biosharing.org
Information Resources

BioSharing – mapping the metadata standards in the life sciences

Peter McQuilton, PhD

BioSharing Content Lead

RDA

UNIVERSITY OF OXFORD

Outline

1. The sea of standards in the life sciences

1. What is BioSharing?

2. Work in progress

3. BioSharing WG and potential collaborations

There are more than 600 standards in the life sciences

Databases and tools implementing Standards; also training material on and around standards

formats

terminologies

guidelines

MAGE-Tab
GCDML
SRAxml
SOFT
FASTA
CML
DICOM
GELML
SBRML
MITAB
MzML
ISA-Tab
SEDML…

AAO
CHEBI
OBI
VO
PATO
ENVO
MOD
TEDDY
XAO
BTO
DO
PRO
IDO…

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

Mapping Standards and Databases in the Life Sciences

"Which standard should I use for this data, considering I'd like to publish in X journal?

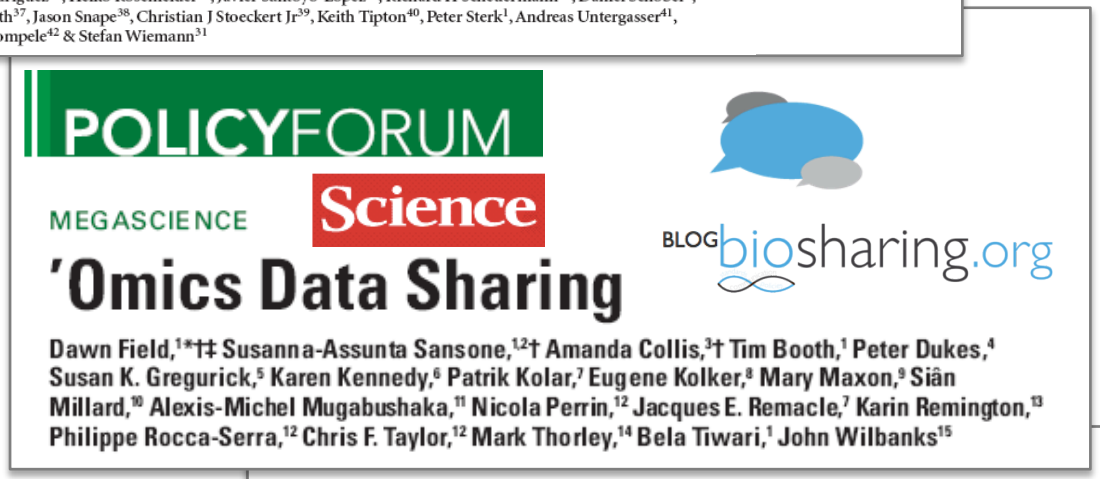
"Are we using the most up-to-date version of this standard?"

"My data is in X format, which databases take that format?

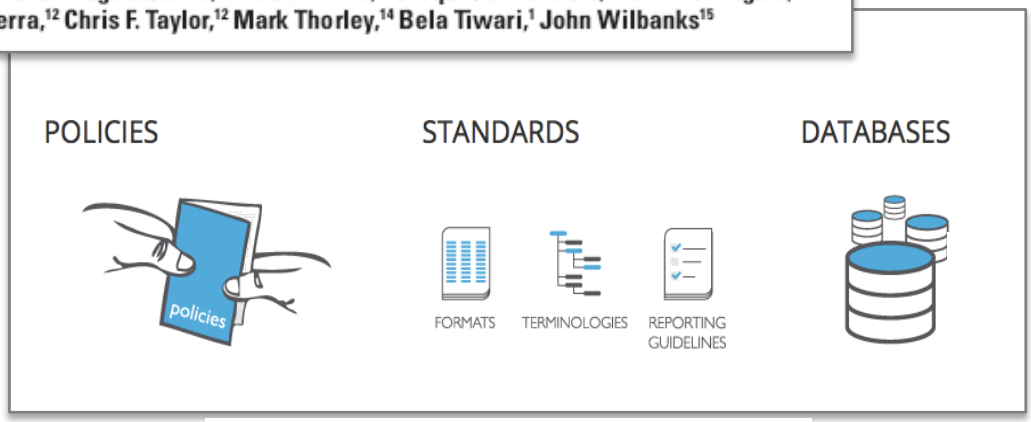
biosharing.org

Information Resources

(2008)



(2009)

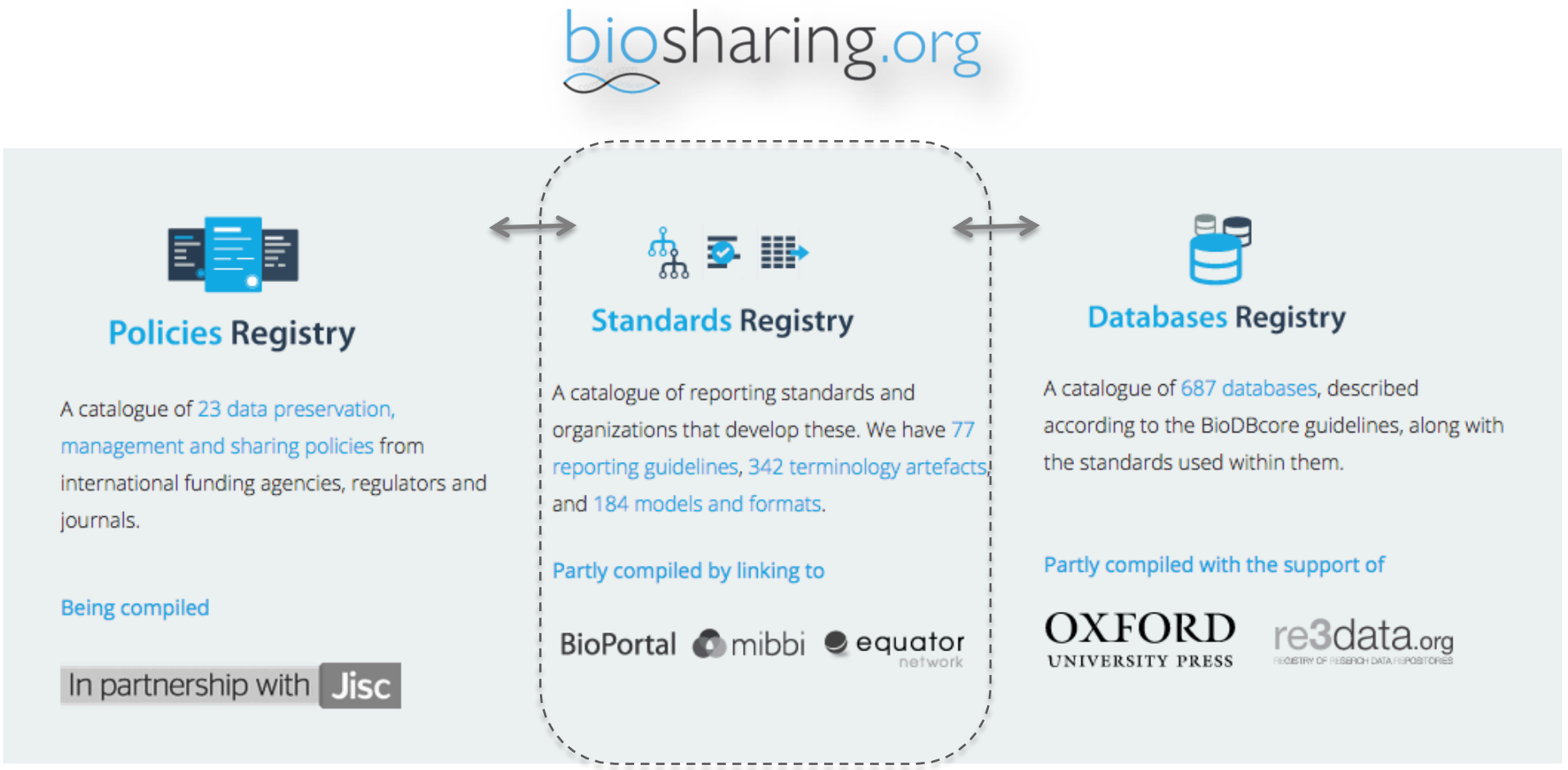


biosharing.org

(2011)

