

Standards: awareness, information, education

Susanna-Assunta Sansone, PhD

Twitter: @SusannaASansone

ORCID: 0000-0001-5306-5690

*Associate Director,
Principal Investigator*

*Consultant,
Founding Academic Editor*



SCIENTIFIC DATA 
nature
research
SPRINGER NATURE

Coconut
FLOUR



dairy-free & gluten-free
ALTERNATIVES

Coconut Milk
YOGURT



Substitute
SOUR CREAM
with



Dairy-free
RANCH



*Too many
cooks in the
standards'
kitchen?*

**CAKE MIX
COOKIES**
12 ways







Standards – a definition

- Agreed-upon conventions for doing ‘something’, established by community consensus or an authority
 - e.g. managing a process or delivering a service



Interoperability standards – as enablers of FAIR

- Agreed-upon specifications, guidelines or criteria designed to *ensure* **data** and any other **digital object** (such as code, algorithms, workflows, models, software, or journal articles) are FAIR



SCIENTIFIC DATA 

OPEN **Comment: The FAIR Guiding Principles for scientific data management and stewardship**

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Bonten, Luiz Bonino da Silva Santos, Philip E Boume, 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, Moris A. Swertz, Mark Thompson, Johan van der Lei, Enk van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons

Open data is about MORE THAN DISCLOSURE it must be **Fair**

- Findable
- Accessible
- Interoperable
- Reusable

<http://www.nature.com/sdata/> nature publishing group 

SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18

Interoperability standards – nuts and bolts

- Enable the operational processes
 - such as exchange, aggregation, integration, comparison etc.
- Automation for both human and machine requires
 - ***metadata***: or descriptors for the digital objects
 - ***identifiers***: unique, resolvable and versionable

Interoperability standards – nuts and bolts

- Enable the operational processes
 - such as exchange, aggregation, integration, comparison etc.
- Automation for both human and machine requires
 - **metadata**: or descriptors for the digital objects
 - **identifiers**: unique, resolvable and versionable...*not the focus on my talk but.....*



THOMSON REUTERS

Open PermID
BETA

May 26, 2015

DOI 10.5281/zenodo.18003

Preprint Open Access

10 Simple rules for design, provision, and reuse of persistent identifiers for life science data

McMurry, Julie; Blomberg, Niklas; Burdett, Tony; Conte, Nathalie; Dumontier, Michel; Fellows, Donal K; Gonzalez-Beltran, Alejandra; Gormanns, Philipp; Hastings, Janna; Haendel, Melissa A; Hermjakob, Henning; Hériché, Jean-Karim; Ison, Jon C; Jimenez, Rafael C; Jupp, Simon; Juty, Nick; Laibe, Camille; Le Novère, Nicolas; Malone, James; Martin, Maria J; McEntyre, Johanna R; Morris, Chris; Muili, Juha; Müller, Wolfgang; Mungall, Christopher J; Rocca-Serra, Philippe; Sansone, Susanna-Assunta; Sariyar, Murat; Snoep, Jacky L; Stanford, Natalie J; Swainston, Neil; Washington, Nicole; Williams, Alan R; Wolstencroft, Katherine; Goble, Carole; Parkinson, Helen

Interoperability standards – invisible machinery

- Identifiers and metadata to be implemented *by **technical experts*** in tools, registries, catalogues, databases, services
 - to find, store, manage (e.g., mint, track provenance, version) and aggregate (e.g., interlink and map etc.) these digital objects
- It is essential to make standards ‘invisible’ to ***lay users***, who often have little or no familiarity with them



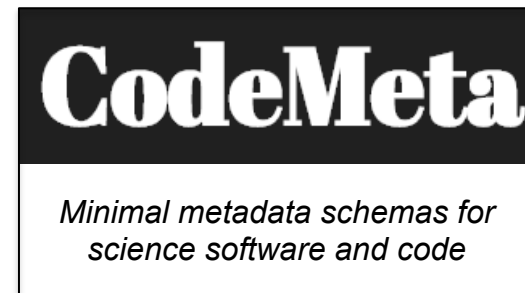
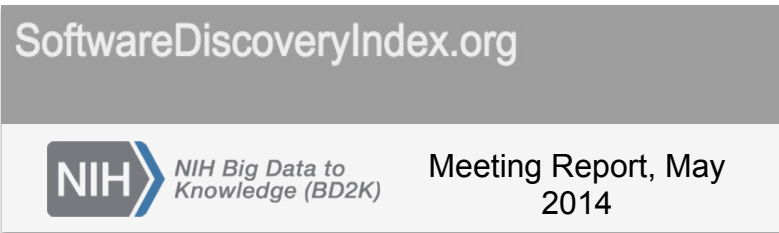
Metadata standards – fundamentals

- Descriptors for a digital object that help to understand what it is, where to find it, how to access it etc.
- The type of metadata depends also on the digital object
- The depth and breadth of metadata varies according to their purpose
 - e.g. *reproducibility* requires richer metadata then *citation*



Metadata standards - software

- Infrastructure to support their *preservation*, *discovery*, *reuse* and *attribution* lags behind that of other digital research outputs
 - Documented needs and efforts in progress, e.g.:

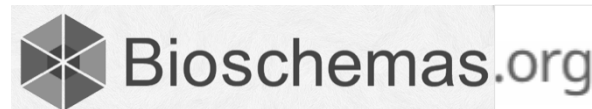


Including academics and



Metadata standards - websites and services

- Increase *discoverability* (e.g. by search engines), *aggregation* (e.g. by indices) and *analysis* of content in different websites and services

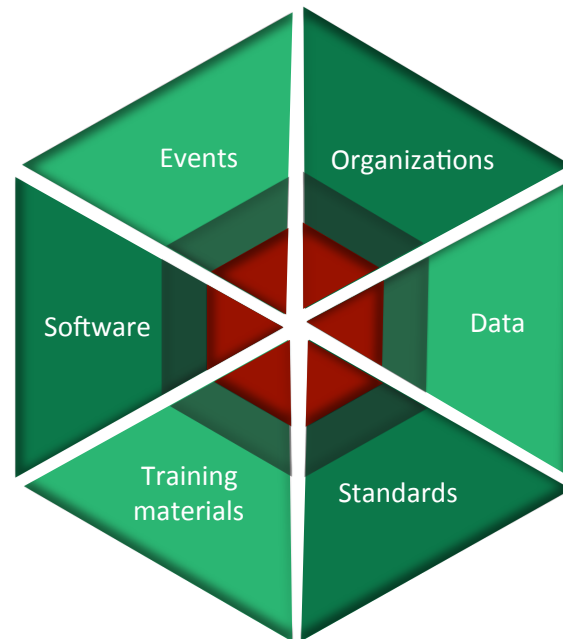


- use of Schema.org structured semantic markup (for web pages' content) by Google, Bing, Yahoo, Yandex
- coordinate its extension, where needed, in the life science area

Gaining traction and support by:



Google



Content standards – deeper metadata for datasets

- Domain-level descriptors that are essential for *interpretation*, *verification* and *reproducibility* of datasets
- The depth and breadth of descriptors vary according to the domain broadly covering the *what*, *who*, *when*, *how* and *why*



Content standards – deeper metadata for datasets

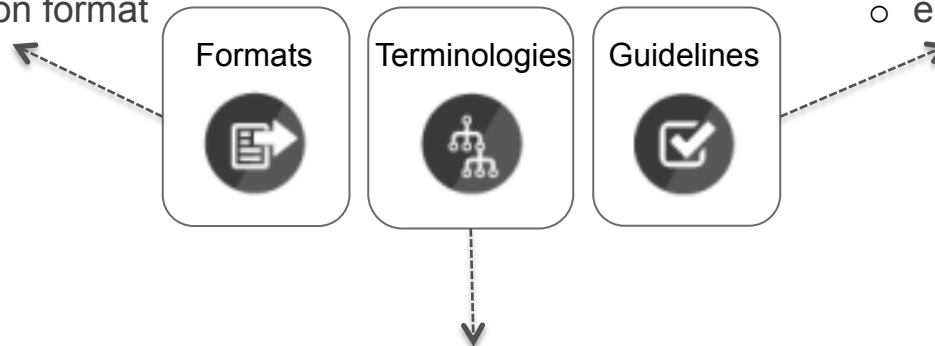
- Domain-level descriptors that are essential for *interpretation*, *verification* and *reproducibility* of datasets
- The depth and breadth of descriptors vary according to the domain broadly covering the *what*, *who*, *when*, *how* and *why* allowing:
 - experimental components (e.g., design, conditions, parameters),
 - fundamental biological entities (e.g., samples, genes, cells),
 - complex concepts (such as bioprocesses, tissues and diseases),
 - analytical process and the mathematical models, and
 - their instantiation in computational simulations (from the molecular level through to whole populations of individuals)

to be harmonized with respect to **structure**, **format** and **annotation**

Types of content standards

Conceptual model, schema, exchange formats etc

- Define the structure and interrelation of information, and the transmission format
- e.g. FASTA



