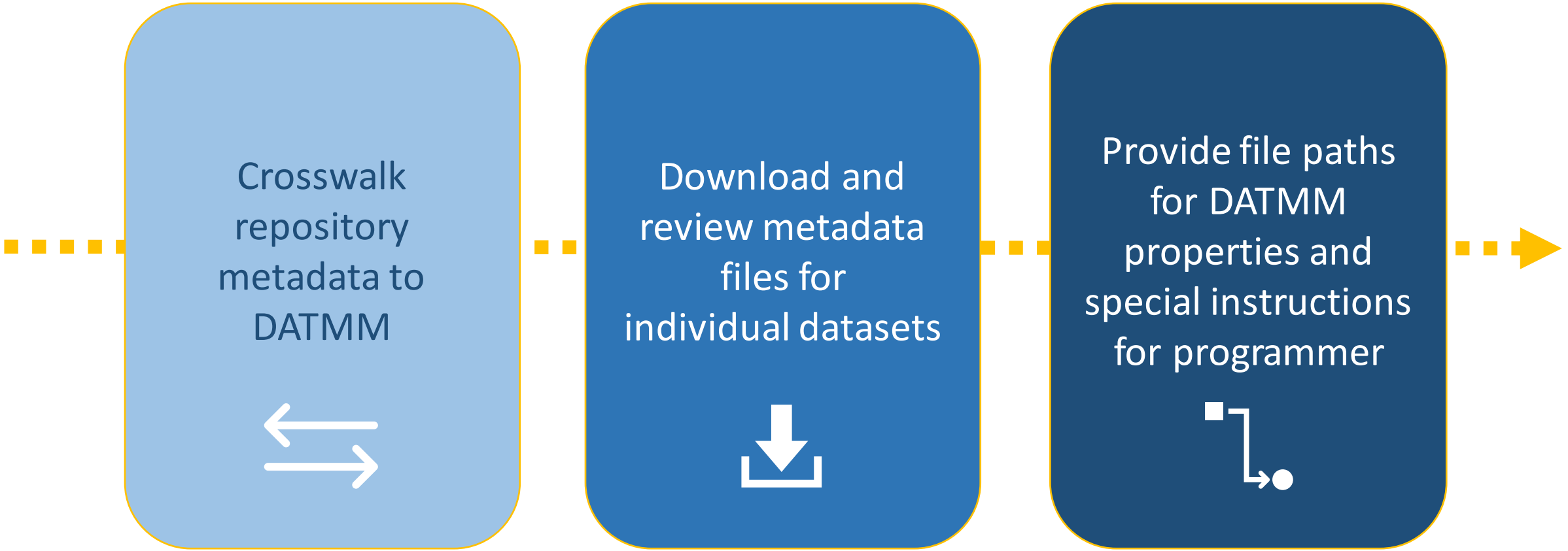


Ingested Repositories

4 repositories
80,000 datasets
and more coming!

dbGaP	*owned by NLM*
ImmPort Shared Data — Your site for searching and downloading shared data	*sponsored by NIH*
DRYAD	*NIH GREI*
HARVARD Dataverse	*NIH GREI, academic*

