ODIN “Big Bang” event, CERN, Thursday, 17 October 2013

www.slideshare.net/SusannaSansone

Data standards, sharing and publication in the life sciences

Susanna-Assunta Sansone, PhD

Associate Director,
Principal Investigator

OXFORD e-Research CENTRE

UNIVERSITY OF OXFORD

Board of Directors

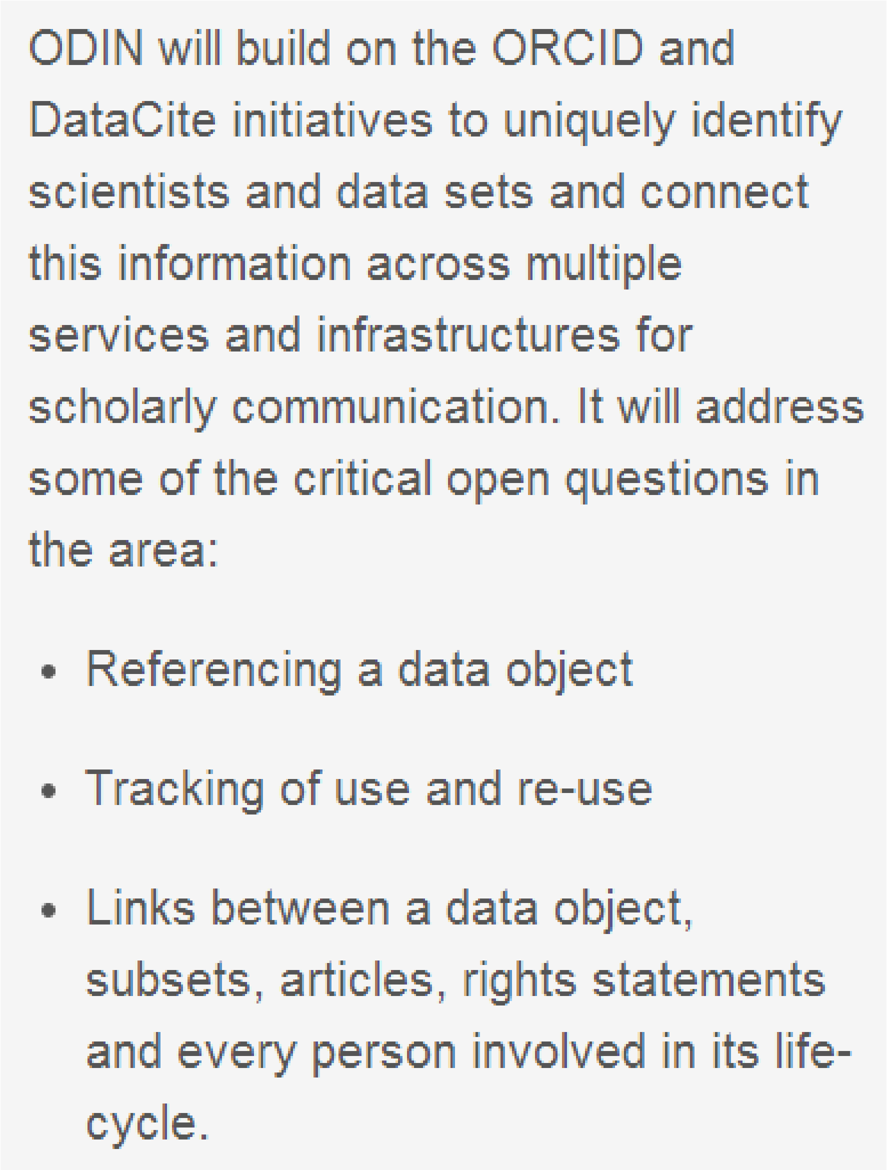
DRYAD

Data Consultant,
Honorary Academic Editor

npg nature publishing group

SCIENTIFIC DATA

ODIN mission



ODIN will build on the ORCID and DataCite initiatives to uniquely identify scientists and data sets and connect this information across multiple services and infrastructures for scholarly communication. It will address some of the critical open questions in the area:

- Referencing a data object
- Tracking of use and re-use
- Links between a data object, subsets, articles, rights statements and every person involved in its life-cycle.

Outline of my talk

***Problem**:*
Identification of datasets in pivotal. But meaningful sharing and (re)use also depend on how well described the datasets are.

Status quo:
In the life sciences there is a wealth of ‘reporting standards’ set to enhance and facilitate the experimental descriptions.

Challenges:
Identify ‘reporting standards’ and their organizations, track their use, usability and impact (e.g. linking them to datasets), credit their developers, users (e.g. curators)...