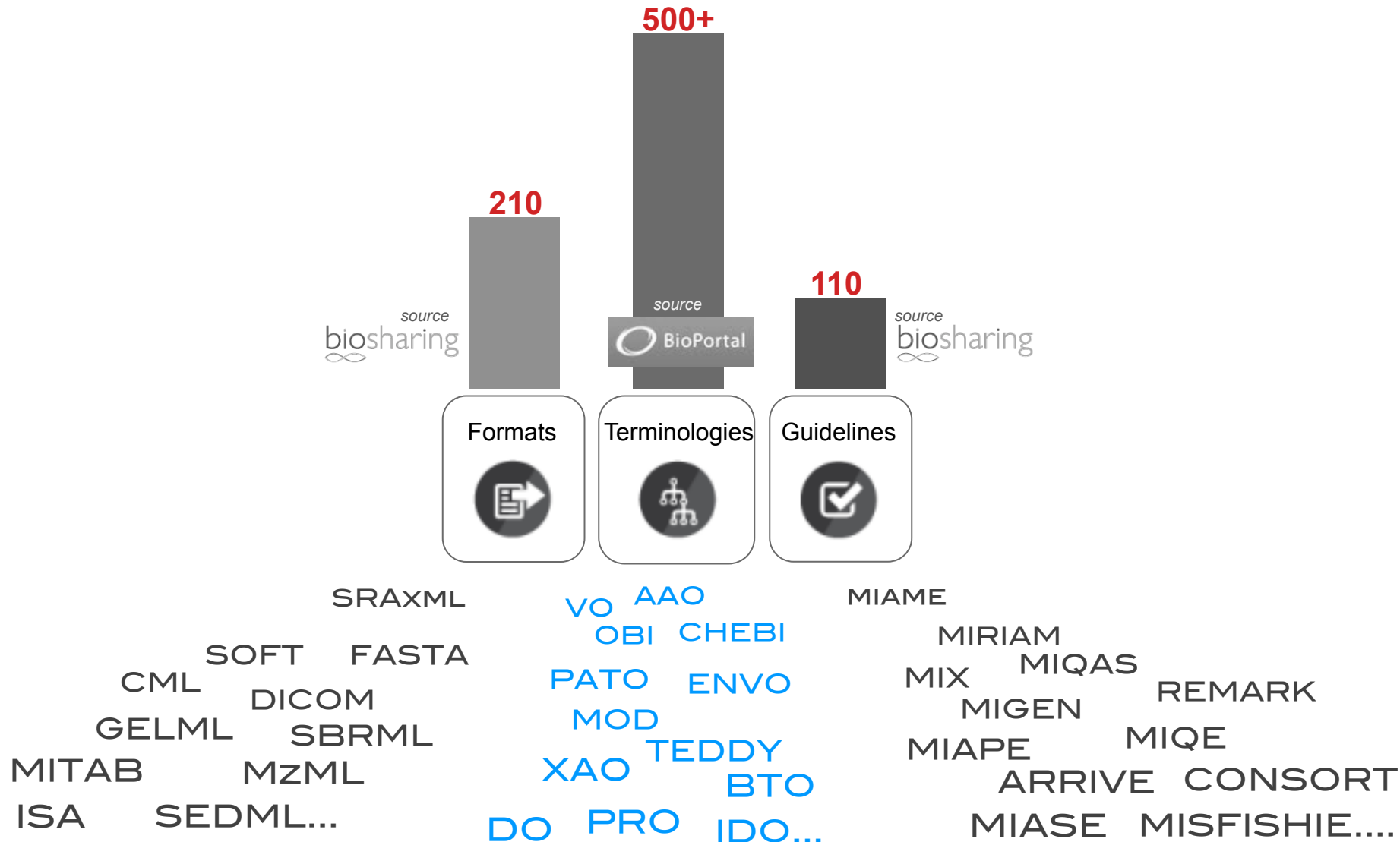
Minimum information reporting requirements, checklists

- Report the same core, essential information
- e.g. MIAME guidelines

Controlled vocabularies, taxonomies, thesauri, ontologies etc.

- Unambiguous identification and definition of concepts
- e.g. Gene Ontology

Content standards in numbers



Improving discoverability of (content) standards

- Producers
 - How do I make my standards visible to others?
- Consumers:
 - How do I find the content standards appropriate for my case?



A curated, informative and educational resource on inter-related data standards, databases, and policies in the life, environmental and biomedical sciences

Find



Recommendations

Standards and/or databases recommended by journal or funder data policies.

Discover



Collections

Standards and/or databases grouped by domain, species or organization.

Learn



Educational

About standards, their use in databases and policies, and how we can help you.



Search

☒ Standards ☒ Databases ☒ Policies ☒ Collections/Recommendations

Advanced Search



Fine grained control over your search.

Search Wizard



Let us guide you to your results.



666 Standards

Terminology Artifact	346
Model/Format	210
Reporting Guideline	110

[View all](#)



811 Databases

Protein	163
Genome	126
DNA	82

[View all](#)



81 Policies

Funder	19
Journal	56
Society	2

[View all](#)

A curated, informative and educational resource on inter-related data standards, databases, and policies in the life, environmental and biomedical sciences

Find



Recommendations

Standards and/or databases recommended by journal or funder data policies.

Discover



Collections

Standards and/or databases grouped by domain, species or organization.

Learn



Educational

About standards, their use in databases and policies, and how we can help you.

Map of the landscape, monitoring development and **evolution** of **standards**, their **use** in **databases** and the adoption of both in data **policies**



666 Standards

Terminology Artifact
Model/Format
Reporting Guideline

346
210
110

[View all](#)



811 Databases

Protein
Genome
DNA

163
126
82

[View all](#)



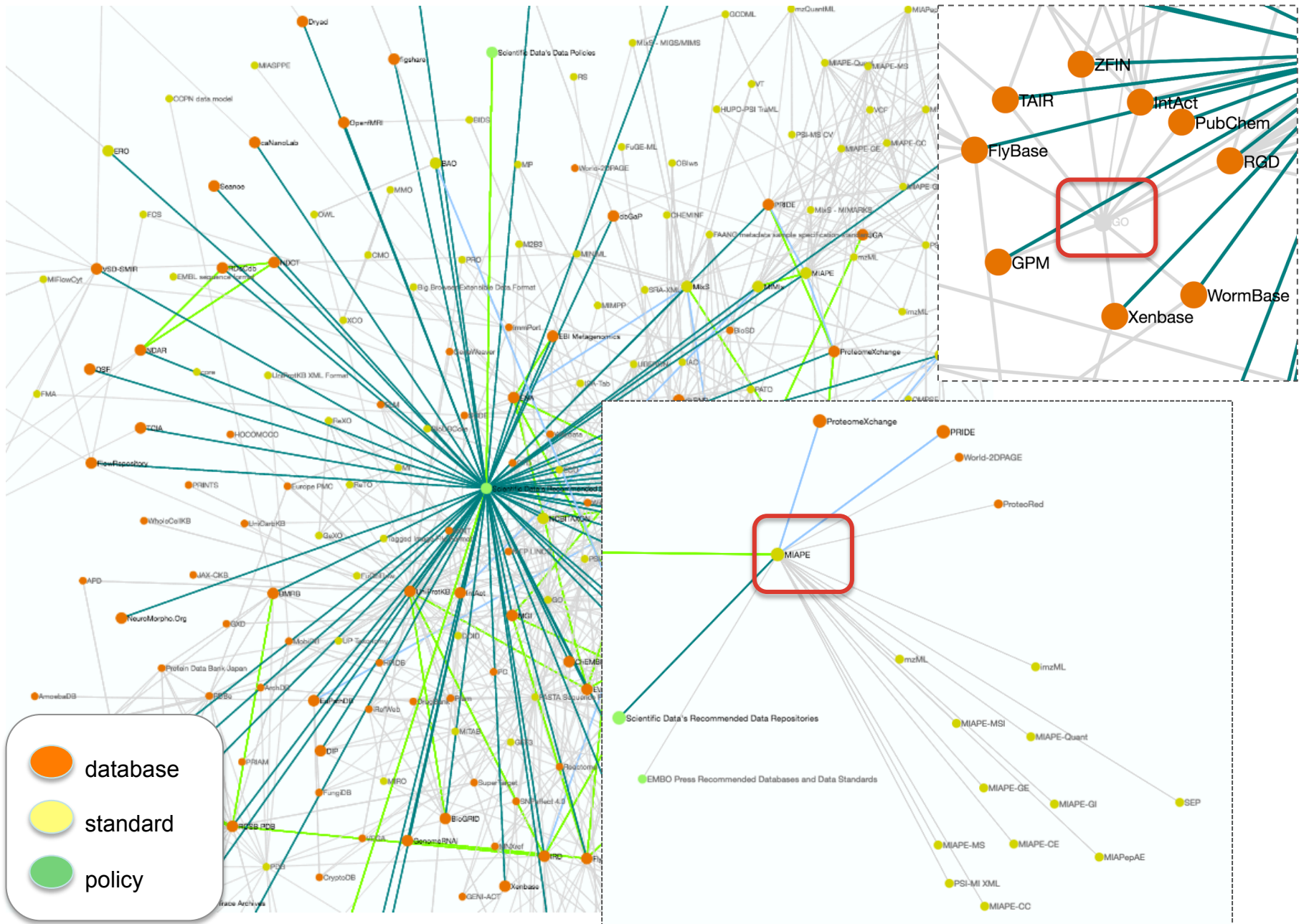
81 Policies

Funder
Journal
Society

19
56
2

[View all](#)

Interactive graph to inform and educate, e.g.