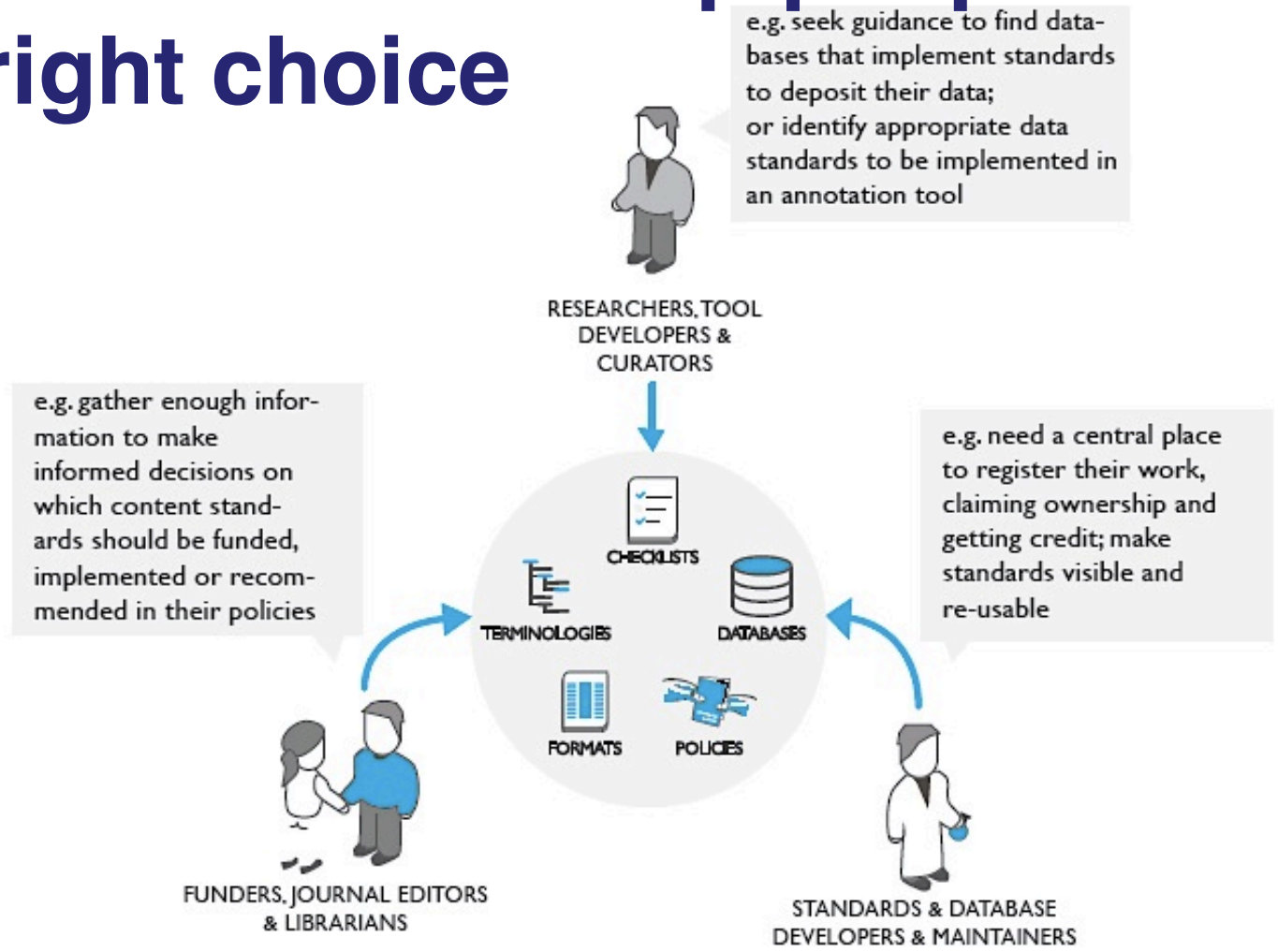
- registry development,
- content collection,
- community engagement,
- funds raising
- creation of a WG in RDA and Force11; Advisory Board
- links with ELIXIR and NIH BD2K

(2015)



A **web-based**, **curated** and **searchable registry** ensuring that biological standards and databases are *registered*, *informative* and *discoverable*; also monitoring the development and evolution of standards, their use in databases and the adoption of both in data policies.

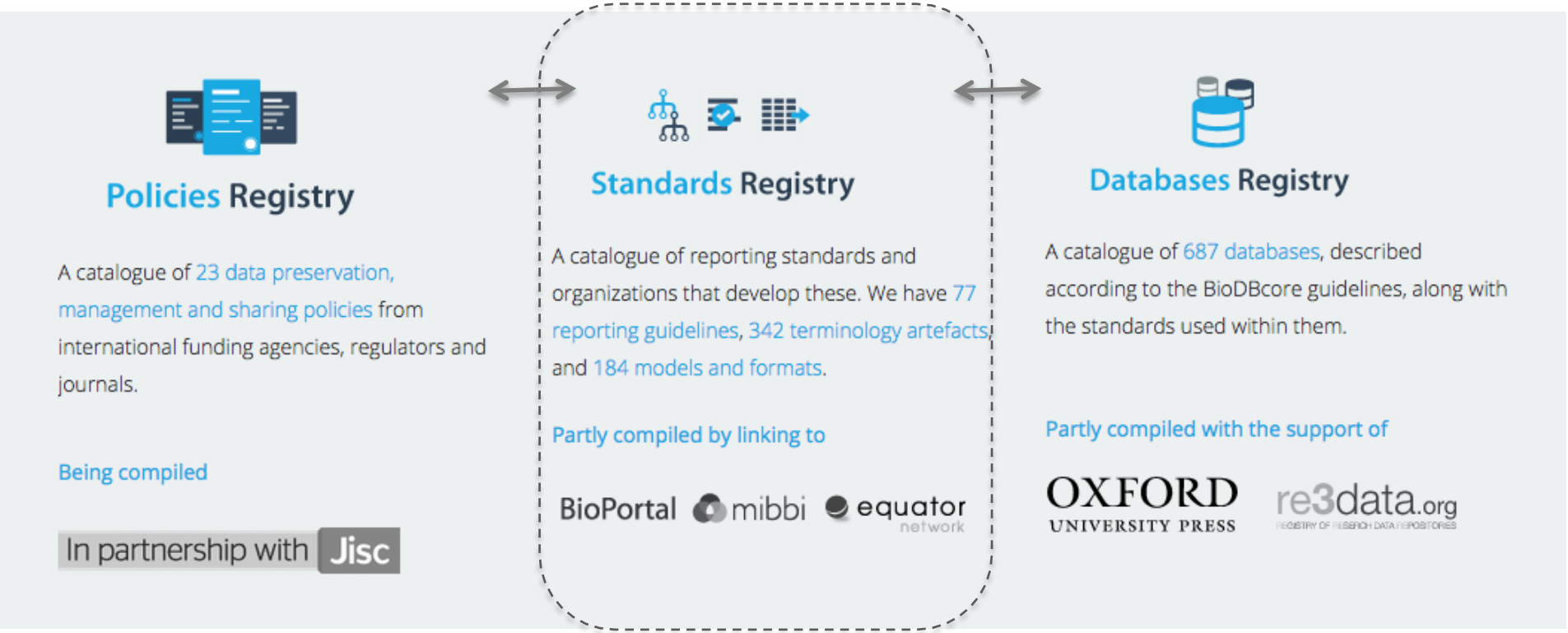
Our mission: To help people make the right choice