My team’s activities and groups we work with

data management, biocuration and publication,
collaborative development of software, database, standards and ontology

- environmental genomics
- metabolomics
- metagenomics
- nanotechnology
- proteomics

SEVENTH FRAMEWORK PROGRAMME

elixir UNITED KINGDOM

bbsrc biotechnology and biological sciences research council

NATURAL ENVIRONMENT RESEARCH COUNCIL

- stem cell discovery
- system biology
- transcriptomics
- toxicogenomics
- environmental health

env

agro

tox/pharma

health

http://www.flickr.com/photos/notbrucelee/8016189356/

CC BY

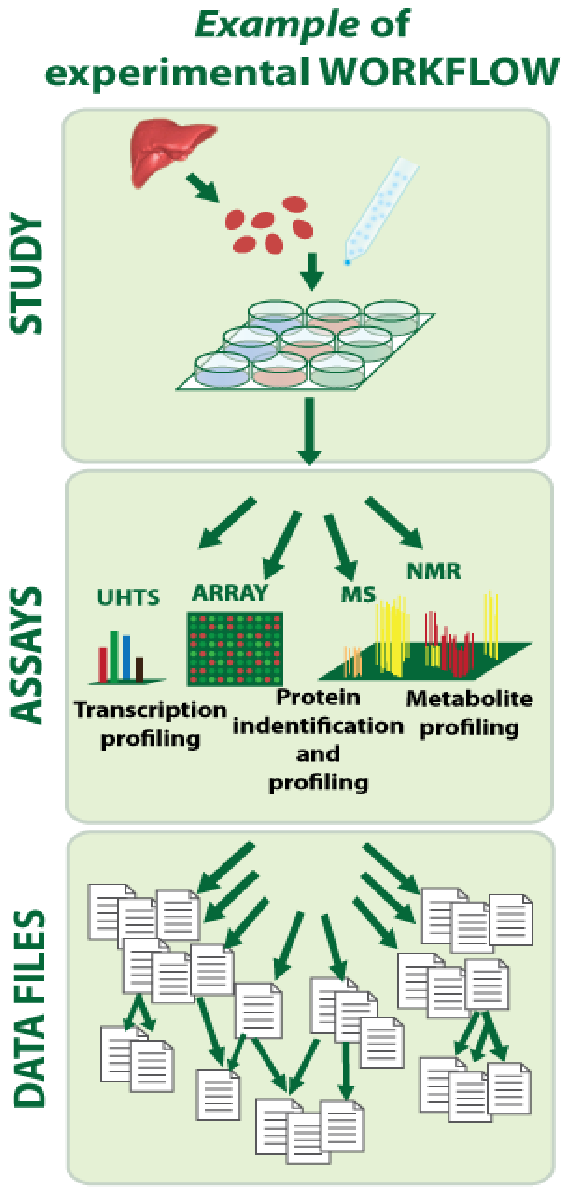
COMPREHENSIBLE

INTEROPERABLE

REUSABLE

http://www.flickr.com/photos/notbrucelee/8016189356/

Growing movement for reproducible research



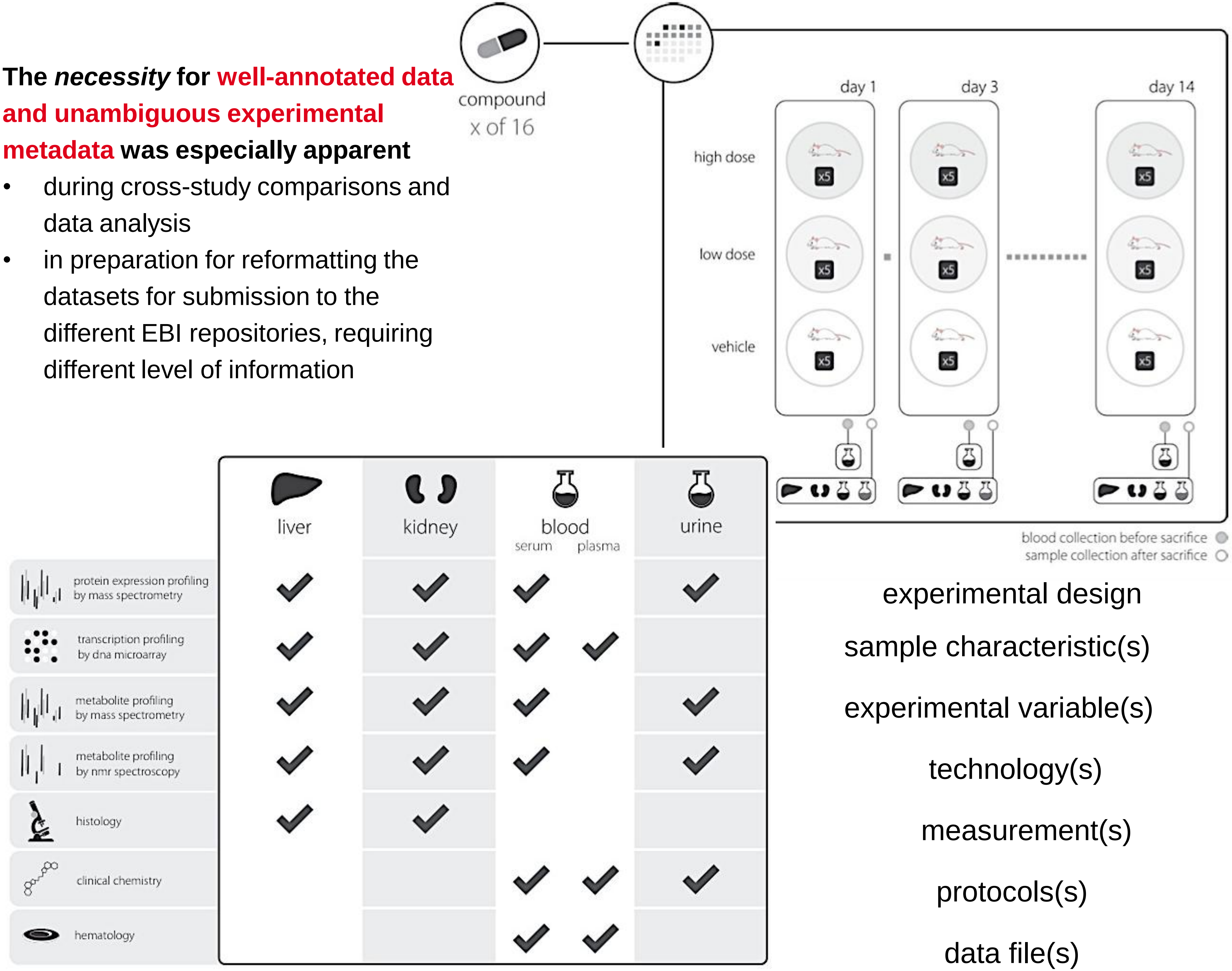
- **Researchers** and **bioinformaticians** in both *academic* and *commercial* arenas, along with **funding agencies** and **publishers**, embrace the concept that to be *comprehensible*, *interoperable* and *reusable* shared datasets we should have **richly described**:
  - entities of interest

    e.g., genes, metabolites, phenotypes, computational models, diseases ...
  - experimental steps

    e.g., provenance of study materials, technology and measurement types, experimentalists and curators ...

The *necessity* for well-annotated data and unambiguous experimental metadata was especially apparent

- during cross-study comparisons and data analysis
- in preparation for reformatting the datasets for submission to the different EBI repositories, requiring different level of information


compound
x of 16
day 1
day 3
day 14
high dose
low dose
vehicle
x5
blood collection before sacrifice ●
sample collection after sacrifice ○

	liver	kidney	blood serum	blood plasma	urine
protein expression profiling by mass spectrometry	✓	✓	✓		✓
transcription profiling by dna microarray	✓	✓	✓	✓	
metabolite profiling by mass spectrometry	✓	✓	✓		✓
metabolite profiling by nmr spectroscopy	✓	✓	✓		✓
histology	✓	✓			
clinical chemistry			✓	✓	✓
hematology			✓	✓	

experimental design

sample characteristic(s)

experimental variable(s)