Variety of community efforts, just few examples:



- **Formal authorities**
 - openness to participations varies
 - standards are sold or licenced (at a costs or no cost)
 - charges apply to advanced training or programmatic access
- **Bottom-up communities**
 - open to interested varies
 - standards are free for use
 - volunteering efforts
 - minimal or little funds for carry out the work, let alone provide training

A complex landscape

- Perspective and focus vary, ranging:
 - from standards with a specific biological or clinical domain of study (e.g. neuroscience) or significance (e.g. model processes)
 - to the technology used (e.g. imaging modality)
- Motivation is different, spanning:
 - creation of new standards (to fill a gap)
 - mapping and harmonization of complementary or contrasting efforts
 - extensions and repurposing of existing standards
- Stakeholders are diverse, including those:
 - involved in managing, serving, curating, preserving, publishing or regulating data and/or other digital objects
 - academia, industry, governmental sectors, and funding agencies
 - producers but also also consumers of the standards, as domain (and not just technical) expertise is a must

Understanding the community process

2013

Frameworks for Community-Based Standards Efforts

September 25, 2013

Co-Chairs: Susanna Sansone, PhD and David Kennedy PhD.

[Workshop Summary](#) | [Workshop Report](#) 

The standards' life cycle:

*Susanna-Assunta Sansone, Leslie K. Derr,
David N. Kennedy and Michael F. Huerta*

[dx.doi.org/10.6084/m9.figshare.3795816.v2](https://doi.org/10.6084/m9.figshare.3795816.v2)



2015

NIH BD2K Workshop on Community-Based Data and Metadata Standards

February 25, 2015

Chairs: Melissa Haendel, Ph.D. and Christopher Chute, M.D., Dr.P.H.

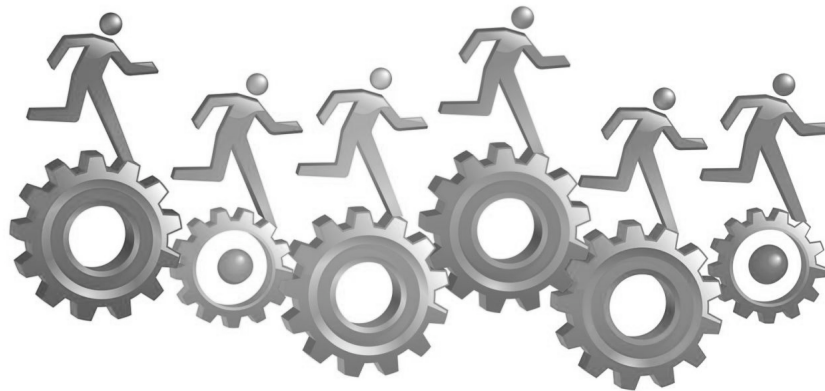
NIH Lead Organizers: Cindy P. Lawler, Ph.D.

Website: https://tools.niehs.nih.gov/conference/community-based_standards/index.cfm

Executive Summary: https://datascience.nih.gov/sites/default/files/bd2k/docs/ExecSumm_CB_DMSworkshopFEB2015.pdf

Life cycle - phases

- Formulation
 - use cases, scope, prioritization and expertise
- Development
 - iterations, tests, feedback and evaluation
 - harmonization of different perspectives and available options
- Maintenance
 - (exemplar) implementations, technical documentation, education material, metrics
 - sustainability, evolution (versions) and conversion modules

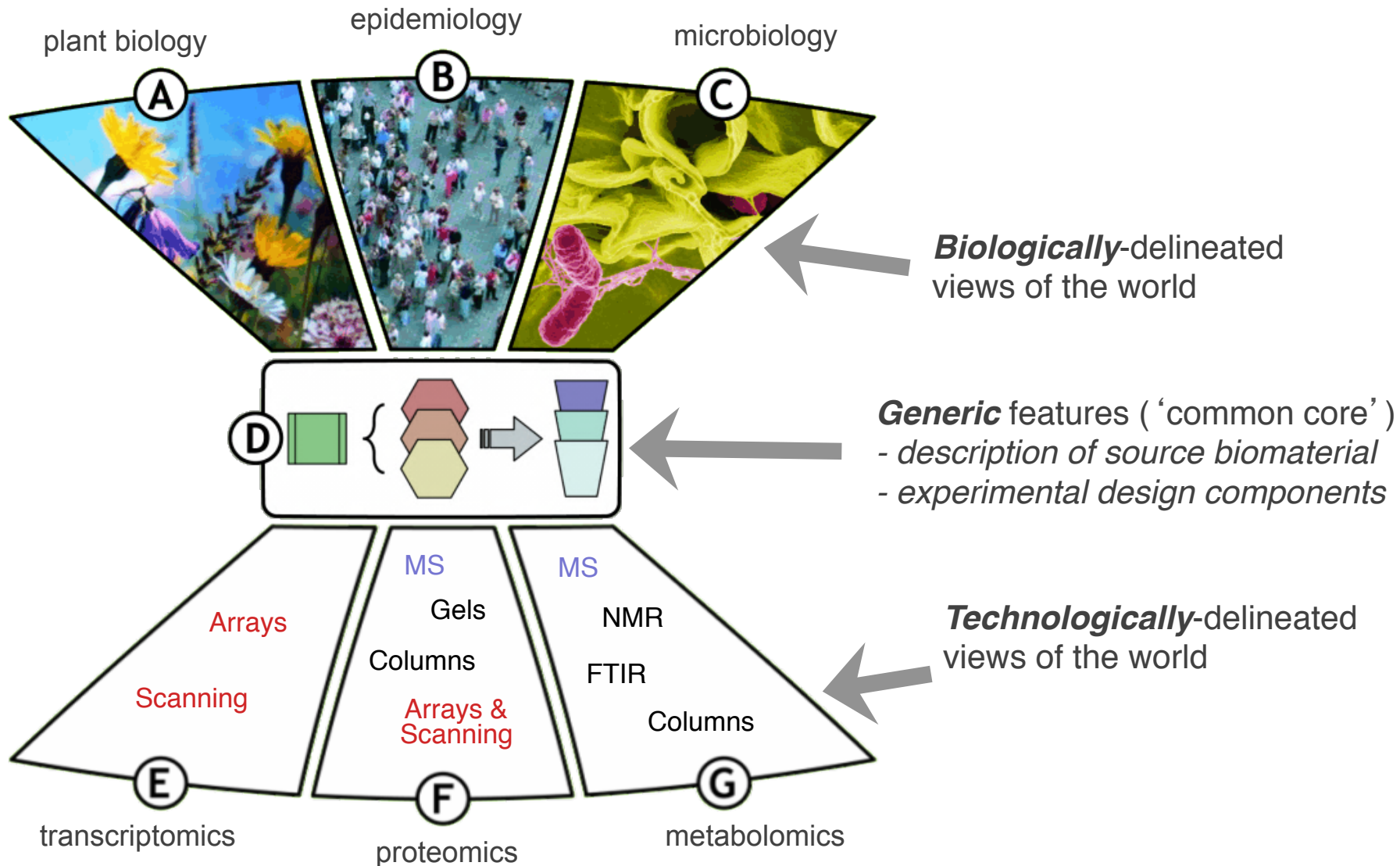


Technical and social engineering – pain points

- Fragmentation
- Coordination, harmonization, extensions
- Credit, incentives for contributors
- Governance, ownership
- Funding streams
- Indicators and evaluation methods
- Implementations: infrastructures, tools, services
- Outreach and engagement with all stakeholders
- Synergies between basic and clinical/medical areas
- Education, documentation and training
- Business models for sustainability



