On behalf of the Monarch Initiative and the TransMed NCATS Data Translator projects, as well as the International Society for Biocuration, we provide the following response to the Request for Information: Metrics to Assess Value of Biomedical Digital Repositories.

We integrate data at scale from numerous public data repositories, develop algorithmic techniques for computing over that data, and we support websites and services that provide access to integrated data and computational results. We are thus familiar with the challenges of assessing the value of biomedical digital repositories and of demonstrating the value of our own. Collectively, our team members cover most of the roles mentioned in this RFI, including biomedical science researcher, bioinformatician, data scientist, standards developer, research repository manager, library/information scientist, and data curator.

Contributors	**Institution**	**Project**
Melissa Haendel	OHSU	Monarch Initiative, NCATS TransMed Data Translator, International Society for Biocuration
Andrew Su	Scripps	NCATS TransMed Data Translator, International Society for Biocuration
Julie McMurry	OHSU	Monarch Initiative, NCATS TransMed Data Translator
Christopher G Chute	JHU	NCATS TransMed Data Translator
Chris Mungall	LBNL	Monarch Initiative, NCATS TransMed Data Translator
Benjamin Good	Scripps	NCATS TransMed Data Translator
Chunlei Wu	Scripps	NCATS TransMed Data Translator
Shannon McWeeney	OHSU	NCATS TransMed Data Translator
Harry Hochheiser	U Pittsburgh	Monarch Initiative
Peter Robinson	JAX-GM	Monarch Initiative, NCATS TransMed Data Translator
Maureen Hoatlin	OHSU	NCATS TransMed Data Translator
Matthew Brush	OHSU	Monarch Initiative, NCATS TransMed Data Translator
Damian Smedley	QMCL, GEL	Monarch Initiative
Monica Munoz-Torres	LBNL	International Society for Biocuration
Guoqian Jiang	Mayo Clinic	NCATS TransMed Data Translator
Hongfang Liu	Mayo Clinic	NCATS TransMed Data Translator
Peter McQuilton	Oxford, UK	International Society for Biocuration
Tom Conlin	OHSU	Monarch Initiative, NCATS TransMed Data Translator

Preface

“Knowledge” is the collection of insights captured by experts, providing an explanatory framework for evaluating new observations. A “knowledge base” makes it possible for that knowledge to be maximally impactful by rendering it findable and computable. Maintaining databases that house scientific knowledge is far more cost-effective than rederiving that knowledge experimentally. Moreover, knowledge bases provide efficiencies beyond those of basic science, they are essential to pharmaceutical R&D and open drug discovery advances as well. Nevertheless, not all databases can be maintained at the same level of support, or for the same duration of time. Therefore, the evaluation and review of biomedical data repositories should be mindful of quality, accessibility, and value of the database resources over time and across the translational divide.

Why traditional metrics fall short. Traditional citation count and publication impact factors as a measure of success or value are known to be inadequate to assess the usefulness of a resource. This is especially true for integrative resources. For example, almost everyone in biomedicine relies on PubMed, but almost no one ever cites or mentions it in their publications. While the Nucleic Acids Research Database issues have increased citation of some databases, many still go unpublished or uncited; even novel derivations of methodology, applications, and workflows from biomedical knowledge bases are often “adapted” but never cited. There is a lack of citation best practices for widely used biomedical database resources (e.g. should a paper be cited? A URL? Is mention of the name and access date sufficient?). Even secondary tracking of the resource’s identifiers is difficult as most researchers do not use such identifiers in their manuscripts. Efforts such as those by Identifiers.org, and N2T.net are working together to help improve consistency and citability of records within such data resources that lack DOIs.

Other measures of impact (e.g. letters of support, patents, etc.), have also been insufficient to rigorously assess impact or value as they essentially only relate to impact. It is clear that evaluating a data or knowledge resource is non-trivial, as evidenced by the large number of evaluation and impact working groups and rubrics. We acknowledge that a one-size-fits-all solution is unrealistic. This RFI is an opportunity to stop “looking for our keys under streetlight because that is where the light happens to be.” Here, we focus exclusively on data access and reuse issues as we feel that these are most important to us as data integrators; moreover these factors may be least likely to be covered extensively in many other responses to this RFI.

We arrange our response according to the commonly cited **FAIR principles -- Findable, Accessible, Interoperable, and Reusable** [PMID:26978244], and have added three additional principles: **Traceable, Licensed, and Connected**. These three additions are of course very closely related to the original principles; however, we call them out as they are still largely overlooked and underappreciated, even within FAIR. It is worth noting that FAIR principles apply not only to the resource as a whole, but also to their key components; this “**fractal FAIRness**” means that even the license, identifiers, vocabularies, APIs themselves must be Findable, Accessible, Interoperable, Reusable, etc.