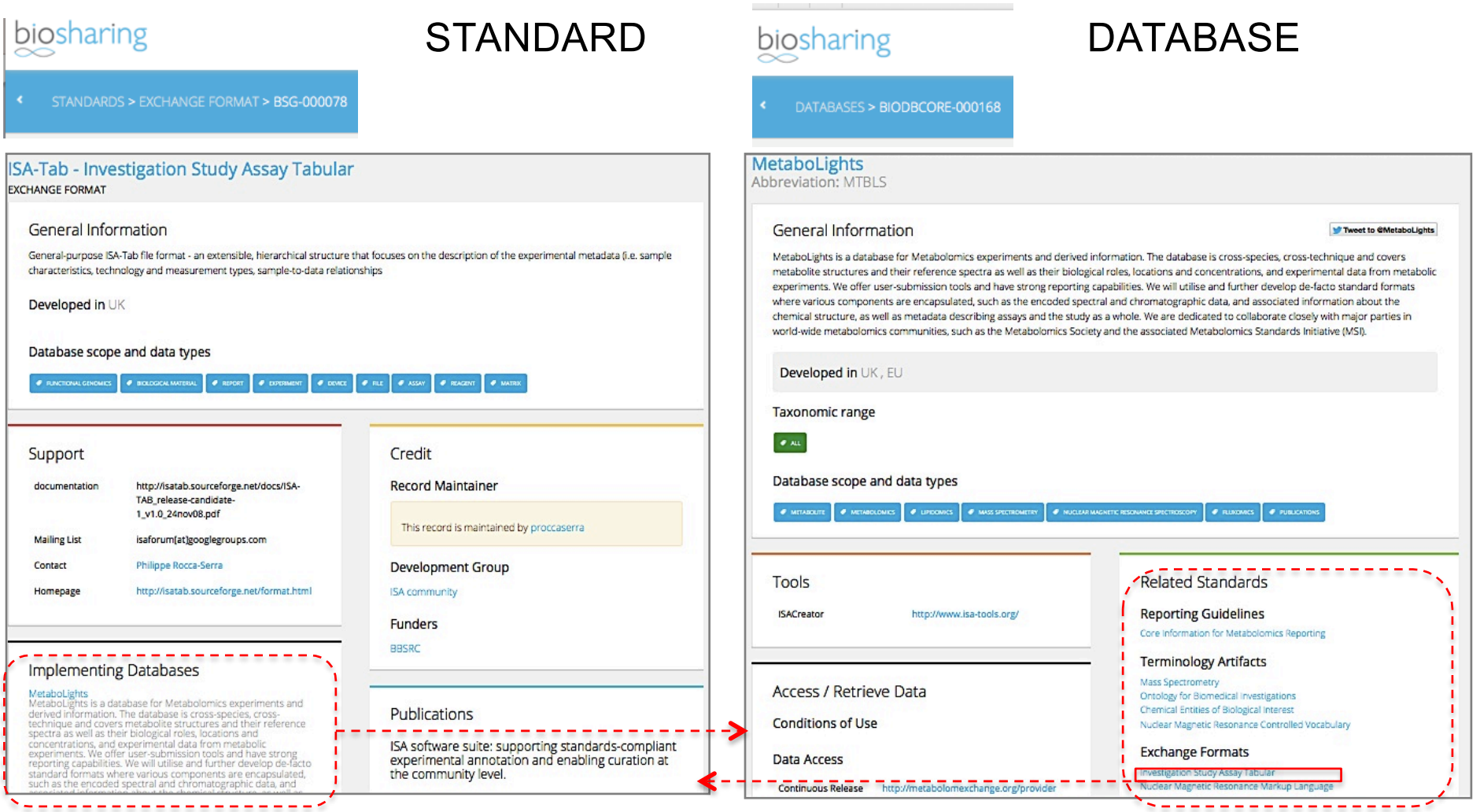
Researchers, ***developers*** and ***curators*** lack support and guidance on how to best navigate and select content standards, understand their maturity, or find databases that implement them;

Funders, ***journals*** and ***librarians*** do not have enough information to make informed decisions on which content standards or database to recommended in policies, or fund or implement