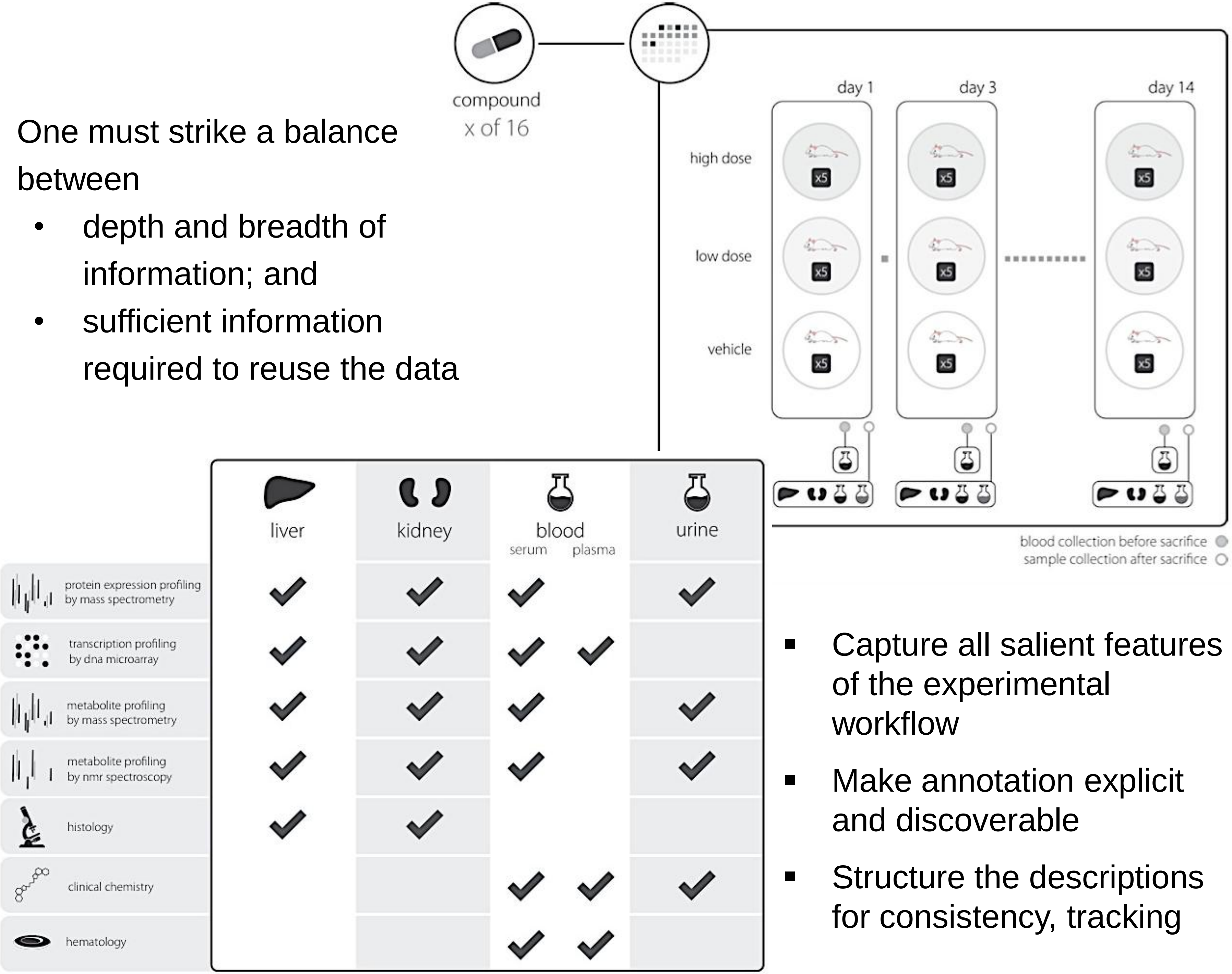
technology(s)

measurement(s)

protocols(s)

data file(s)

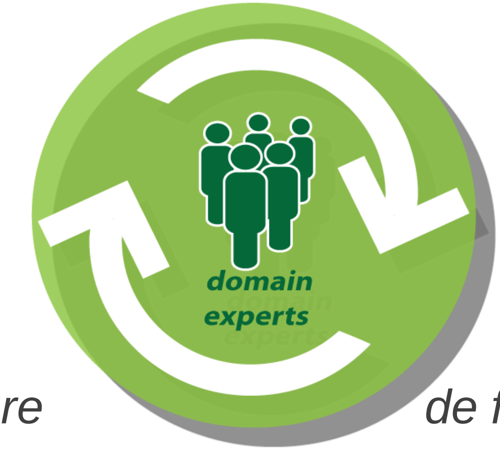
- One must strike a balance between
  - depth and breadth of information; and
  - sufficient information required to reuse the data


compound
x of 16
day 1
day 3
day 14
high dose
low dose
vehicle
x5
blood collection before sacrifice
sample collection after sacrifice

	liver	kidney	blood serum	blood plasma	urine
protein expression profiling by mass spectrometry	✓	✓	✓		✓
transcription profiling by dna microarray	✓	✓	✓	✓	
metabolite profiling by mass spectrometry	✓	✓	✓		✓
metabolite profiling by nmr spectroscopy	✓	✓	✓		✓
histology	✓	✓			
clinical chemistry			✓	✓	✓
hematology			✓	✓	

- Capture all salient features of the experimental workflow
- Make annotation explicit and discoverable
- Structure the descriptions for consistency, tracking