Mapping a Repository

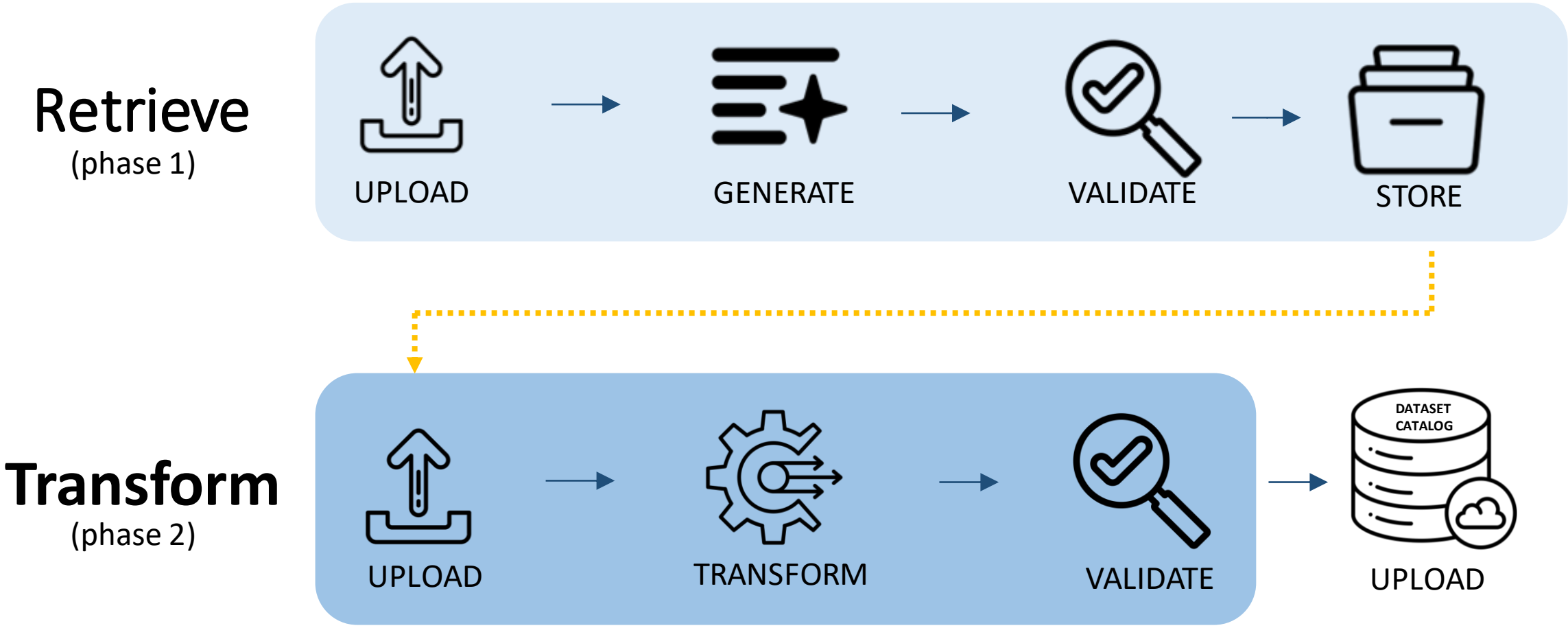


Implementing DATMM

datmm: Dataset – A discrete collection of data gathered for use in research

Property	Definition	Constraint	Cardinality	File Path	Metadata Value	Comments
bf:contribution	Agent and its role in relation to the resource.	Required	1:N		DATMM URI to Contribution	
pav:createdWith	The software/tool used by the creator when making the digital resource, for instance a word processor or an annotation tool.	Optional	1:N	datasetVersion/metadataBlocks/citation/fields/{loop}[typeName]=software/value/0/softwareName/value	MATLAB	
dct:description	An account of the resource.	Required	1:1	datasetVersion/metadataBlocks/citation/fields/{loop}[typeName]=dsDescription/value/0/dsDescriptionValue/value	<p>In previous work, we describe healthcare utilization as a time-series signal from 6 months ...	Clean up HTML coding

Generative AI and the Dataset Catalog



NLM Dataset Catalog - BETA

The Dataset Catalog will ***improve discoverability and reuse of research data***.

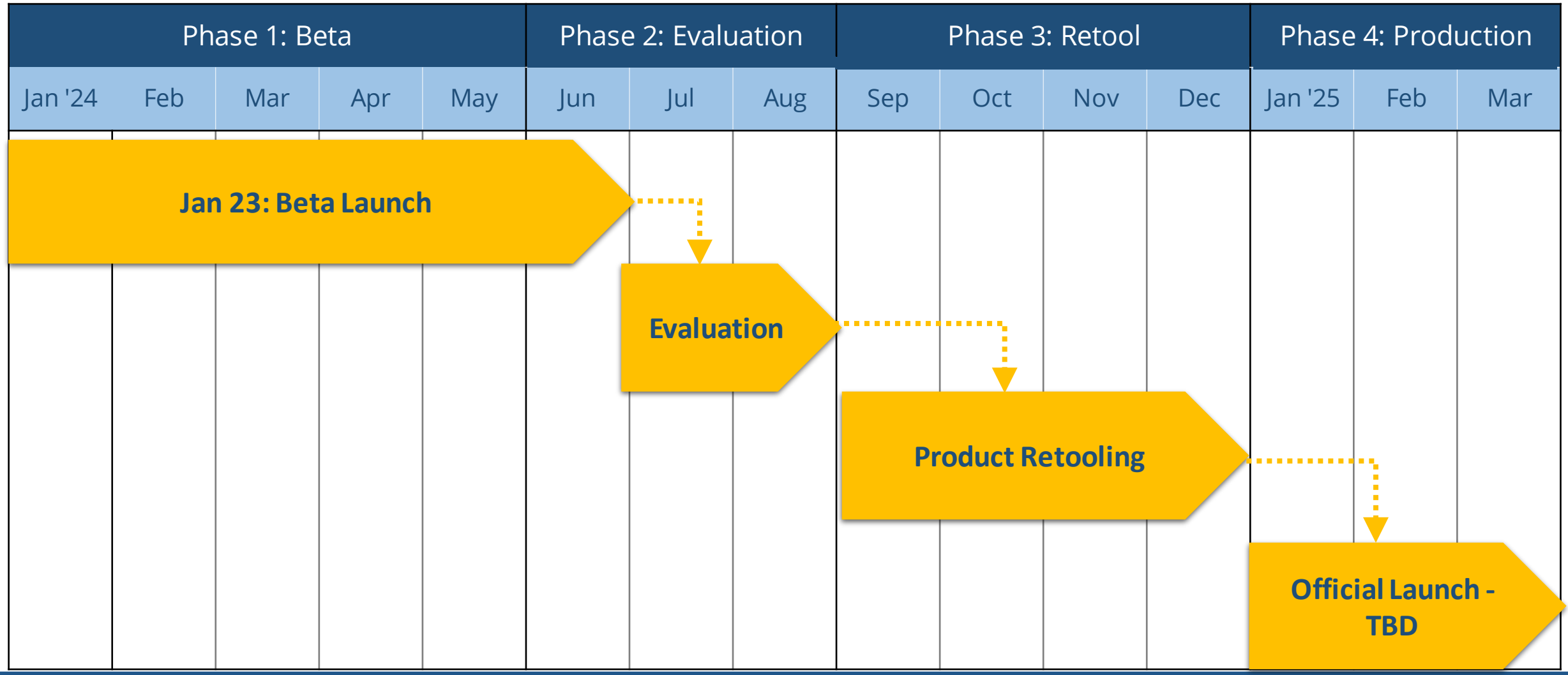
Referred to as the “PubMed of datasets”, the tool is a ***catalog of biomedical datasets*** from selected publicly available repositories.

Underlying the Dataset Catalog is a metadata ***schema developed by NLM called DATaset Metadata Model (DATMM)***.

We will continue to ***add repositories and datasets*** to the Dataset Catalog during the Beta launch.

User feedback is a key driver of the proposed phases of product development, especially during the beta phase planned January through June of 2024.

Dataset Catalog Timeline



Product Demonstration

NLM's Dataset Catalog: datasetcatalog.nlm.nih.gov