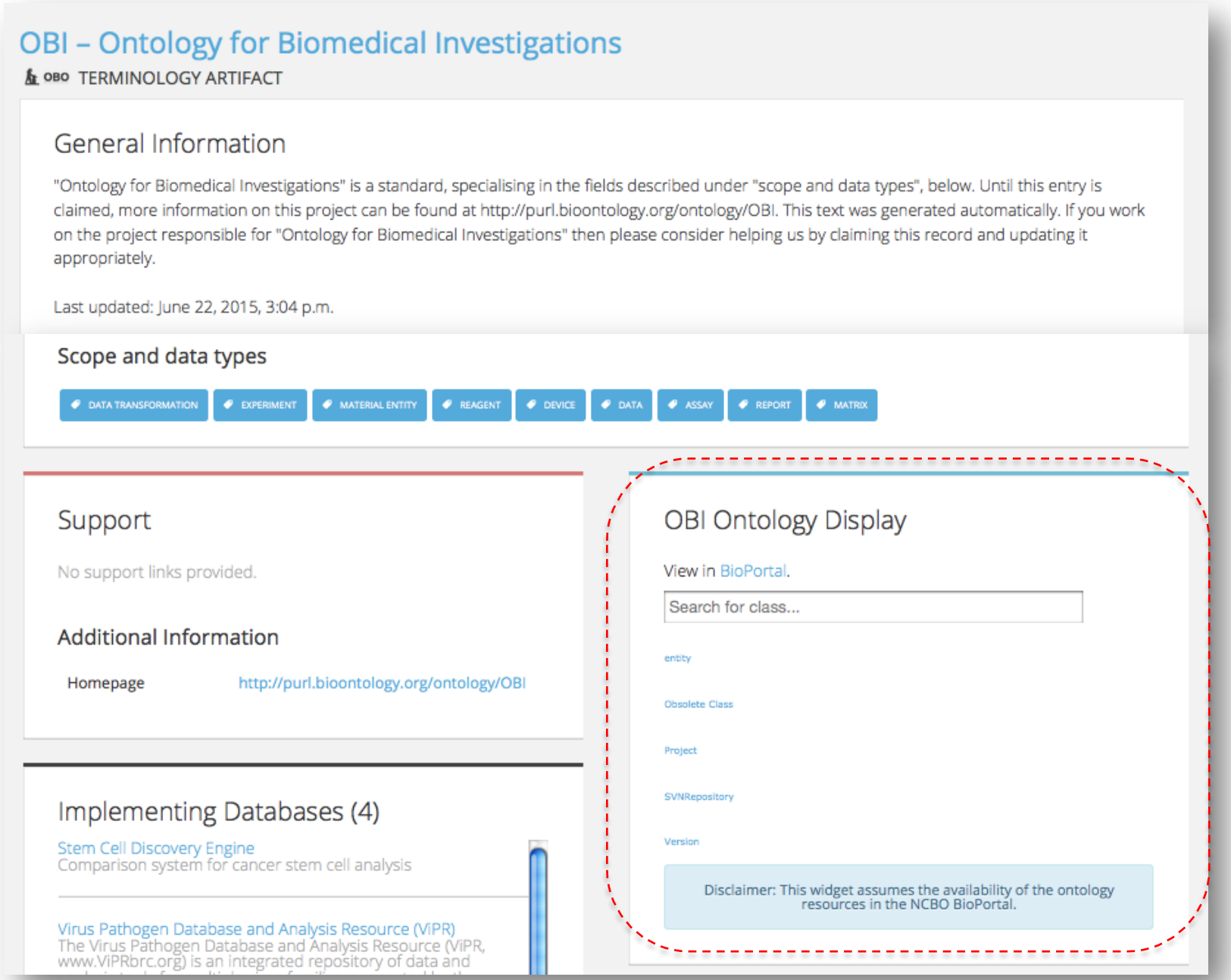
Three interlinked registries



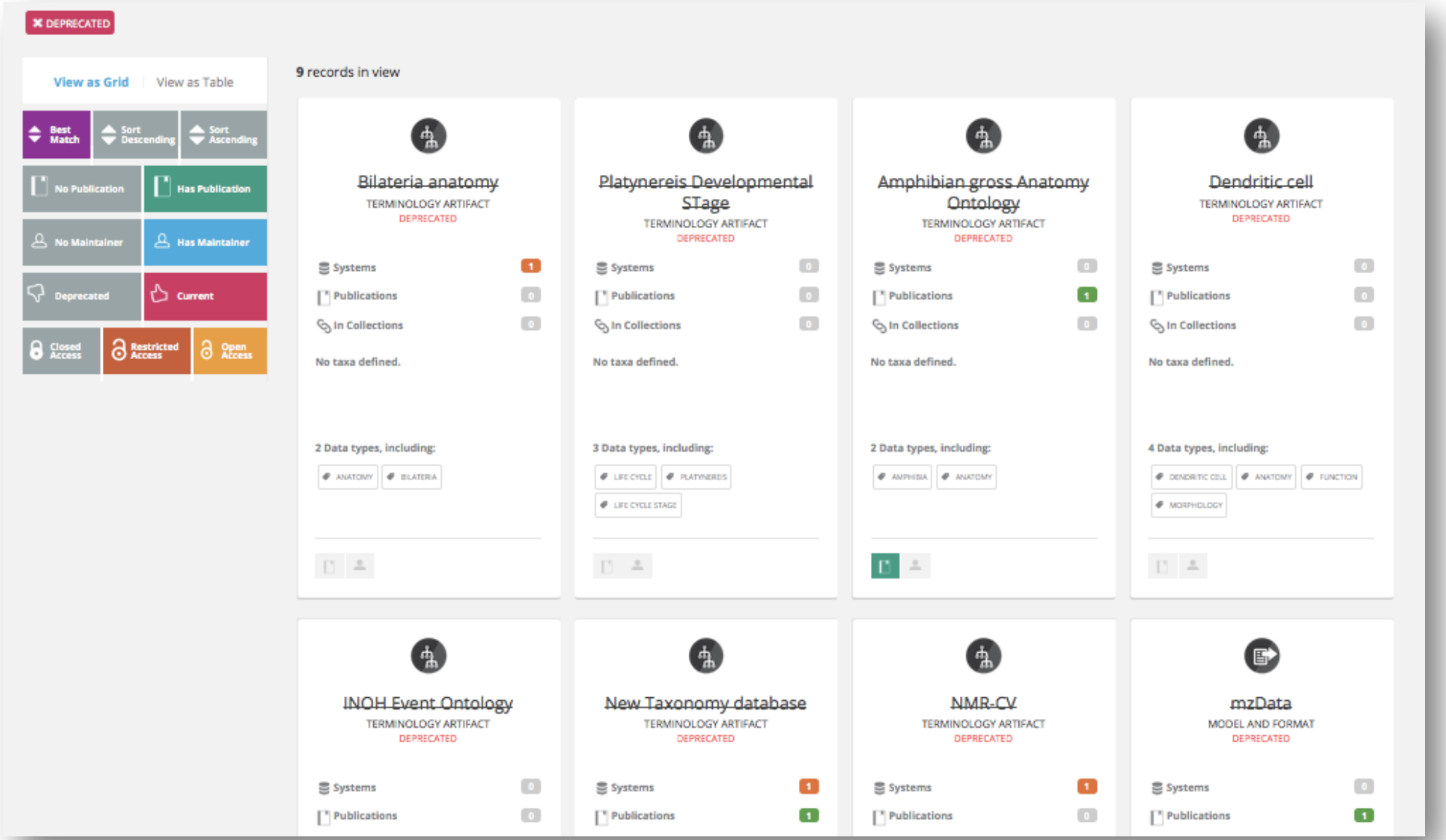
Standards and databases cross-linked



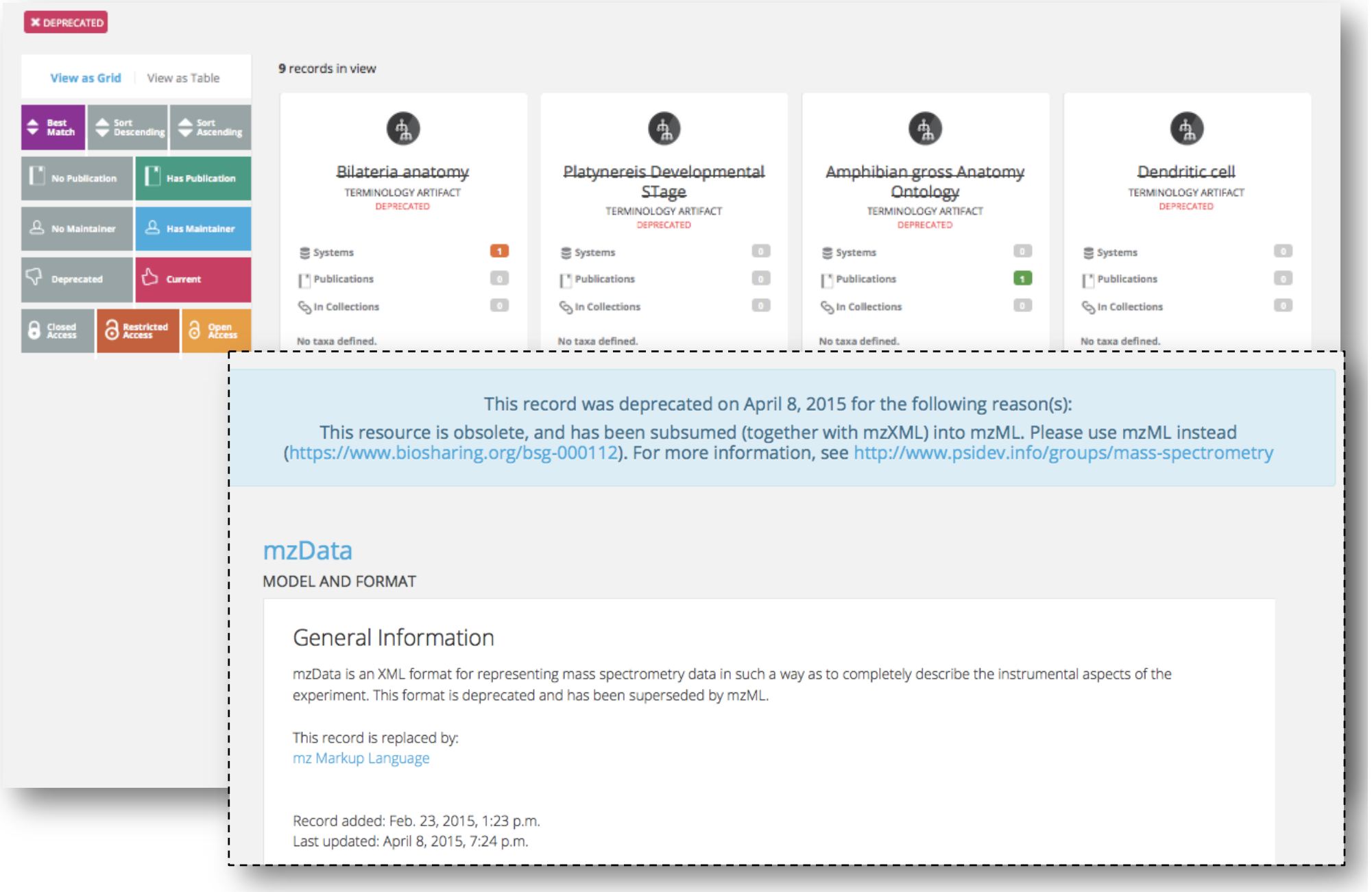
Terminologies linked to NCBO BioPortal



Tracking evolution, e.g. deprecations and substitutions



Tracking evolution, e.g. deprecations and substitutions



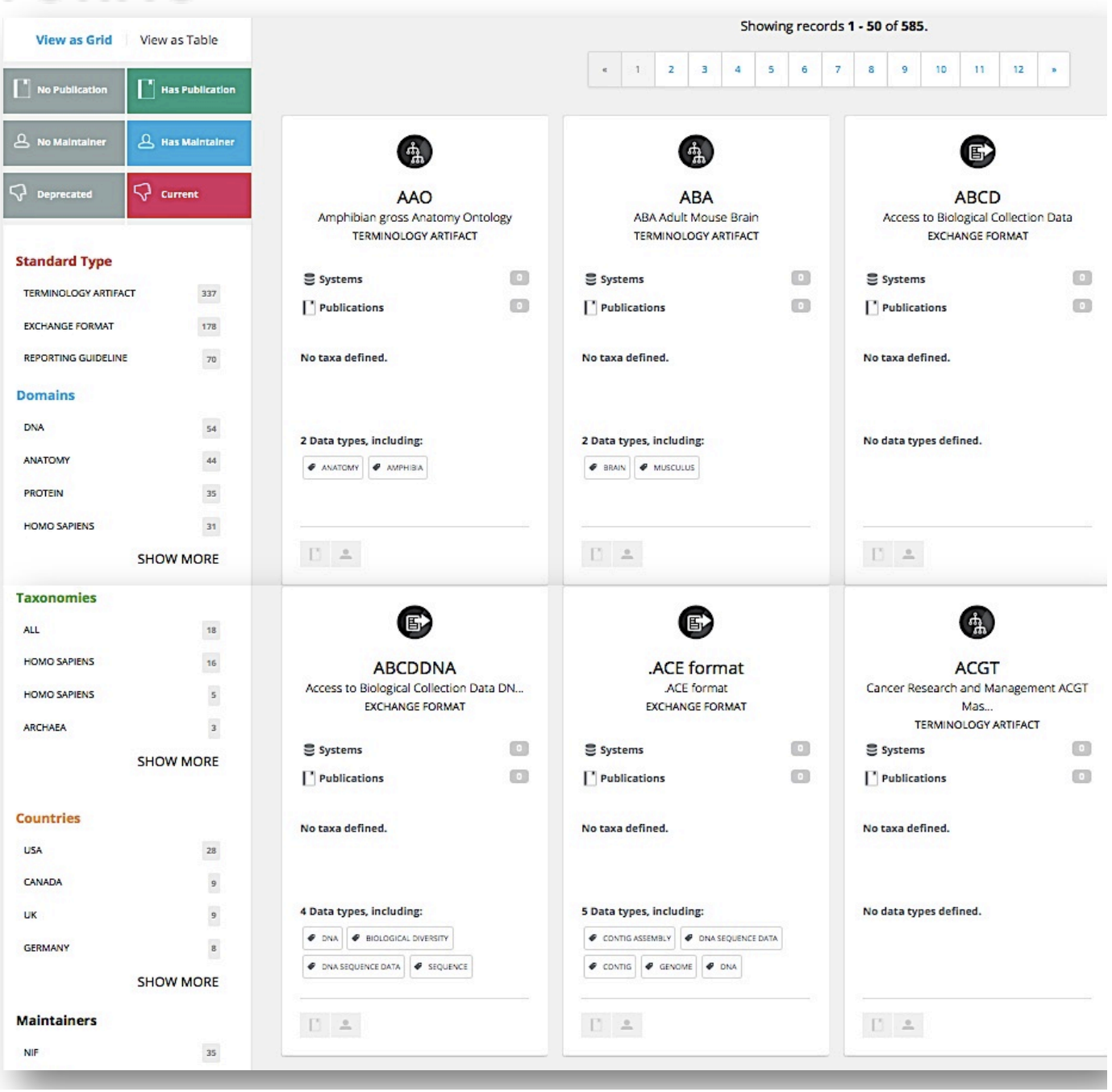
Search filter, refine