A community mobilization to develop standards, e.g.:



de jure

de facto

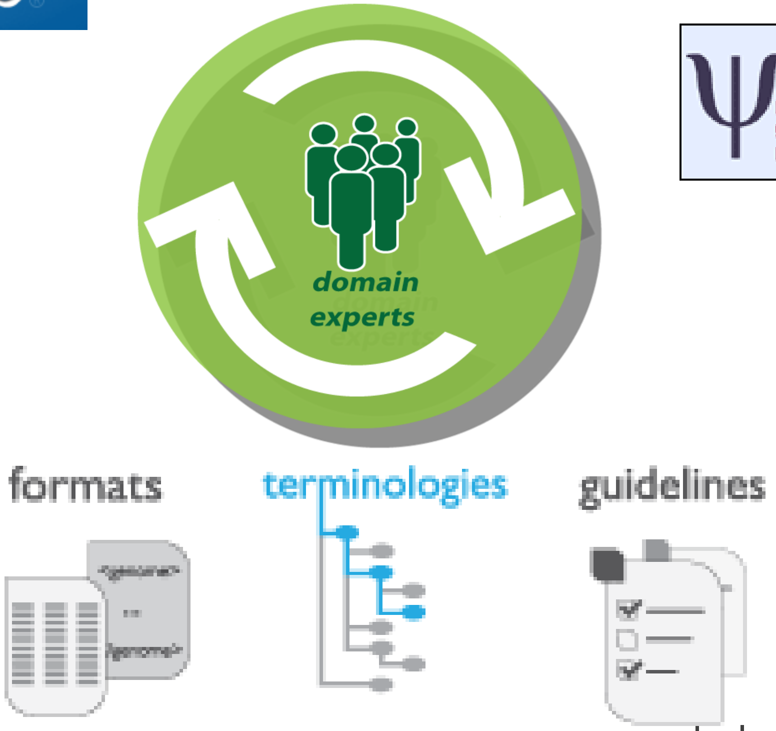
standard organizations

grass-roots groups

Nanotechnology Working Group

- Structural and operational differences
  - organization types (open, close to members, society, WG etc.)
  - standards development (how to formulate, conduct and maintain)
  - adoption, uptake, outreach (link to journals, funders and commercial sector)
  - funds (sponsors, memberships, grants, volunteering)