FINDABILITY

Ensuring that a biomedical data resource is actually discoverable is often overlooked. The resource and its data/components are only useful if they can be found.

F1: Discoverable through various external mechanisms

Concrete metrics:

- **Registered:** The resource should be registered, for example, using a BioDBcore ID via BioSharing, PathGuide for pathway data, Re3Data, etc.
- **Discoverable via search engines/applications:**
  - **By name:** Resource names should easily discovered by web search engines (e.g. easy to spell, highly ranked, and branded/named in a way that is distinct from similar resources. Resource names (and acronyms) should not conflict with other existing resources likely to be used in the same context.
  - **By features:** People that know about a resource search for it using its name, but those who don't know about a resource should be able to find it by searching for its features (e.g. "variant databases"). Provision of more precise metadata (see below) can help ensure that a resource has *accurate* ranking in search engines such as Google as well as more targeted applications such as BioCaddie's DataMed or Re3Data.
- **Linked from external resources:** The resource and its underlying components should be linked from/to other relevant online resources. This could be via a variety of mechanisms, such as NCBI's LinkOut feature.

F2: Contents/components are well documented and searchable

Concrete metrics:

- **Metadata documented:** can you learn enough from the metadata to know if further effort with the dataset is warranted i.e timestamps, versions, counts etc.
- **Indexed**: Contents are indexed and ideally optimized to support the most common types of queries
- **Searchable using various mechanisms**: Search boxes and APIs are not only present, they are also documented, and populated with examples
- **Contactable:** Contact information is readily discoverable; responses are prompt

ACCESSIBILITY

Access to the underlying data in a biomedical data resource is not always available.

A1: Diverse data access mechanisms

Where practical, the resource should provide the option to download all data via one or more well documented mechanisms. Evidence of dissemination mechanisms that enable the community to use knowledge and data in innovative and reproducible ways should be evident.

Concrete metrics:

- **Dumps:** Whole database dumps are available (where appropriate)
- **Query:** Query interfaces or Mart-style exports, where possible
- **Downloads:** Slices of the database and individual records can be downloaded (e.g. as JSON/XML/tab delimited, etc.)
- **API:** Application Programming Interface (API) for the data exists

A2: Well structured and provisioned APIs

If the resource provides an API, the following highly-recommended implementation guidelines are recommended. Direct database endpoints (e.g. MySQL, SPARQL etc) can be valuable; however, expertise in using these varies. Therefore, it is important to also wrap these with an API wherever possible. A summary of important REST principles is below; see also SSI REST best practices here.

Concrete metrics:

- **RESTful**: Follow RESTful API pattern
- **JSON:** Return JSON if possible, TSV if not
- **Retrieval:**
  - Allow retrieval of a single record by using its identifier
  - Allow batch retrieval of a list of data entities using a list of identifiers
- **Paging:** Provide a query interface to return matching data entities with paging support
- **Versioned:**
  - Provide versioned URL pattern for future API changes
  - Document policies for change management
- **Uptime:**
  - Provide an API uptime report (third-party services are available to reduce the implementation burden)
- **Access:**
  - Grant access requests (e.g. new accounts or API keys) promptly and efficiently
  - Grant write access to trusted partners to make contributions, corrections, suggestions to records

A3: Understandable data and scope

A clear description of what the repository covers in terms of content and target audience.

Concrete metrics:

- **Audience:** Target audience and use cases are well defined and obvious from the homepage
- **Content:** The content types included (e.g. genes, variants, species, protein structures, RNAseq data, etc.) are obvious from the homepage
- **Browsable:** Data is browsable through high-level categories / visualizations
- **Documented:** The data model, schema, data dictionaries, etc. are well documented
- **Tutorials:** Tutorial available for novice users, and literature cited that previously used the repository

INTEROPERABILITY

The impact of the data repository only increases with data reuse, but this is hampered when not planned for in terms of interoperability. This is related to identifier provisioning, but extends beyond.

I1: Identifiers

How the data is referenced and stewarded is crucially important for a biomedical data repository (see this community declaration, excerpted below regarding how to evaluate identifiers).

Concrete metrics:

- **Credit** any derived content using its original identifier
- Help local identifiers travel well: **document prefix and patterns**
- **Design** new identifiers for diverse uses by others
- **Avoid embedding meaning**, or relying on it for uniqueness
- Opt for **simple, durable web resolution**
- Implement an identifier **version-management policy**
- Make **URIs clear and findable**
- **Do not reassign or delete** identifiers
- **Document** the identifiers you issue and use
- **Reference and display** responsibly

I2: Vocabularies, Ontologies, and exchange standards

Resources should document which ontologies, terminologies, and value sets are used to type information and how the information is structured and exchanged.