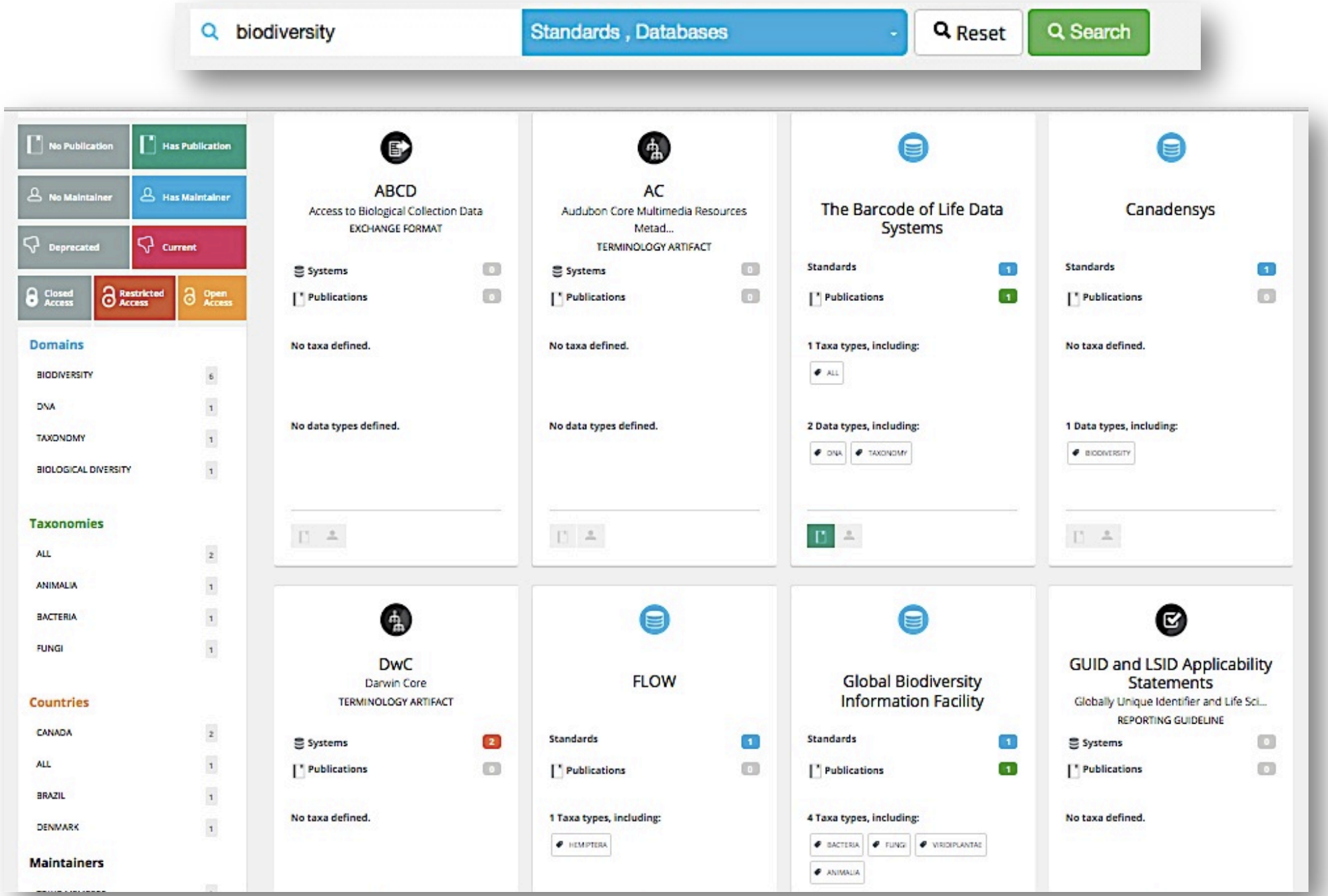
Core functionalities:

- search and filtering, e.g. by funder, domain, type of standard
- Refine by publication, maintainer etc.
- add new records, edit existing records
- “claim” records
- person’s profile (as maintainer of records) associated to the ORCID profile (for credit)
- visualization and views of content and linking

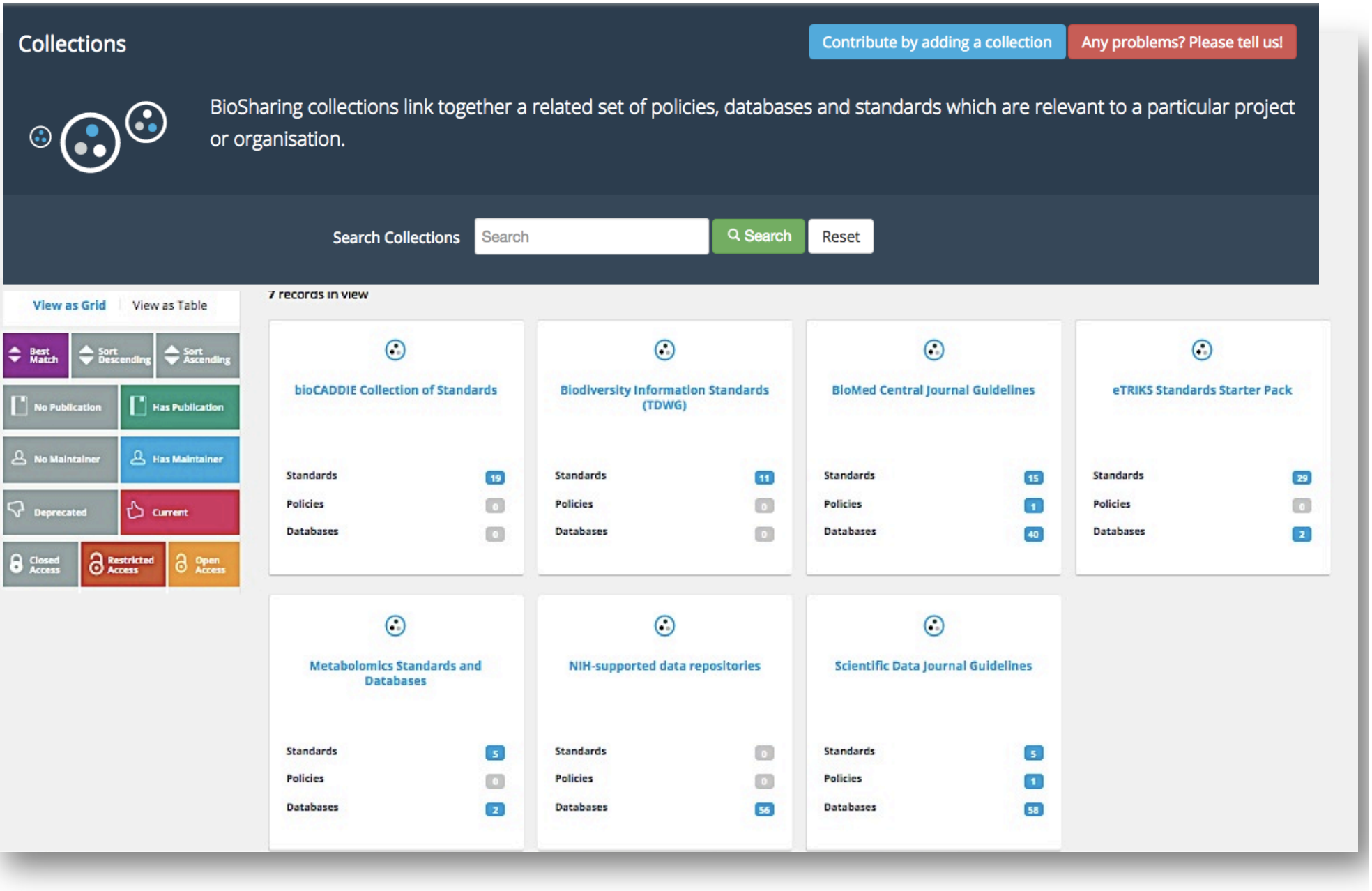
Annotation Sources:

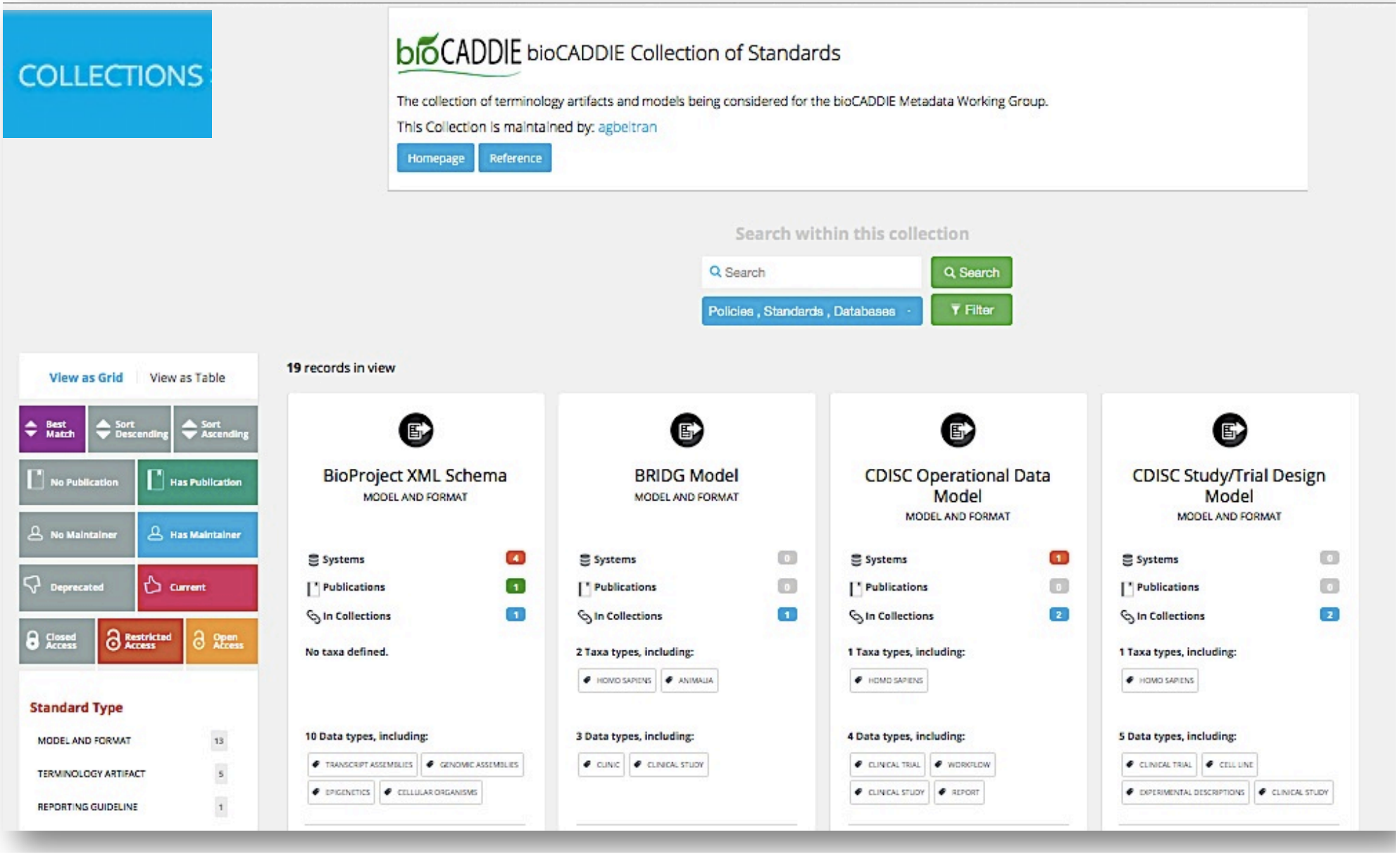
- 4 axis: (material, process, quality, information)
- NIF,OBI,CL,GO,IAO,EDAM