Fragmentation of content standards



Fragmentation of content standards

_computational
BIOLOGY

nature
biotechnology

COMMENTARY

Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project

Chris F Taylor^{1,2}, Dawn Field^{2,3}, Susanna-Assunta Sansone^{1,2}, Jan Aerts⁴, Rolf Apweiler¹, Michael Ashburner⁵, Catherine A Ball⁶, Pierre-Alain Binz^{7,8}, Molly Bogue⁹, Tim Booth², Alvis Brazma¹, Ryan R Brinkman¹⁰, Adam Michael Clark¹¹, Eric W Deutsch¹², Oliver Fiehn¹³, Jennifer Fostel¹⁴, Peter Ghazal¹⁵, Frank Gibson¹⁶, Tanya Gray^{2,3}, Graeme Grimes¹⁵, John M Hancock¹⁷, Nigel W Hardy¹⁸, Henning Hermjakob¹, Randall K Julian Jr¹⁹, Matthew Kane²⁰, Carsten Kettner²¹, Christopher Kinsinger²², Eugene Kolker^{23,24}, Martin Kuiper²⁵, Nicolas Le Novère¹, Jim Leebens-Mack²⁶, Suzanna E Lewis²⁷, Phillip Lord¹⁶, Ann-Marie Mallon¹⁷, Nishanth Marthandan²⁸, Hiroshi Masuya²⁹, Ruth McNally³⁰, Alexander Mehrle³¹, Norman Morrison^{2,32}, Sandra Orchard¹, John Quackenbush³³, James M Reecy³⁴, Donald G Robertson³⁵, Philippe Rocca-Serra^{1,36}, Henry Rodriguez²², Heiko Rosenfelder³¹, Javier Santoyo-Lopez¹⁵, Richard H Scheuermann²⁸, Daniel Schober¹, Barry Smith³⁷, Jason Snape³⁸, Christian J Stoeckert Jr³⁹, Keith Tipton⁴⁰, Peter Sterk¹, Andreas Untergasser⁴¹, Jo Vandesompele⁴² & Stefan Wiemann³¹

doi: 10.1038/nbt.1411



110

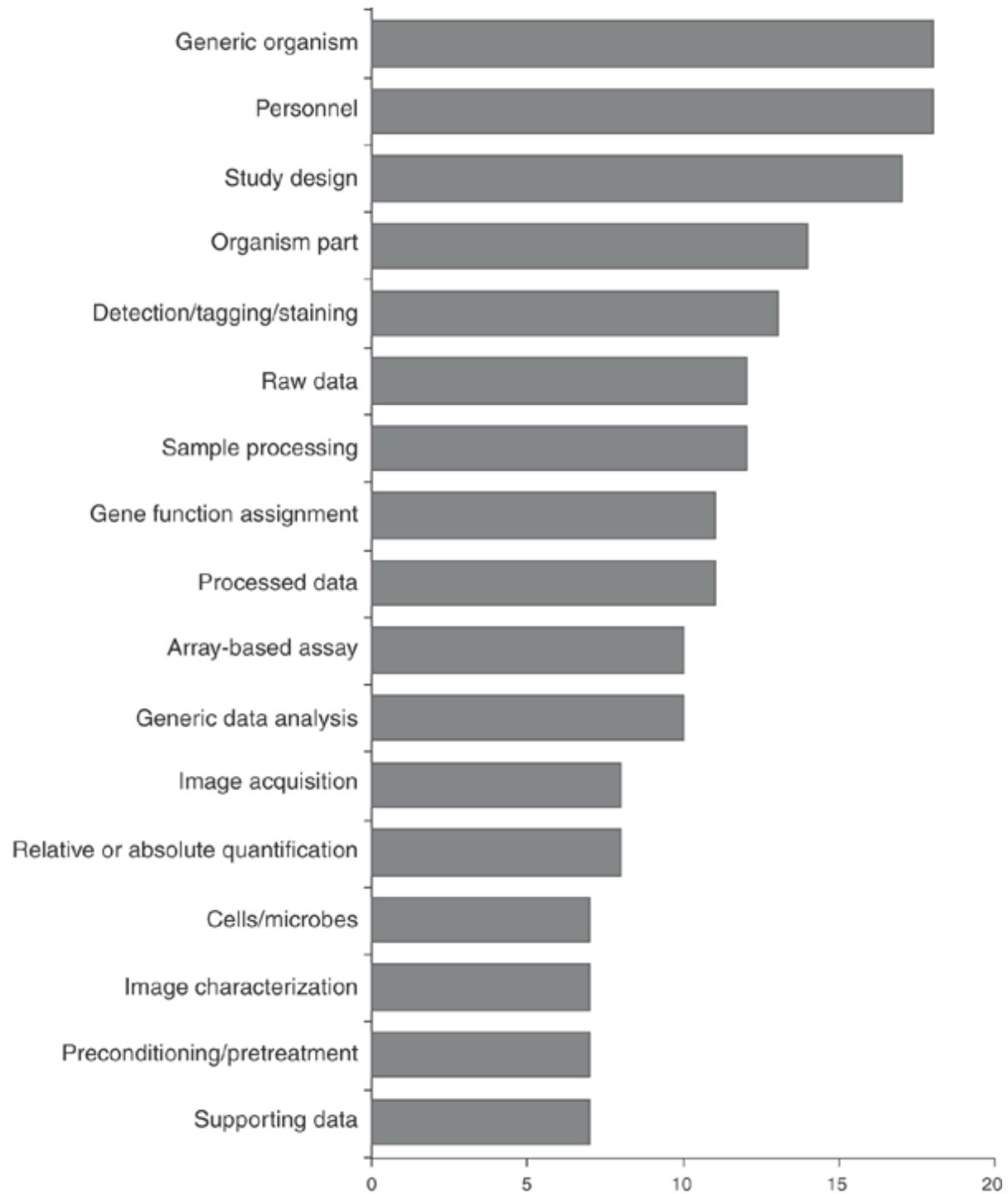
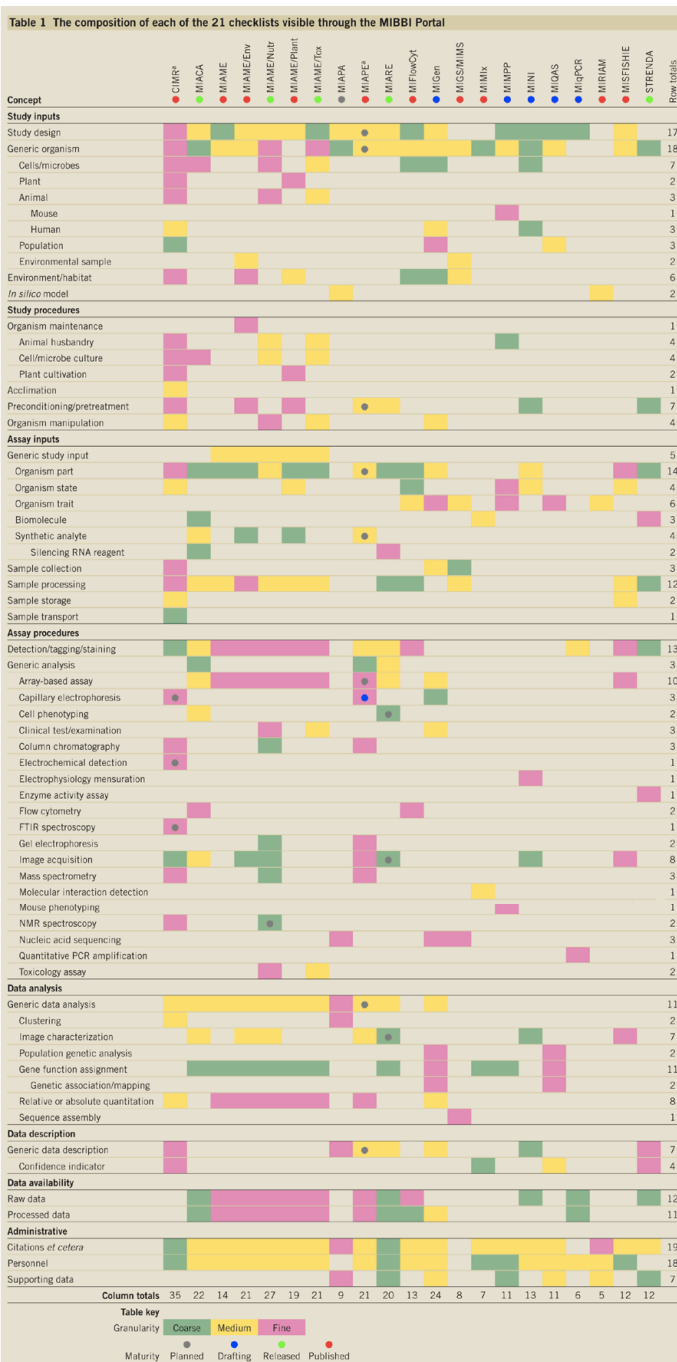


Guidelines



MIAME

MIRIAM
MIX MIQAS
MIGEN REMARK
MIAPE MIQE
ARRIVE CONSORT
MIASE MISFISHIE....