Types of reporting standards



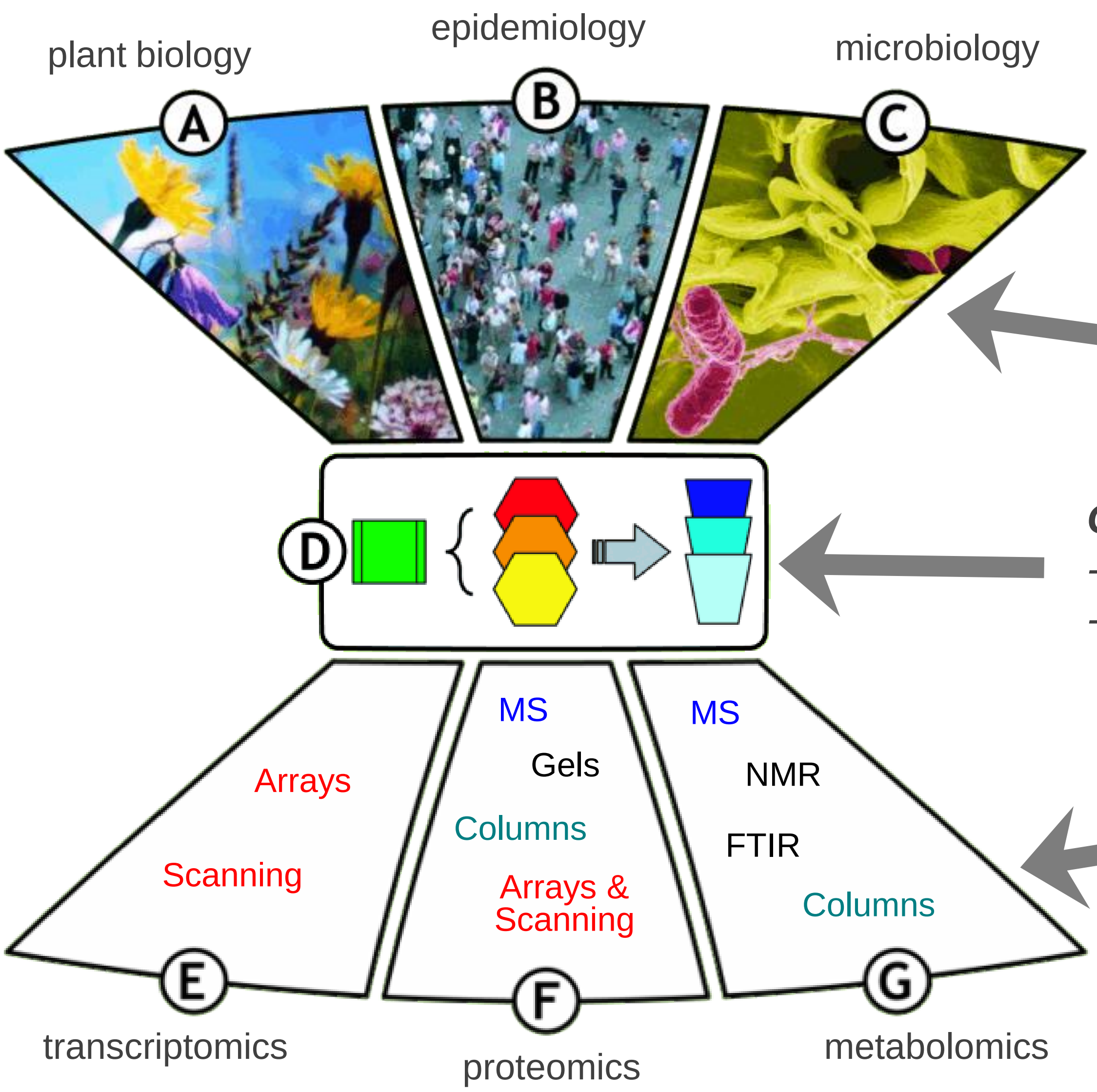
Nanotechnology Working Group

Including ***conceptual model***, ***conceptual schema*** from which an exchange format is derived to allow data to flow from one system to another

Including ***controlled vocabularies***, ***taxonomies***, ***thesauri***, ***ontologies*** etc. to use the same word and refer to the same ‘thing’

Including ***minimum information reporting requirements***, or ***checklists*** to report the same core, essential information

Fragmentation, duplications and gaps



***Biologically*-delineated views of the world**

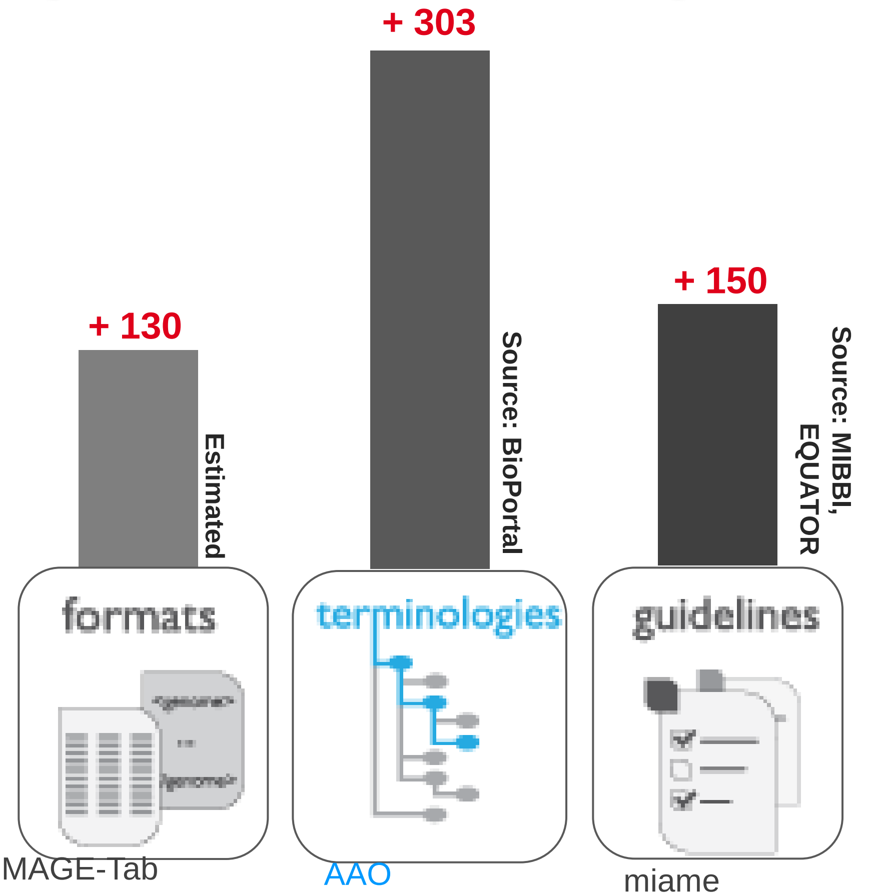
***Generic* features (‘common core’)**
- *description of source biomaterial*
- *experimental design components*