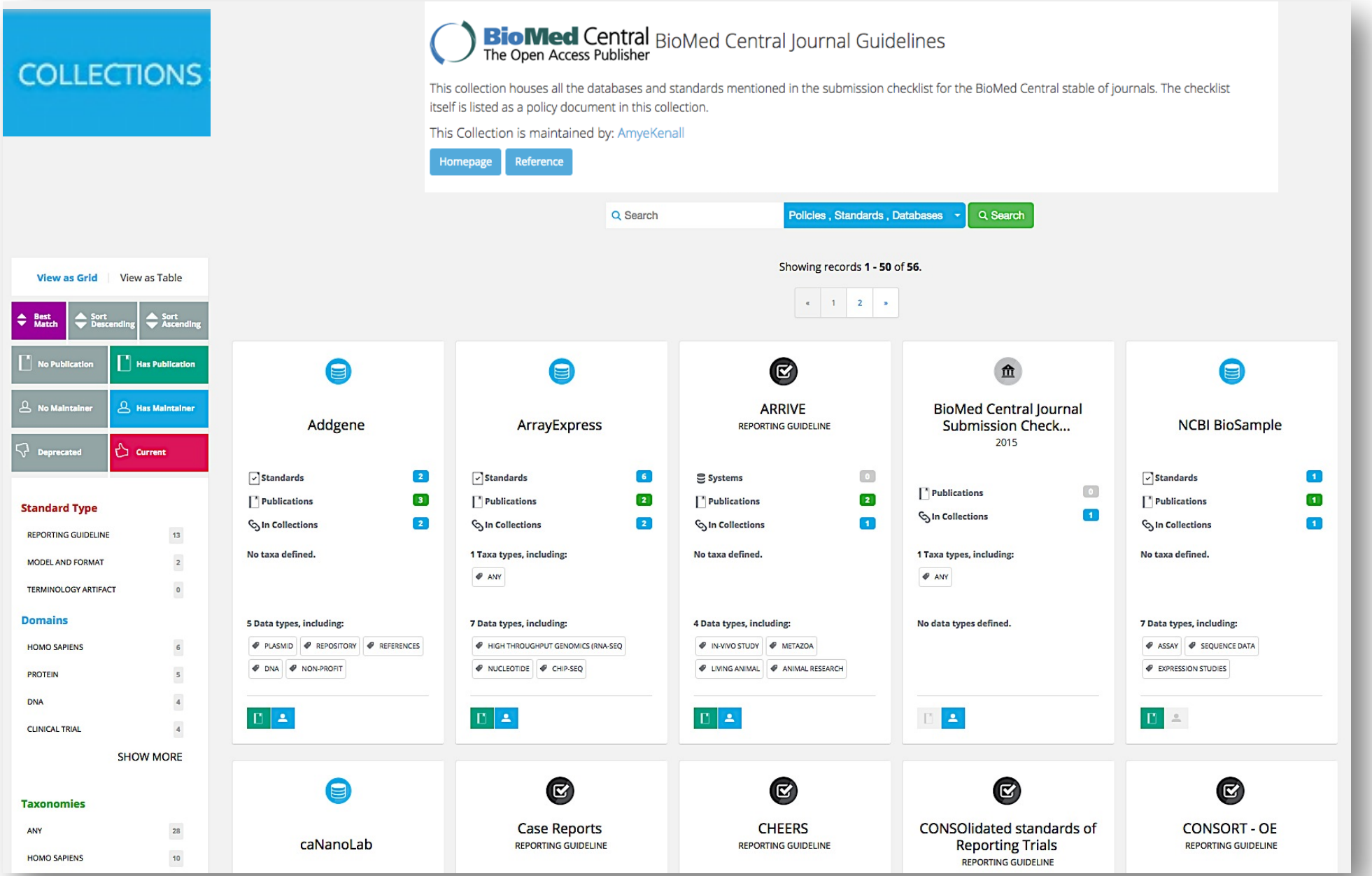



biodiversity
Standards , Databases
Reset
Search
No Publication
Has Publication
No Maintainer
Has Maintainer
Deprecated
Current
Closed Access
Restricted Access
Open Access
Domains
BIODIVERSITY 6
DNA 1
TAXONOMY 1
BIOLOGICAL DIVERSITY 1
Taxonomies
ALL 2
ANIMALIA 1
BACTERIA 1
FUNGI 1
Countries
CANADA 2
ALL 1
BRAZIL 1
DENMARK 1
Maintainers
ABCD
Access to Biological Collection Data
EXCHANGE FORMAT
Systems 0
Publications 0
No taxa defined.
No data types defined.
AC
Audubon Core Multimedia Resources Metad...
TERMINOLOGY ARTIFACT
Systems 0
Publications 0
No taxa defined.
No data types defined.
The Barcode of Life Data Systems
Standards 1
Publications 1
1 Taxa types, including:
ALL
2 Data types, including:
DNA
TAXONOMY
Canadensys
Standards 1
Publications 0
No taxa defined.
1 Data types, including:
BIODIVERSITY
DwC
Darwin Core
TERMINOLOGY ARTIFACT
Systems 2
Publications 0
No taxa defined.
FLOW
Standards 1
Publications 0
1 Taxa types, including:
HEMIPTERA
Global Biodiversity Information Facility
Standards 1
Publications 1
4 Taxa types, including:
BACTERIA
FUNGI
VIRIDIPLANTAE
ANIMALIA
GUID and LSID Applicability Statements
Globally Unique Identifier and Life Sci...
REPORTING GUIDELINE
Systems 0
Publications 0
No taxa defined.

Create your own Collection



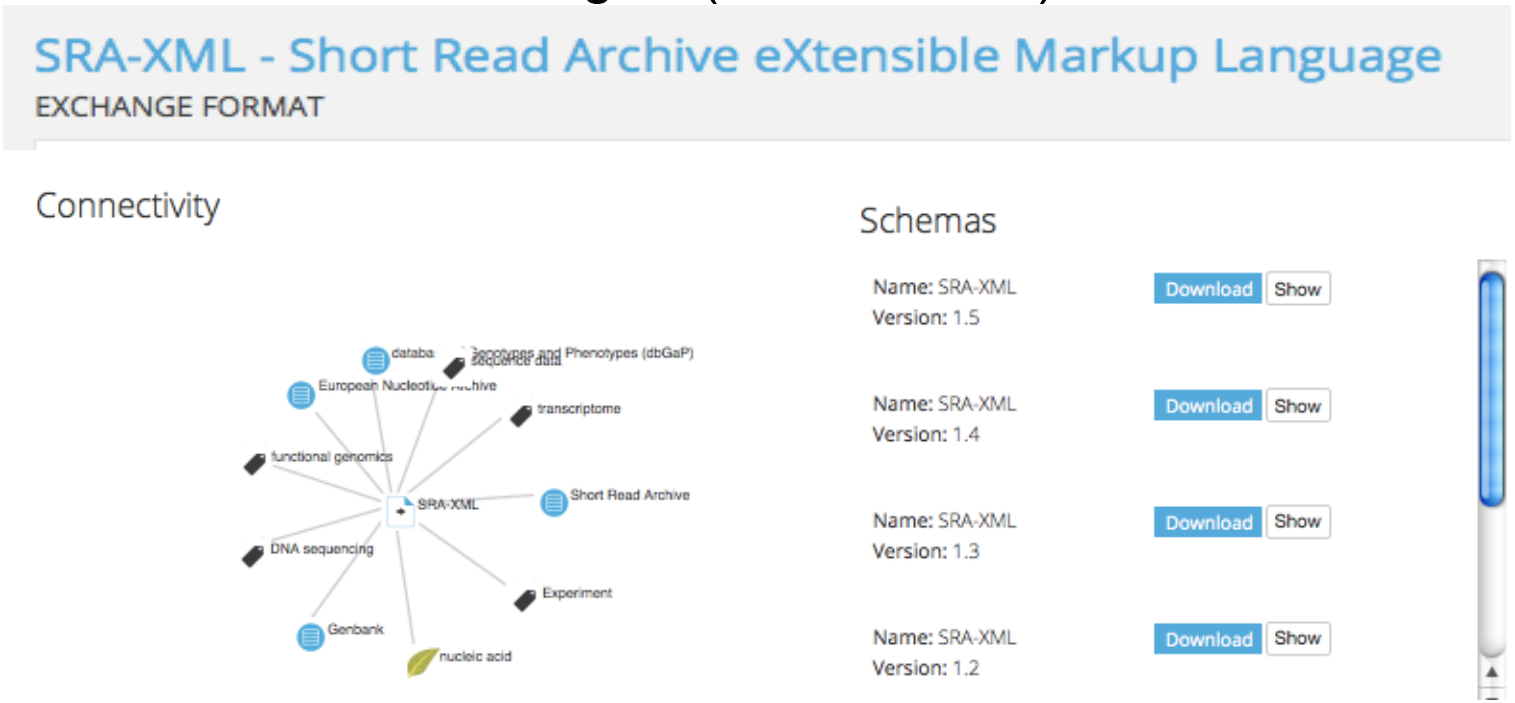

COLLECTIONS
bioCADDIE bioCADDIE Collection of Standards
The collection of terminology artifacts and models being considered for the bioCADDIE Metadata Working Group.
This Collection is maintained by: agbeltran
Homepage
Reference
Search within this collection
Search
Search
Policies , Standards , Databases
Filter
View as Grid
View as Table
Best Match
Sort Descending
Sort Ascending
No Publication
Has Publication
No Maintainer
Has Maintainer
Deprecated
Current
Closed Access
Restricted Access
Open Access
Standard Type
MODEL AND FORMAT 13
TERMINOLOGY ARTIFACT 5
REPORTING GUIDELINE 1
19 records in view
BioProject XML Schema
MODEL AND FORMAT
Systems 4
Publications 1
In Collections 1
No taxa defined.
10 Data types, including:
TRANSCRIPT ASSEMBLIES
GENOMIC ASSEMBLIES
EPIGENETICS
CELLULAR ORGANISMS
BRIDG Model
MODEL AND FORMAT
Systems 0
Publications 0
In Collections 1
2 Taxa types, including:
HOMO SAPIENS
ANIMALIA
3 Data types, including:
CLINIC
CLINICAL STUDY
CDISC Operational Data Model
MODEL AND FORMAT
Systems 1
Publications 0
In Collections 2
1 Taxa types, including:
HOMO SAPIENS
4 Data types, including:
CLINICAL TRIAL
WORKFLOW
CLINICAL STUDY
REPORT
CDISC Study/Trial Design Model
MODEL AND FORMAT
Systems 0
Publications 0
In Collections 2
1 Taxa types, including:
HOMO SAPIENS
5 Data types, including:
CLINICAL TRIAL
CELL LINE
EXPERIMENTAL DESCRIPTIONS
CLINICAL STUDY


COLLECTIONS
BioMed Central The Open Access Publisher
BioMed Central Journal Guidelines
This collection houses all the databases and standards mentioned in the submission checklist for the BioMed Central stable of journals. The checklist itself is listed as a policy document in this collection.
This Collection is maintained by: AmyeKenall
Homepage
Reference
Search
Policies , Standards , Databases
Search
View as Grid
View as Table
Best Match
Sort Descending
Sort Ascending
No Publication
Has Publication
No Maintainer
Has Maintainer
Deprecated
Current
Standard Type
REPORTING GUIDELINE 13
MODEL AND FORMAT 2
TERMINOLOGY ARTIFACT 0
Domains
HOMO SAPIENS 6
PROTEIN 5
DNA 4
CLINICAL TRIAL 4
SHOW MORE
Taxonomies
ANY 28
HOMO SAPIENS 10
Showing records 1 - 50 of 56.
« 1 2 »
Addgene
Standards 2
Publications 3
In Collections 2
No taxa defined.
5 Data types, including:
PLASMID REPOSITORY REFERENCES DNA NON-PROFIT
ArrayExpress
Standards 6
Publications 2
In Collections 2
1 Taxa types, including:
ANY
7 Data types, including:
HIGH THROUGHPUT GENOMICS (RNA-SEQ NUCLEOTIDE CHIP-SEQ
ARRIVE
REPORTING GUIDELINE
Systems 0
Publications 2
In Collections 1
No taxa defined.
4 Data types, including:
IN-VIVO STUDY METAZOA LIVING ANIMAL ANIMAL RESEARCH
BioMed Central Journal Submission Check...
2015
Publications 0
In Collections 1
1 Taxa types, including:
ANY
No data types defined.
NCBI BioSample
Standards 1
Publications 1
In Collections 1
No taxa defined.
7 Data types, including:
ASSAY SEQUENCE DATA EXPRESSION STUDIES
caNanoLab
Case Reports
REPORTING GUIDELINE
CHEERS
REPORTING GUIDELINE
CONSOlidated standards of Reporting Trials
REPORTING GUIDELINE
CONSORT - OE
REPORTING GUIDELINE