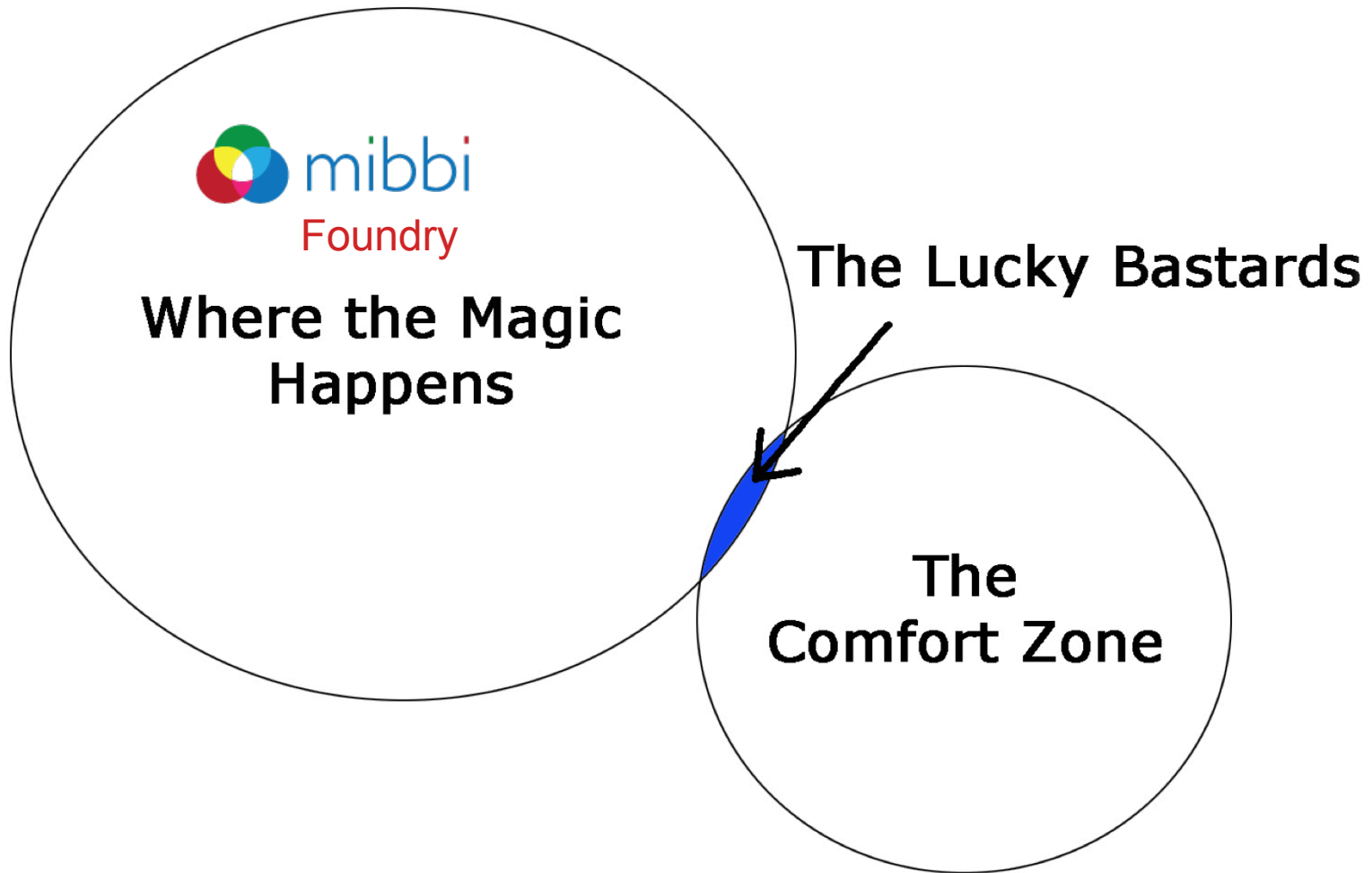


Concepts (row headings) were derived as described in the Methods section, constructed exclusively for the purpose of identifying overlaps between checklists. Because the concepts vary widely in breadth of scope, the number of concepts addressed by any one project is not necessarily indicative of the extent of that project's actual guidelines. Color codings of cells and bullets indicate granularity of coverage and developmental status, respectively. Some bullets have been placed within the matrix itself to provide a free-grained view of developmental status, cells lacking bullets inherit the overall status of their project (that is, the bullet at that column's head). Row and column totals (counting presence or absence only) are provided in the rightmost column and bottom row. Analyses of these data are provided in [Figure 1](#) and in [Supplementary Figures 1 and 2](#).

*The specification is provided as a suite of related documents

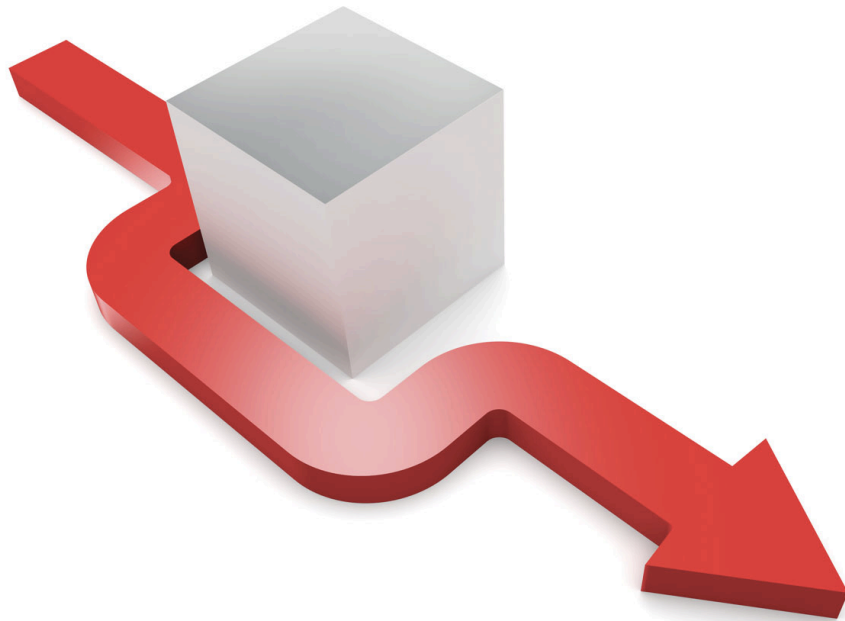


Harmonization is the magic word...until it is not



Working in/across multiple domains is challenging

- Requires
 - Mapping between/among heterogeneous representations





MAGE-TAB

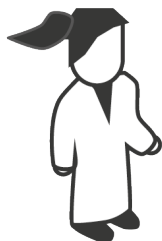


ARRAYEXPRESS

Transcriptomics data files &
relevant experimental descrip-
tors



MAGE-TAB



PRIDE ML

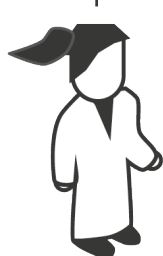


PRIDE

Proteomics data files & relevant
experimental descriptors



PRIDE ML



SRA XML



ENA 
European Nucleotide Archive

Sequence data files & relevant
experimental descriptors



SRA XML



Sample-Tab



 **BioSamples**

Biological sample information



Sample-Tab





The NIH BD2K biomedical and health**C**are **D**ata **D**iscovery **I**ndex **E**cosystem
Do for data what PubMed did for literature



To help users **find** and **access** shared datasets
available in public databases



under grant 1U24AI117966-01.