Concrete metrics:

- **Semantics/data structure:**
  - Data dictionary is provided
  - Defined schema or data model is provided
  - Services are well aligned to the model and consistent across various access mechanisms
  - Structure, format, architecture, and metadata for the repository is consistent with community norms or shared specifications (for example, use of the W3C Dataset Description)
- **Exchange standards:**
  - Data are made accessible using common exchange formats, if applicable (for example, use of the HL7 FHIR standard for exchanging healthcare information electronically)
  - Data elements are well-defined using metadata standards (e.g., ISO/IEC 11179, DDI and SDMX/ISO17369)
  - Value set services and value set definition services using the Common Terminology Services 2 (CTS2) standard
- **Ontologies:**
  - All ontologies in use are documented in one place and are consistently applied to the data
    - Novel ontologies, if any, are registered in public standards repositories (such as the OBO Foundry Library) and released via standard well documented mechanisms (for examples ROBOT or the OBO Starter Kit)
    - Appropriate community standards/vocabularies are used to record metadata; preferably standards that are: a) designated or *de facto* standards within the relevant domain, and b) free to use, see also Licensure section
    - Version of the ontologies used is indicated
    - Ontologies are attributed according to community best practice

I3: Versioning

Concrete metrics:

- Data versioning and/or change history is **well documented**
- **Prior versions** of each database release (or each record, if appropriate) are accessible

REUSE

Evidence of community reuse of the data means the data has value to the community. **Reusability** is necessary for **reuse** and, reuse in turn, is necessary for **impact**; however these are measured in different ways. Knowing what data is reused, where it is reused, and how extensively it is relied upon helps establish a measure of community trust.

R1: Use

There should be evidence that the resource is in demand.

Concrete metrics:

- The resource should have a large and engaged user base (those referencing/using the resource). This is not just the **size of the user community**, but the **value** this community gets from the resource.
- Metrics in terms of pageviews, time on site, and new and return visitors can be self-reported by repositories using third-party tools (e.g., Google Analytics) or server logs. However, reporting of such statistics in proposals and progress reports should be accompanied by a precise description of how these statistics are tabulated. Time course data, rather than single snapshots, are most representative.
- Whenever possible, access metrics should be **stratified by key user groups** (e.g., human users through the web, programmatic access by analysts, programmatic access by third-party applications).
- Evidence of community contribution and/or interaction (for instance, user-submitted tickets)

R2: Impact

This is a measure of the “reach” of the resource in terms of scientific contribution. Degree to which data is *actually* used, reused, or derived and redistributed. Lists of all downstream consumers of data and resources should be compiled.

Concrete metrics:

- Degree of **uniqueness**, e.g. an assessment of scope overlap with other existing resources
- **Embeddedness:**
  - **Links:** Number of *other* biomedical community tools and contexts that link out to the resource (most importantly context-appropriate links to specific records, rather than links to the whole resource)
  - **Algorithmic uses:** External algorithms, tools, and knowledge bases that are using the data
  - **Integrative queries:** Inasmuch as they can be measured (for example, in open platforms like Wikidata), frequency of integrative queries by the community that utilize particular data elements or resources
  - **Interdisciplinary uses:** Evidence of adoption/use by disciplines outside the originally intended one
- **Publications, case studies** by others
- **Testimonials** detailing unique contribution of resource

R3: Awareness and responsiveness to key user needs

Concrete metrics:

- **Trackers:** Existence of public issue trackers where members of the community can see all outstanding bugs/feature requests and comment or log a new one of their own.

- **Advisers:** Existence of an advisory board
- **Mailing lists:** Existence of a mailing list and number of subscribers
- **Responsiveness:** Turnaround time (eg. for feature requests, bug fixes, or, in the case of repositories, turnaround time to data publication)
- **Release notes:** Practice of regular releases with release notes summarizing major improvements

R4: Quality of data content and service

Usability testing to quantify more nuanced determinations such as “easy to find” or “well documented” should be encouraged.

Concrete metrics:

- **Currency:** Data is as up-to-date as it needs to be to be useful
- **Uptime:** Resource is as reliable as it needs to be in order to be useful
- **Comprehensiveness:** Resource is as comprehensive as it needs to be in order to be useful
- **Data Quality**
    - Mechanism for data quality assessment is declared and clear
    - Transparency around outcomes of that assessment (e.g. data that may be present but not yet QC’d is presented with a flag)
- **Probability**: Transparent with regard to probabilities, where relevant (for example, text mined associations)

TRACEABILITY

T1: Provenance

The data’s provenance is well documented and attributed (the data within the resource)

Concrete metrics:

- Derived content is credited using its original identifier and linked using some persistent mechanism (eg. PURL, identifiers.org etc)
- Data processing/transportation provenance is tracked using systems such as PROV or the W3C Dataset description, where relevant.
- For complex integrated data, provenance information should be available via APIs, as graphs, or other mechanisms

T2: Attribution

The contributions to the content (data, tools, algorithms, sources, etc.) are clearly declared.