***Technologically*-delineated views of the world**

To compare and integrate data we need interoperable standards

Growing number of reporting standards

To track provenance of the information and ensure richness of data and experimental metadata descriptions, to maximize reusability



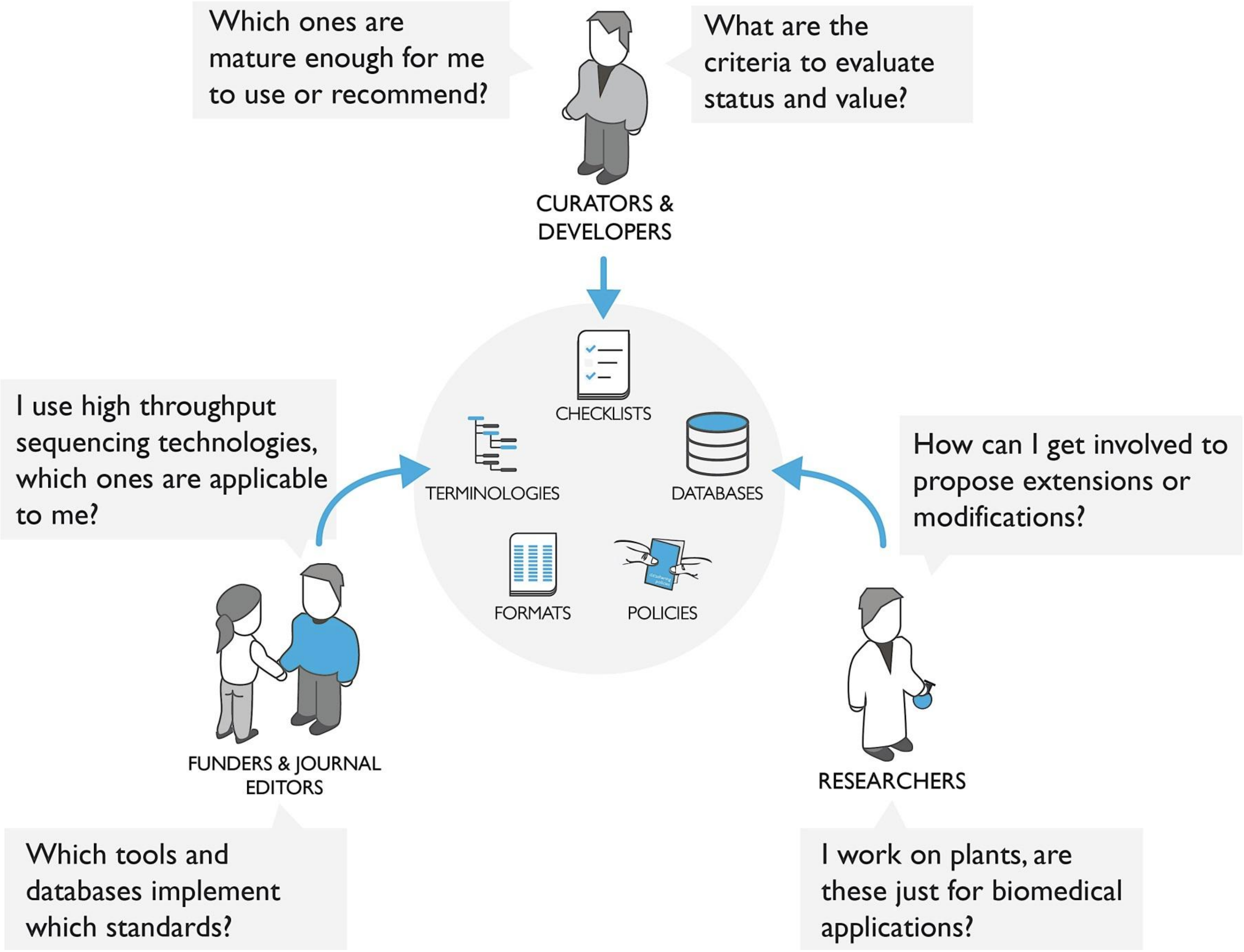
MAGE-Tab
GCDML
SRAxml
SOFT
FASTA
CML
DICOM
GELML
SBRML
MITAB
MzML
ISA-Tab
SEDML…

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

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

Databases, annotation, curation tools

But how much do we know about these standards



Which ones are mature enough for me to use or recommend?

What are the criteria to evaluate status and value?