WIP – surveys, advisory board & WGs

1. Maximize categorization to improve searching, continue curation, add records, increase maintainers/vetting of records

2. Schema views: progressively visualize formats and reporting guidelines as tested for terminologies (via BioPortal)



3. Export of metadata elements: as part of

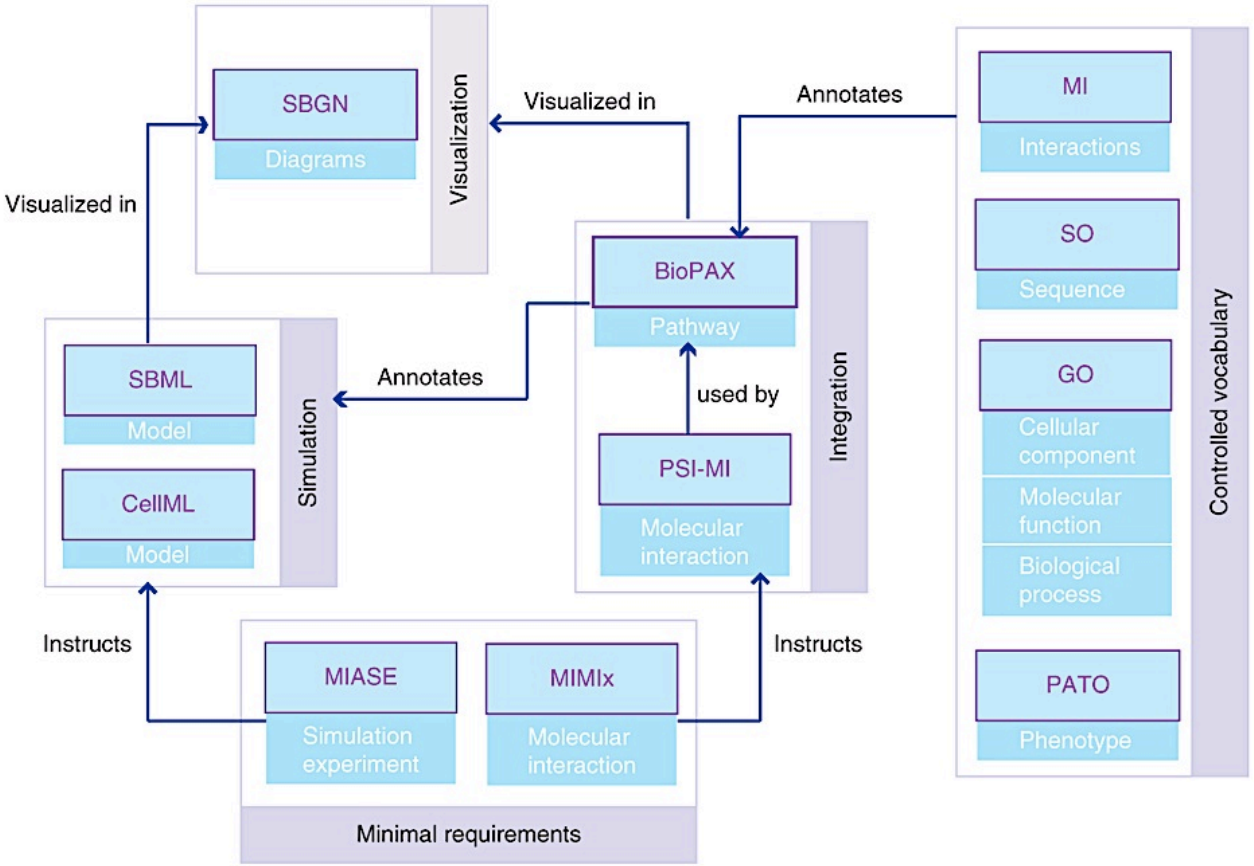
NIH Big Data to Knowledge (BD2K)

CEDAR

elixir UNITED KINGDOM

WIP – surveys, advisory board & WGs

4. Create relation or "usage maps and guides", e.g.:



The relationship among popular standard formats for pathway information.

Demir, *et al*., **The BioPAX community standard for pathway data sharing**, Nat Biotech. 2010.

5. Metrics of maturity, usability and popularity, kick-started by Tenenbaum, Sansone, Haendel. A sea of standards for omics data: sink or swim? ***J Am Med Inform Assoc*** (2014).

6. Embed in the ecosystem of complementary registries

The BioSharing Registry: connecting data policies, standards & databases in life sciences

Green = 0 - 6 months
WG is just starting

Group details

Status: Recognised & Endorsed
Chair(s): Susanna-Assunta Sansone, Rebecca Lawrence, & Simon Hodson
Secretariat Liaison: Herman Stehouwer
TAB Liaison: Peter Wittenberg and Simon Cox
Case Statement: Download

The products are

- ***principles*** for linking information about **databases**, **content standards** and journal / funder **policies** in the **life sciences**
- and a curated ***registry*** to access and cross-search the information, on which a variety of stakeholders can base their decisions

This is a joint RDA FORCE11 WG

The work will benefit from the existing biosharing.org Information Resources also part of elixir UNITED KINGDOM

Collaboration and communication

Key points:

- We all provide information to those seek guidance to use, do or recommend something (databases, tools, standards, training material etc.)
- Designed with a slightly different focus and user needs; focus on curating/cataloguing certain information
- Have different entry points and UIs to fit the specific scope/user needs is good, so federation is key, e.g. cross-referencing
- Pool resources where appropriate

Operational Team – based at the Oxford e-Research Centre

Susanna
Sansone, PhD

Philippe
Rocca-Serra, PhD

David
Johnson, PhD

Allyson
Lister, PhD

Peter
McQuilton, PhD

Alejandra
Gonzalez-Beltran, PhD

Milo
Thurston, PhD

Eamonn
Maguire, DPhil / *Contractor*

Where we are and present

@biosharing

WED - Session 1

- RDA/WDS Publishing Data Workflows

THUR - Session 6

- Joint meeting of IG Metadata, WG Metadata Standards Catalog & WG BioSharing Registry: Connecting data policies, standards & databases in life sciences"

FRI - Sessions 8 and 9

- Joint meeting of IG ELIXIR Bridging Force & WG BioSharing Registry: Life Sciences and Sensitive Data"
- Joint meeting of IG Domain Repositories, IG Agriculture Data Interest Group (IGAD), WG BioSharing Registry, IG RDA/CODATA Materials Data, Infrastructure & Interoperability & WG Wheat Data Interoperability"

Proponents, core WG members and adopters

Jessica Tenenbaum (Duke University, USA; Chair Elect for the American Medical Informatics Association's Genomics WG), *Co-Chair of the Advisory Board*

Melissa Haendel (OHSU, USA; Force11; Monarch Initiative), *Co-Chair of the Advisory Board*

Simon Hodson (CODATA), *Co-Chair of the RDA/Force11 WG*

Rebecca Lawrence (F1000), *Co-Chair of the RDA/Force11 WG*

Michael Ball (BBSRC)

Theo Bloom (BMJ)

Jennifer Boyd (OUP)

Scott Edmunds (GigaScience, BGI)

Emma Ganley (PLOS)

Jeff Grethe (USCD; NIF, NIDDK Information Network and NIH BD2K bioCADDIE)

Rafael Jimenez (ELIXIR)

Amye Kenall (BMC)

Varsha Khodiyar(NPG)

Thomas Lemberger (EMBO Press)

Jennifer Lin (CrossRef)

Jonathan Tedds (Ubiquity Press Open Health Data journal; University of Leicester, UK)

Todd Vision (UNC Chapel Hill & NESCent, USA; Dryad)

Michael Witt (re3data; Purdue University Libraries)

Joe Shell (Elsevier/Mendeley)

Linda Naughton (JISC)