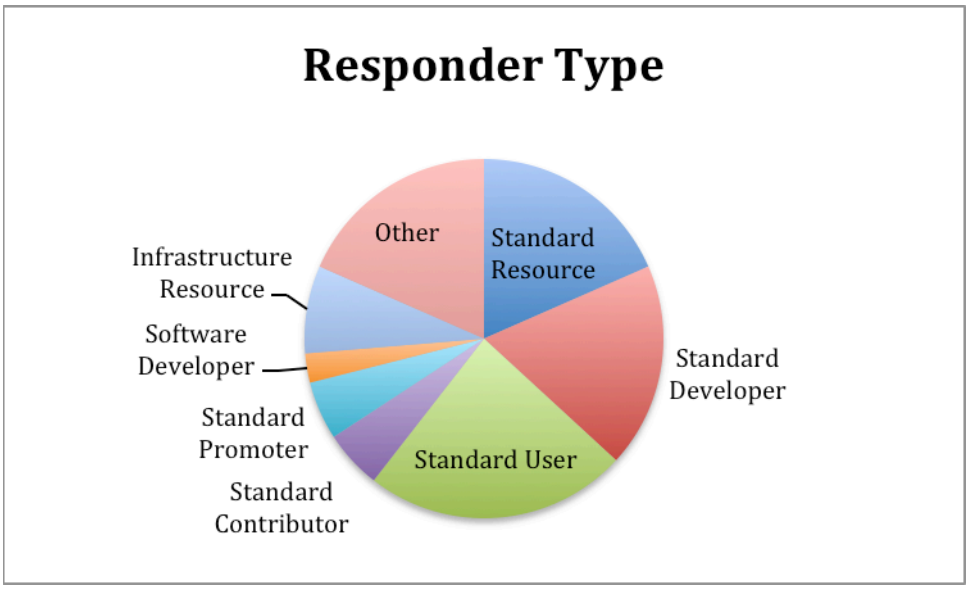
Updated 16 Jan 2015

Overview of Responses to RFI NOT-CA-14-054
Input on Information Resources for Data-Related Standards Widely Used in Biomedical Science

In this Request for Information (RFI), the NIH invited comments and ideas from interested persons to inform the consideration of an NIH Standards Information Resource (NSIR) that would collect, organize, and make available to the public trusted, systematically organized, and curated *information about* data-related standards. The two major questions in the published RFI were related to the Content of an NSIR, i.e. types of information and metadata that are appropriate; and then knowledge about relevant current existing efforts, resources, and lessons learned. The NSIR is a potential initiative of the NIH Big Data to Knowledge (BD2K) program as part of efforts to facilitate the broad use of biomedical research data. It is envisioned that such data standards themselves may become specific, citable digital objects within the digital research ecosystem, or Commons, as envisioned by the BD2K.

Respondent Analysis

The RFI produced 30 responses from a diverse group representing basic and clinical research, health care, federal agencies and others with interest in data related standards. Responses were characterized into eight different "types" of respondent: Standard Resource creator/ maintainer, Standard Developer, Standard User, Standard Contributor, Standard Promoter, Software Developer, Infrastructure Resource, and Other. The "Other" category consisted of a mélange of biomedical researchers, bioinformaticians, curators, database managers, engineering firms, and public advisory groups. Several responses were categorized under two different types, such as both a Standard Resource and a Standard Developer. The distribution is summarized in the pie chart below. (List of respondents included on "Response Summary" table, included as supporting information.)



Updated 16 Jan 2015

Response Highlights

A highlighted selection of responses is included here to illustrate some of the themes and major points. For further detail, please see Response Summary Table and/or original responses.

COMBINE and BioPAX are a Standard Resource and Standard Developer respectively. Both submitted responses. COMBINE is the Standards "home" for Computational Biology and includes the standards: BioPAX, CellML, SBOL, SBGN, etc. (See: http://co.mbine.org/) COMBINE is an initiative to coordinate the development of standards and formats for computational models, which are interoperable and non-overlapping. The COMBINE response described a resource exhibiting a number of similar features as a potential NSIR, and they included a useful knowledge management section on "lessons learned", for instance a recommendation on dealing with identifiers. It also illustrates an important consideration in thinking about creating the NSIR. Namely, that there are existing Ecosystems of standards already; the NSIR will need to be able to inform potential users about these ecosystems in addition to the individual standards.