University of California, San Diego

University of Texas Houston

University of Michigan

University of Oxford

Engaging The Community Toward a Data Discovery Index (v1.0)

Search For Data Through BioCADDIE



☒ Search for data set ☐ Search for repository

[Advanced Search](#) [help](#)

Statistics



23 REPOSITORIES



10 DATA TYPES

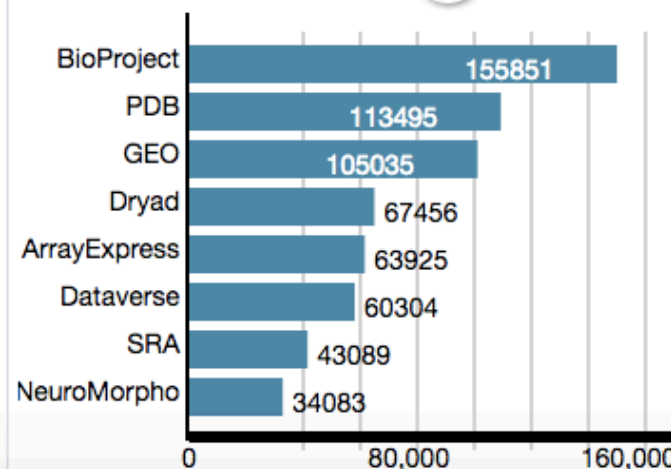


649055 DATASETS

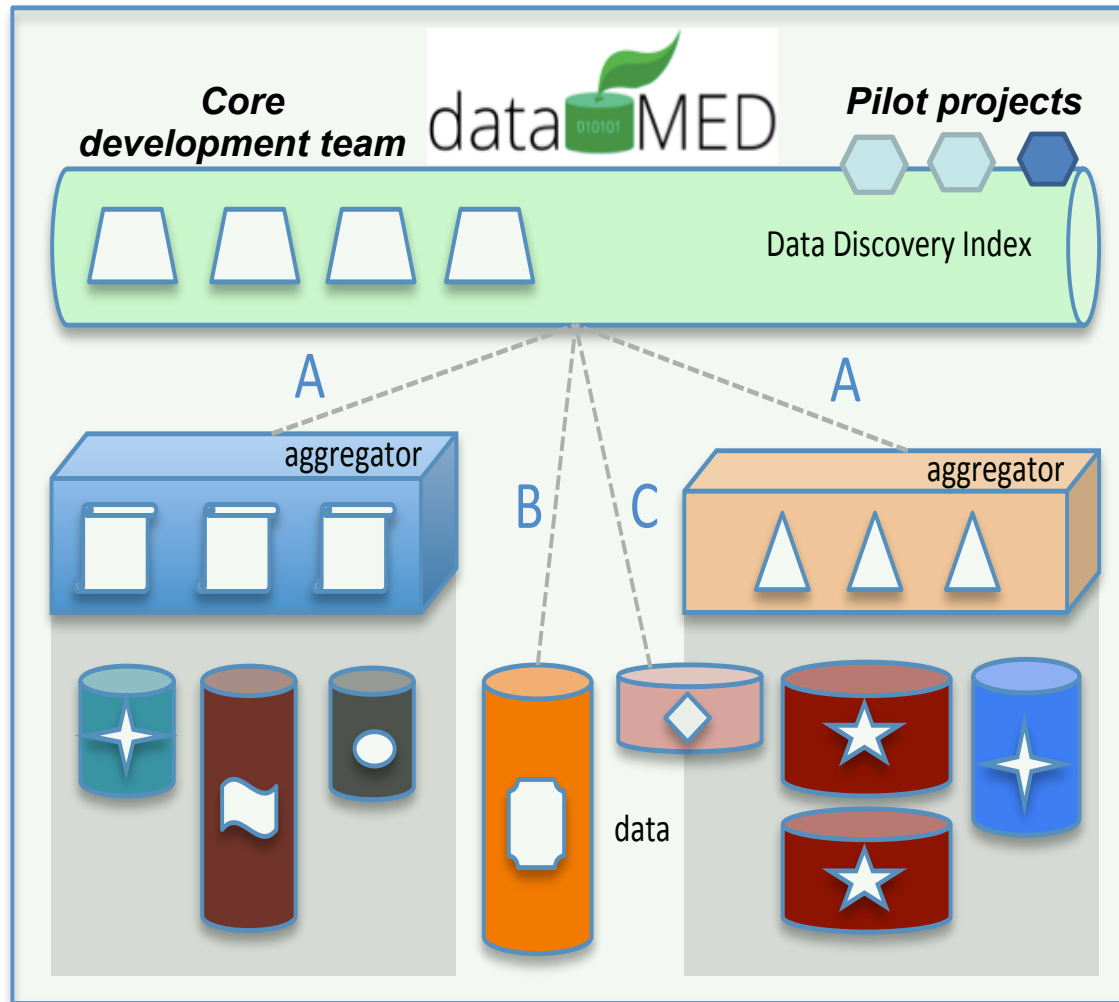


4 PILOT PROJECTS

Top 8 Repositories



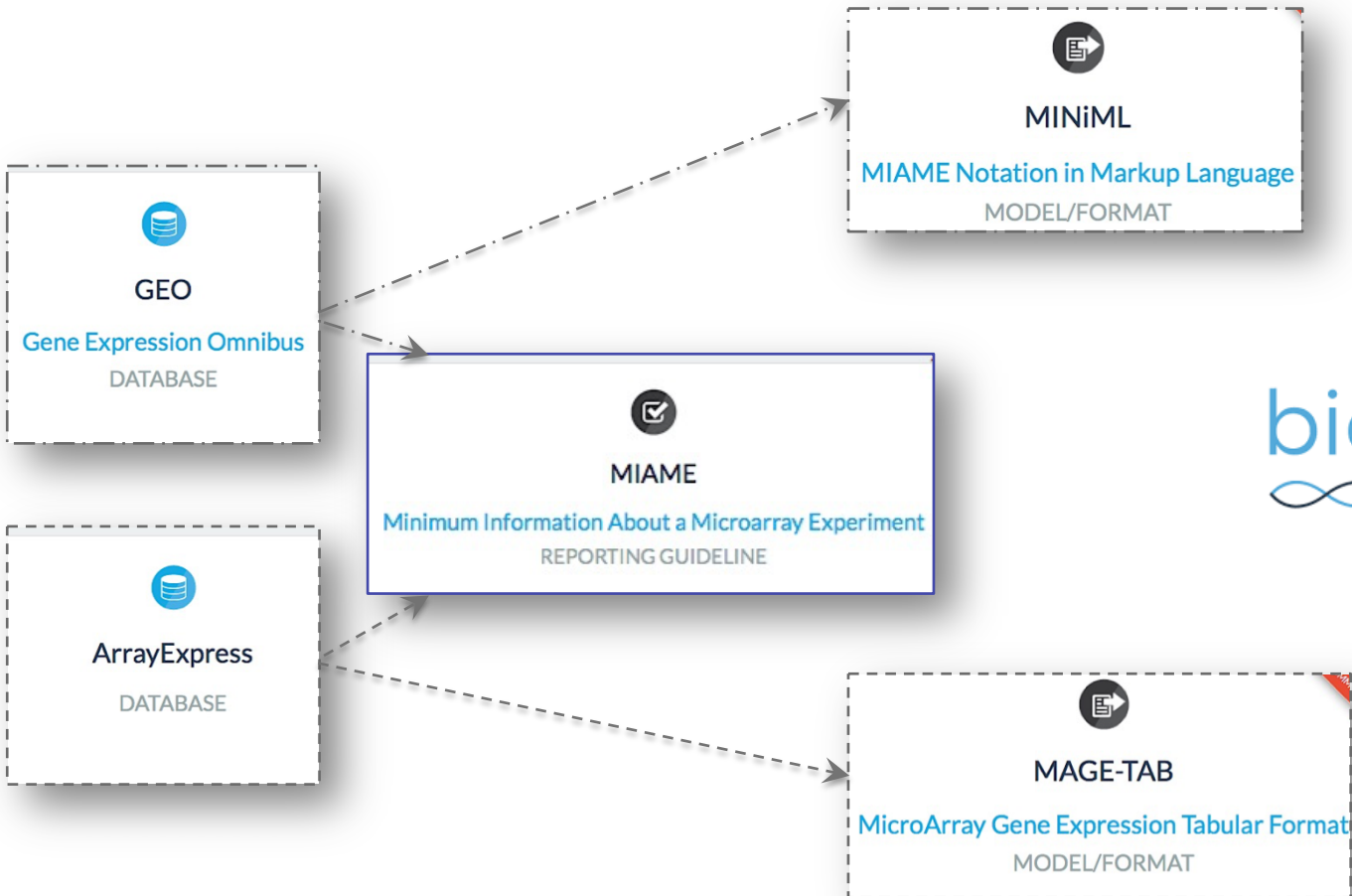
Mapping, mapping, mapping and more mapping