Concrete metrics:

- The contributor, author or data source’s organizations are attributed using identifiers, logos, and other references to source content
- Individual people / institutions / grants etc. are referenced with identifiers where relevant, such as from Wikidata; some examples are also: ORCID or ResearcherID for people; Digital Science GRID or OCLC for organizations
- Community standards are followed for attribution where declared, see above example for ontologies.
- Documentation exists for how to cite a record from the resource or how to cite the whole resource

LICENSURE

Not all data resources are free to use, derive, and redistribute, even if they are publicly funded and seemingly publicly available. This is true for almost all existing NIH-funded resources. Some widely-used examples of resources that are commonly thought of as “open” but in practice cannot easily be derived and redistributed are:

> *ClinVar: “This site contains resources which incorporate material contributed or licensed by individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws…..All persons reproducing, redistributing, or making commercial use of this information are expected to adhere to the terms and conditions asserted by the copyright holder.”*
>
> *OMIM: “This site contains resources which incorporate material contributed or licensed by individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws…..All persons reproducing, redistributing, or making commercial use of this information are expected to adhere to the terms and conditions asserted by the copyright holder.”*
>
> *PharmGKB: “PharmGKB grants use of its contents for research purposes. The use of this data and knowledge is NOT available for redistribution... This content is freely available to researchers in academia and industry for RESEARCH PURPOSES. Absent the issuance of a license by Stanford, the content shall not be used in any non-research commercial application in any form."*

We believe that there needs to be better awareness of the impacts of data license choices among both resource providers and NIH program staff. Moreover, few databases produce just data; most also produce software source code, algorithms, and applications. There should be licenses explicitly covering each of these products.

L1: Documented, clear, standard, minimally restrictive, contactable

We propose a license rating of 1 to 5 stars based on the following issues:
Concrete metrics:

1. **Documented:** Explicit data use terms (ideally formal licenses) should be defined by the resource providers and easy to find
2. **Clear:**
   a. At a minimum, licenses/data use agreements must be clear and easy to understand. A variety of specific examples of data use/reuse conditions should be included.
   b. Licenses should not require negotiation and licenses themselves should be legally re-distributable without engaging legal counsel
3. **Minimally restrictive:** The licenses and/or data use agreements should explicitly permit downstream data reuse, derivation, and re-dissemination
4. **Standard licenses.** We note that considerations for data are significantly different than those for software and they must be considered separately (see this blog for example).
   a. **Standard data license:** For data, ideally CC0.
   b. **Standard software license:** For software, ideally Apache version 2. Note that software license choices are the subject of much community discussion especially regarding “copy-left” approaches and there are other valid standard options available (such as GPLv2, GPLv3, AGPLv3, etc.)
5. **Contactable:** There should be an appropriate person available for contact with questions about licensure; this person’s contact information should be easy to find

L2: Transparent about flowthrough implications

If others' data is redistributed, clarity about the licensing implications of the redistribution is critically important.

Concrete metrics:

- Documentation about which source resources/data, if any, come with flowthrough implications
- Links to the original licenses/data use terms of all redistributed content. It is currently commonplace that such terms do not exist; in such cases, it should be clearly stated that license/terms could not be found.
- If specific authorization has been obtained for redistribution

CONNECTEDNESS

Having diverse data in the same warehouse can be a good starting point, but it does not make the data inherently more usable or integrated. Data connectedness can be a measure of computational power across diverse data.

C1: Connectedness

- For repositories that have many content types, a measure of the degree of connectivity between the types. For instance,using graph or link association measures to evaluate degree of data integration and complexity. If you view a given resource as a sub-graph, you could measure it in terms of nodes, edges, and quantify how well-connected it is to other sub-graphs. Less connectivity is only an indicator of quality in cases where the information is highly novel.
- The resource provides qualified links between related entities in other systems. For instance, unqualified database cross references ("DB xrefs") could mean that an entry is related to another, derived from another, more general or more specific than another, etc. Lack of description of *why/how* records are related has led to others integrating based on false assumptions.
- The data model complexity is appropriate to the described use cases / target audiences; the API allows the data complexity be put to its full use
- Clinical and basic research data is related across the translational divide in some manner