BioPAX is a community driven process that makes pathway data substantially easier to collect, index, interpret and share. BioPAX's response mentioned that these data are hampered by current fragmentation of pathway information across many databases with incompatible formats. Using BioPAX, millions of interactions, organized into thousands of pathways, from many organisms, are available from a growing number of databases. This response suggested including BioPAX in the NSIR.

Dr. Michael Hucka responded on behalf of the Systems Biology Markup Language (SBML) editors, as a Standard Developer and Standard User. SBML, also part of COMBINE, is the de facto standard format used to store and exchange computational models in systems biology. Their response illustrates the need to include different types of standards, such as markup languages and models, and to not lose the information about standards ecosystems. The SBML response recommends to not duplicate COMBINE, a common theme among multiple respondents.

Reflecting a different set of interests, the Health Sciences Library at NYU School of Medicine is a Standard User and Standard Promoter. They proposed a metadata schema for data standards. This schema draws upon one recently created by this Library for a catalog of datasets which in turn, was based on schemas from Dryad [1], DataCite [2], W3C Data Catalog Vocabulary [3], and the minimal metadata elements that were created to inform the NIH Data Discovery Index at the Big Data to Knowledge (BD2K) workshop held in August 2013. This might be used for input in creating a testable set of metadata for the NSIR resource after the project has begun.

Metadata Element	Description of metadata
unique ID	A unique identifier assigned to the standard.
title	The title of the standard.
alternate title	Any alternate titles used to describe the standard including acronyms.
description	A detailed description of the standard which should include: • What area of research is covered in the standard • What the standard is used for
date of creation	The date when the standard was created. Should follow standard guidelines (e.g. ANSI/NISO Z39.85-2012: YYYY-MM-DD).
date last updated	The date when the standard was last updated. Should follow standard guidelines (e.g. ANSI/NISO Z39.85-2012: YYYY-MM-DD).
versioning (if applicable)	The specific version of the standard. This version should be connected to all other versions of the standard -- especially the original to ensure that the provenance of the standard is maintained. (e.g. Version 1, 1.1, 1.2, etc.).
publisher/creator name	The person(s), institute(s), organization(s), or entity(ies) responsible for creating the standard.
publisher/creator type	A categorically assigned type to the publisher/creator responsible for creating the standard. This will assist in filtering information when searching/browsing through standards. (e.g. consortium, individual researcher, government agency, academic institution). This will also help distinguish between standards organizations and grassroots standards.
publisher/creator URL	A direct link in the form of a URL that points to the web page or profile of the person(s), institute(s), organization(s), or entity(ies) responsible for creating the standard.
funder	The person(s), institute(s), organization(s), or entity(ies) responsible for funding the standard.
access path URL	A direct link (via URL) to the standard and all of its associated components.
access instructions	Detailed instructions on how to implement the standard, including the software/programming/tools required. This section should also

	provide links to official readme documentation, instruction manuals, FAQs, etc.
format of standard	The format(s) of the standard in terms of how it can be used and/or downloaded. (e.g. xml, csv, etc)
associated publication(s)	Links to associated publications that describe or discuss the standard. (e.g. connect standard with related PMIDs)
related publication(s)	Links to publications that use and/or cite the standard. (e.g. connect standard with related PMIDs)
related standard(s)	UIDs of other standards that are related in some way, such as covering overlapping types of research or being interoperable.
related standard(s) note	A note providing information about how the standard listed in the "related standards" field are related.
related dataset(s)	Links to datasets that use and/or cite the standard. (e.g. link to UID in the NIH Data Discovery Index)
domain	A controlled vocabulary term that represents that category of research where the standard can best be used. This will also serve as a filter when searching/browsing. (e.g. Alzheimer's, Traumatic Brain Injury)
subject	A controlled vocabulary of more specific terms that can be used for searching, as well as for establishing linkages with PubMed and the NIH Data Discovery Index. Medical Subject Headings (MeSH) should be used if possible to maintain interoperability with other NIH discovery systems.
field categories	Specific categories that will help to break down the fields within a standard into specific parts for improved searchability. (e.g. Demographics, Family History, Vital Signs)
standard open or proprietary	A simple indication of whether the standard is open and available or proprietary.
tools for standard	A listing of all of the tools required to implement the standard, read the data in a standard format, or facilitate the use of the standard.