Aggregators:
repositories or various indices

Data:
digital research objects

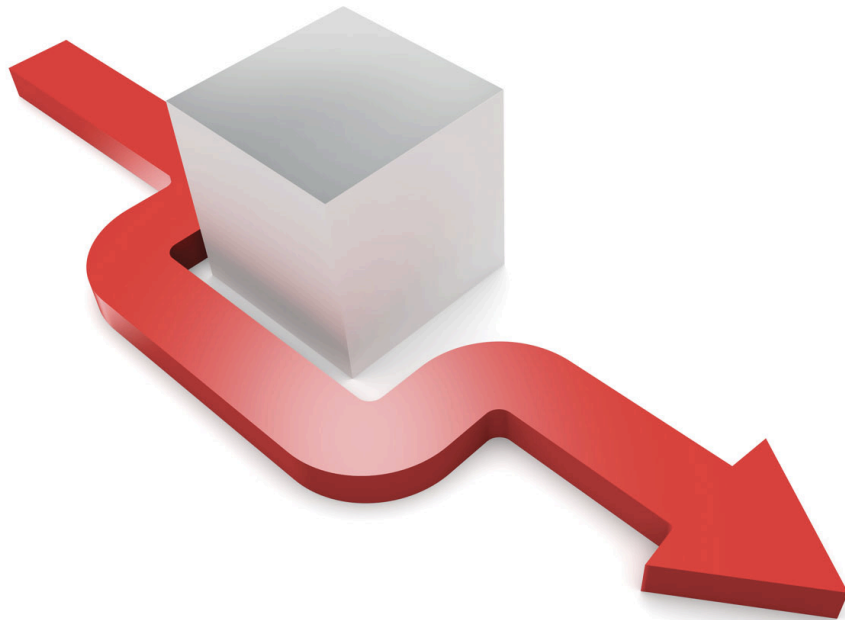
Knowing which (content) standard a repository implement is key - just a simplified example:

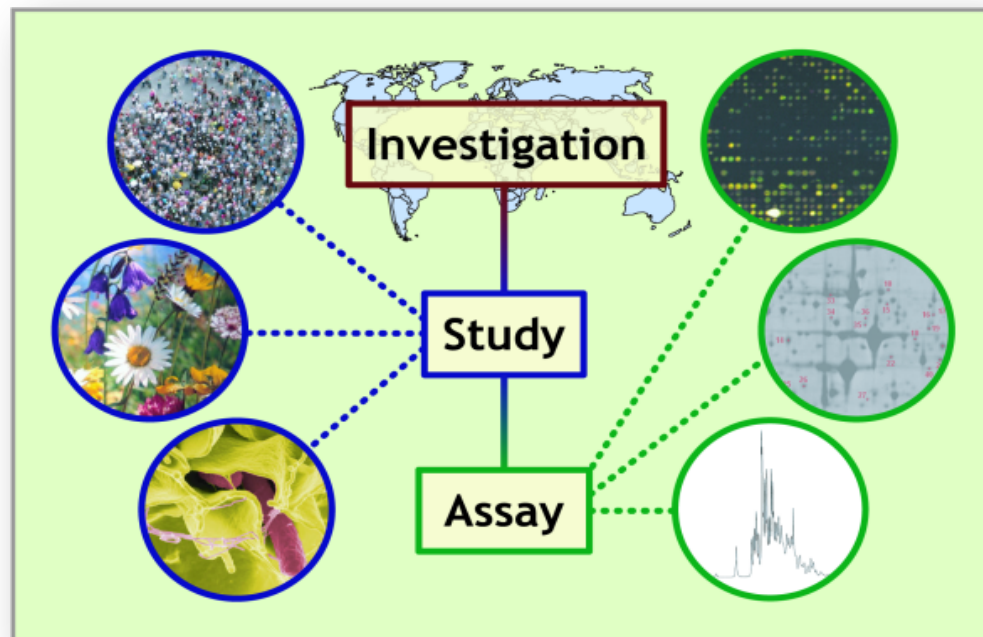


biosharing.org
standards, databases, policies


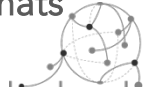
Working in/across multiple domains is challenging

- Requires
 - Mapping between/among heterogeneous representations
 - Conceptual modelling framework to encompass the domain specific content standards
 - Tools to handle customizable annotation, multiple conversions and validation