The NYU Health Science Library response also mentioned a framework describing a number of groups working to harmonize conflicting standards (e.g. the Joint Initiative on SDO Global Health Informatics Standardization); a similar effort coming from NIH, with the promise of ongoing support, could coordinate and greatly aid these efforts, they said.

Along with COMBINE, described above, several other respondents were classified as Infrastructure Resources. These responses offer a broad view that aids in planning for the potential NSIR. Among them were the International Neuroinformatics Coordinating Facility (INCF) and Bioconductor. INCF has implemented the globally coordinated INCF Standards for Data Sharing in Electrophysiology Task Force. This Task Force is working out the requirements for a standard for storing electrophysiological data and related metadata, with the aim to enable the efficient sharing of these types of data. The INCF task forces also work on data and metadata re-use, sharing, provenance, and harmonization. Their standard will be based on HDF5 (http://www.hdfgroup.org/HDF5/) and specifies what needs to be stored for commonly used electrophysiology data types. A response was also submitted representing Bioconductor, the 'largest open-source software project dedicated to biological data analysis with an active user base numbering several thousand. It is a simple, extensible system for describing key features of biological software and is necessary to make software discoverable, maintainable, and useable.' This response is probably more relevant to the Software Index effort, however.

Three other "resources" submitted coordinated responses dealing with various aspects relevant to an NSIR: (1) NIF and the Monarch Initiative; (2) Biosharing.org, Research Data Alliance and collaborators; and (3) CEDAR, the new BD2K Center, which includes Mark Musen as PI (Bioportal) with Biosharing.org and IMMPort. The three respondents represent a number of organizations that collaborate and are all also very active in developing standards, integrating data and providing information about standards and data; they submitted detailed and thoughtful responses that should be reviewed carefully in planning the NSIR. NIF, for instance, "advances neuroscience research by enabling discovery and access to public research data and tools worldwide through an open source, networked environment. NIF is based on three main indexes: (1) the Registry, a PubMed-like listing of basic attributes and descriptions of databases, software tools, core facilities and biobanks; (2) the NIF Data Federation, a PubMed Central-like meta-index that searches continuously updated content of over 200 databases all integrated via the NIF DISCO framework; and (3) the NIF literature, which contains PubMed and the open access portion of PubMed Central." Similarly, Biosharing.org is a well-known standards resource. "BioSharing works to map the landscape of community developed standards in the life sciences - See more at: http://www.biosharing.org/#sthash.9IxITy7X.dpuf". These groups also provided information about relevant metadata, a list of relevant activities and other useful considerations for an NSIR, including not duplicating existing resources.

On the medical care and research side, responses were received from: American Society for Clinical Oncology (ASCO), CDISC, and C-PATH among others. ASCO detailed the work on a variety of oncology interoperability standards and suggested areas it thought needed attention, for example, "There is a critical need for standards for genomic data, with raw data entry,

interpretation including storage algorithms and statistical algorithms, and storage to keep raw primary data for future analysis". Peter Yu, ASCO President, also offered to talk further with NIH regarding this project.

CDISC's (Clinical Data Interchange Standards Consortium) response identified metadata about standards that it considers important, and identified other relevant efforts, such as CFAST (Coalition for Accelerating Standards and Therapies), a collaboration with C-PATH; SHARE (CDISC's metadata repository effort – Shared Health and Research Electronic Library) and the IMI consortium European Translational Information & Knowledge Management Services (eTRIKS - http://www.etriks.org/). CDISC also recommended referencing other resources rather than duplicating, and provocatively added: "It is our belief that an NSIR would be most valuable if it is curated and contains complementary, non-redundant standards."

In summary, some of the responses will be useful when we are considering more details about how the NSIR should be constructed, and others will be more useful to review later as to whether they fit the criteria for inclusion for the resource. The NSIR RFI workgroup was pleased to see the variety of thoughtful responses with interesting resources/ facts and recommendations worth digging into. There were multiple statements that the NSIR should not replicate content but should take advantage of/ link to, existing resources. Given the number of standards communities and ecosystems – and we assume there are others that did not send responses -- we will need a simple and creative approach to making relevant standards and resources accessible via links, while including a layer of useful guidance to potential users, but not duplicating content. Means for including identifiers as part of the metadata, and potentially links to data (and possibly software) resources indexed in the DDI using those standards, should be included as one of the tasks in the early planning/development of the NSIR.