isatools
isa-tools.org

isa model and related formats
isatab  **isatordf**  **linkedisa**



collect and curate, following standards



store and browse, locally or publicly



submit to public repositories



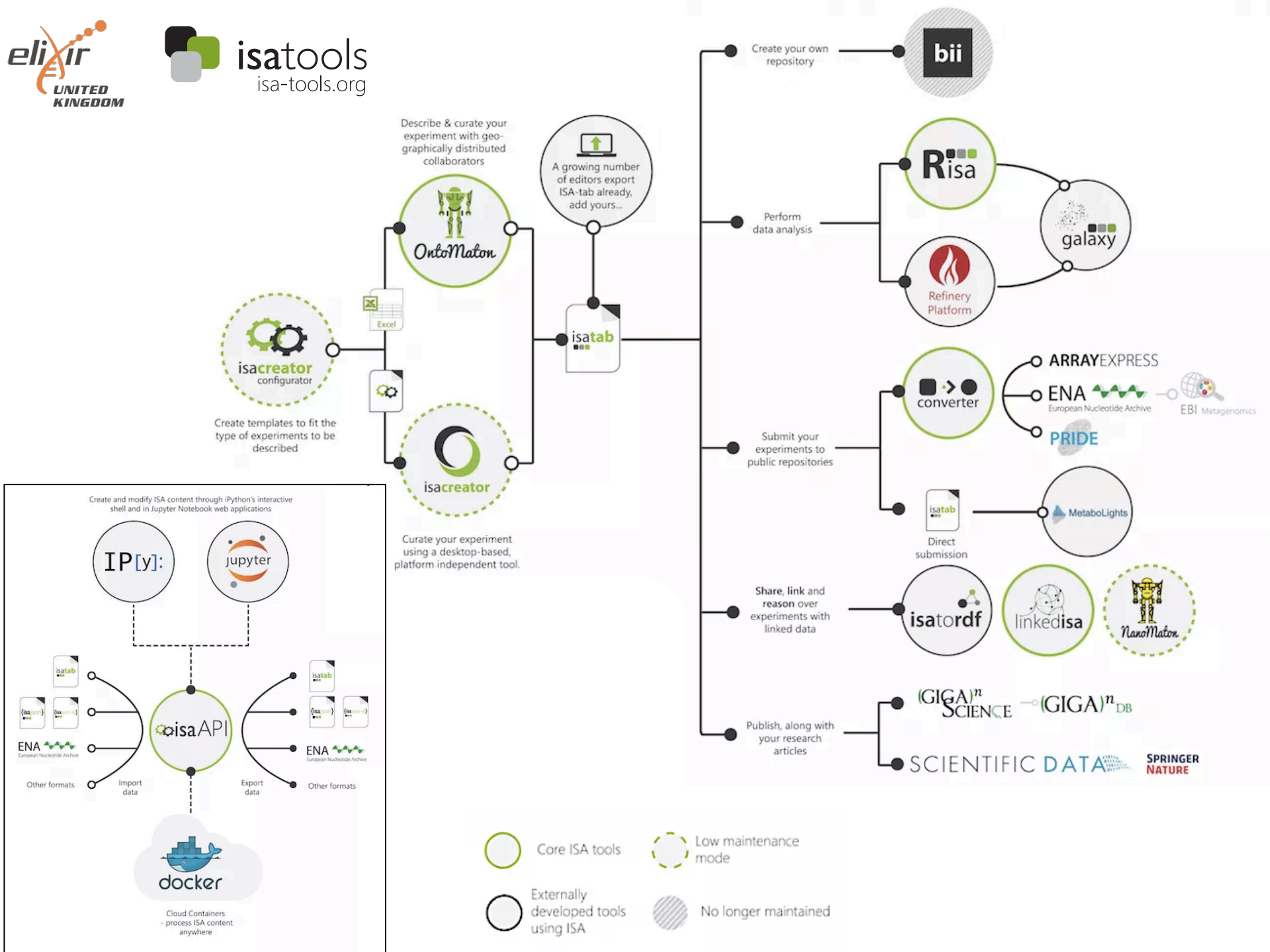
analyse with existing tools



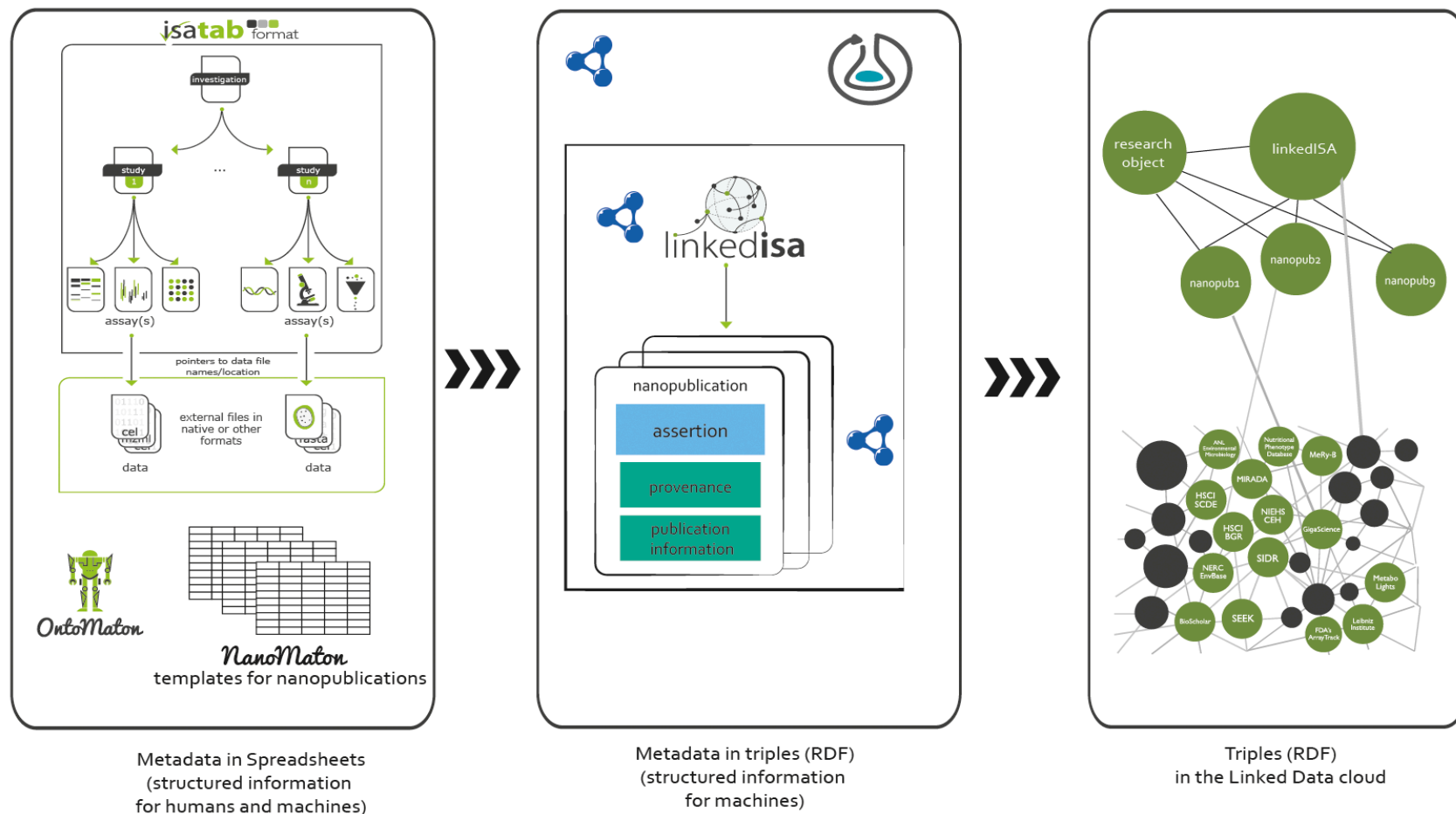
release, reason and nanopublish



publish data along your article



Complementary roles of RO, ISA and nanopublications



From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing:
The Complementary Roles of Data Models and Workflows in
Bioinformatics



Search



E-alert



Submit

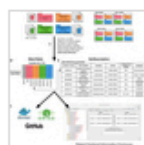


Login

Extending the studyforrest resource for brain imaging research

Data Descriptor | 25 October 2016 | [OPEN](#)

A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey



Chirag J. Patel, Nam Pho [...] Paul Avillach

Announcement

An open approach to Huntington's disease research

Oct 19 | Rachel Harding explains why she is working in the open, how openness can speed scientific progress. ... [show more](#)



Announcement

Data Matters: Interview with Ben Lehner

Oct 19 | Ben Lehner talks about his experiences accessing and using human genome data, and argues that a change in... [show more](#)



Find out more about Scientific Data

Read our Aims
& Scope

SCIENTIFIC DATA

Discover data associated with our content

 **isaexplorer**
a data discovery tool

Find the right repository for your data

Access our Recommended
Repositories list

SCIENTIFIC DATA

Search



Data Repositories

figshare 36

Dryad Digital Repository 31

Gene Expression Omnibus 17

[Show all](#) [Show next 5](#)

Designs

observation design 58

time series design 33

data integration objective 25

[Show all](#) [Show next 5](#)

Measurement Types

transcription profiling assay 16

protein expression profiling 8

nuclear magnetic resonance assay 8

[Show all](#) [Show next 5](#)


Technology Types

Developed and hosted at




 20/12/2013 

Carine Poussin et al 


The species translation challenge—A systems biology perspective on human and rat bronchial epithelial cells 


 Data Repositories 

2

CELL BIOLOGY MOLECULAR EVOLUTION
RESEARCH DATA GENE REGULATORY NETWORKS 


 22/12/2014 

Katherine E. Battle et al 

Global database of matched Plasmodium falciparum and P. vivax incidence and prevalence records from 1985–2013 


 Data Repositories 

1

EPIDEMIOLOGY MALARIA LITERATURE MINING 


 11/04/2014 

Ricco Lindner et al 

DNA methylation temporal profiling following peripheral versus central nervous system axotomy 

 21/03/2014 

Julian Maclaren et al 

Reliability of brain volume measurements: A test-retest dataset 

 Data Repositories 

1

Search



Data Repositories

figshare	36
Dryad Digital Repository	31
Gene Expression Omnibus	17

[Show all](#) [Show next 5](#)

Designs

observation design	58
time series design	33
data integration objective	25

[Show all](#) [Show next 5](#)

Measurement Types

transcription profiling assay	16
protein expression profiling	8
nuclear magnetic resonance assay	8

[Show all](#) [Show next 5](#)

Technology Types

Developed and hosted at



20/12/2013

Carine Poussin et al

The species translation challenge—A systems biology perspective on human and rat bronchial epithelial cells

Data Repositories

figshare	36
Dryad Digital Repository	31
Gene Expression Omnibus	17

[Show all](#) [Show next 5](#) [Res](#)

Designs

observation design	58
time series design	33
data integration objective	25

Measurement Types

transcription profiling assay	16
protein expression profiling	8
nuclear magnetic resonance assay	8

[Show all](#) [Show next 5](#)

Technology Types

data collection method	14
DNA microarray	9
MRI Scanner	8

22/12/2014

Katherine E. Battle et al

Global database of matched Plasmodium falciparum and P. vivax incidence and prevalence records from 1985–2013

Data Repositories

Factor Types

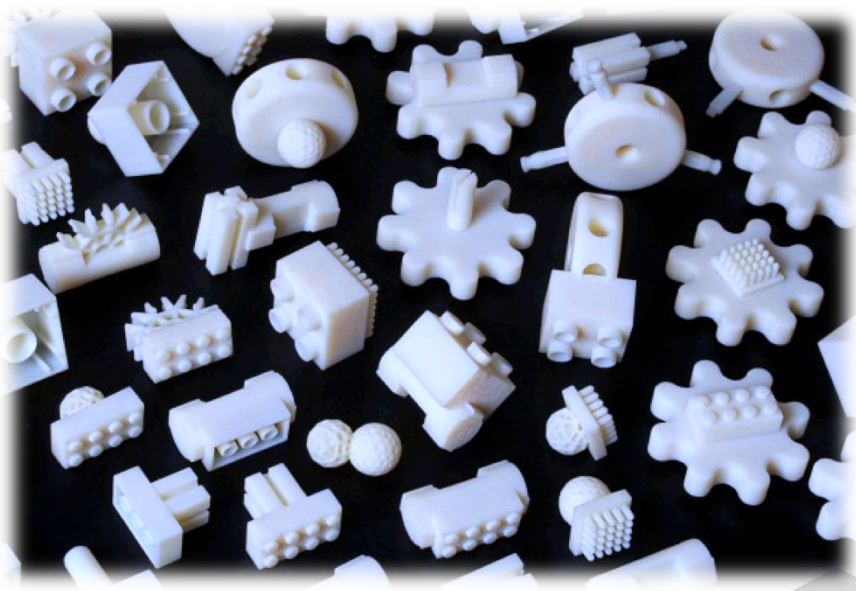
biological replicate	6
observation period	6
developmental stage	4

[Show all](#) [Show next 5](#)

Organisms

Homo sapiens	47
Mus musculus	18
Rattus norvegicus	9

[Show all](#) [Show next 5](#)



October 2016

Interoperability Standards - Digital Objects in Their Own Right

Susanna-Assunta Sansone and Philippe
Rocca-Serra



“As Data Science culture grows, digital research outputs (such as data, computational analysis and software) are being established as first-class citizens.

This cultural shift is required to go one step further: to recognize interoperability standards as digital objects in their own right, with their associated **research, development and educational activities**”.

Sansone, Susanna-Assunta; Rocca-Serra, Philippe (2016). Interoperability Standards - Digital Objects in Their Own Right. Wellcome Trust”

<https://dx.doi.org/10.6084/m9.figshare.4055496.v1>