CURATORS & DEVELOPERS

CHECKLISTS

TERMINOLOGIES

DATABASES

FORMATS

POLICIES

I use high throughput sequencing technologies, which ones are applicable to me?

How can I get involved to propose extensions or modifications?

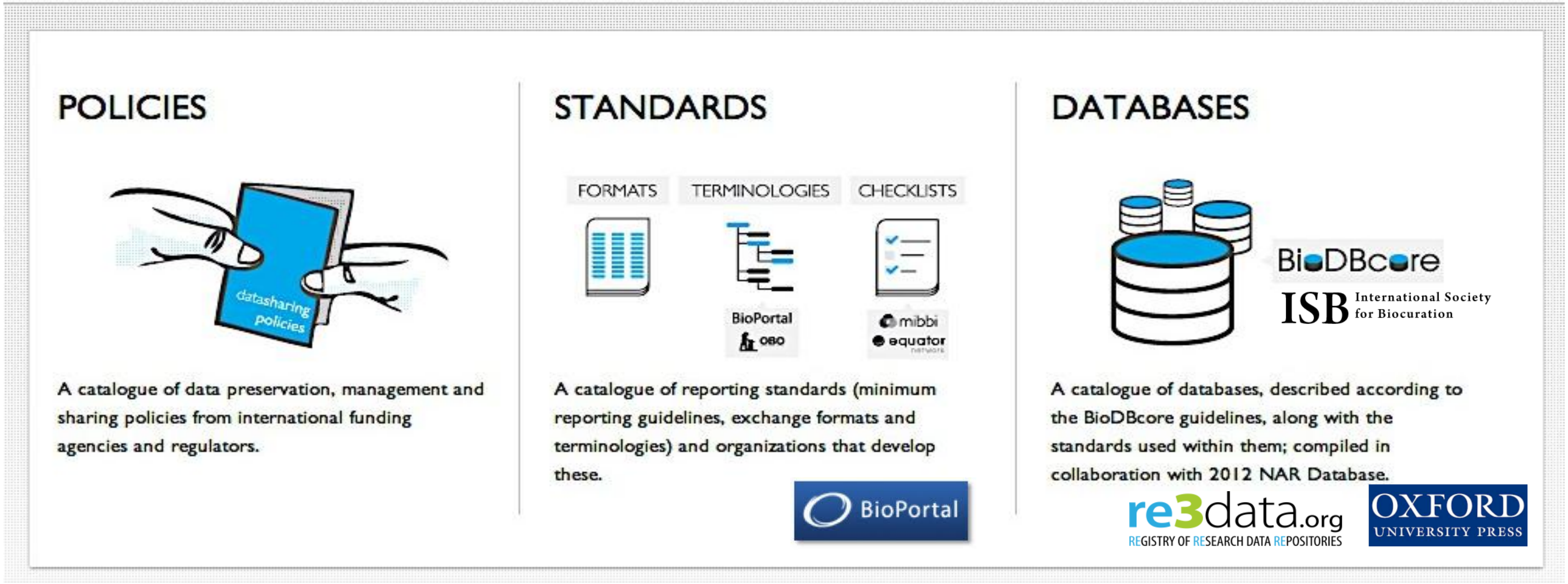
FUNDERS & JOURNAL EDITORS

RESEARCHERS

Which tools and databases implement which standards?

I work on plants, are these just for biomedical applications?

biosharing.org



POLICIES

A catalogue of data preservation, management and sharing policies from international funding agencies and regulators.

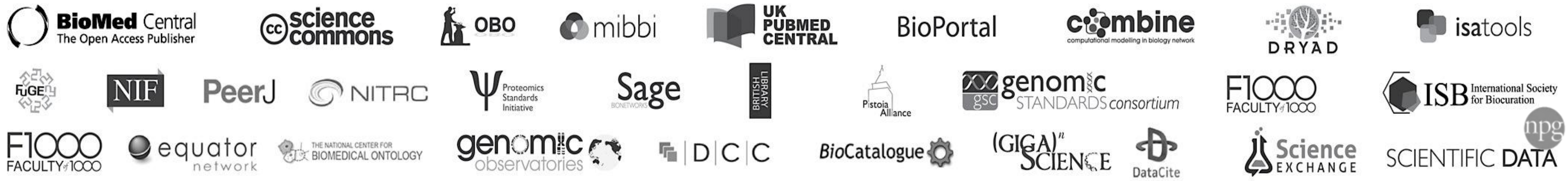
STANDARDS

A catalogue of reporting standards (minimum reporting guidelines, exchange formats and terminologies) and organizations that develop these.

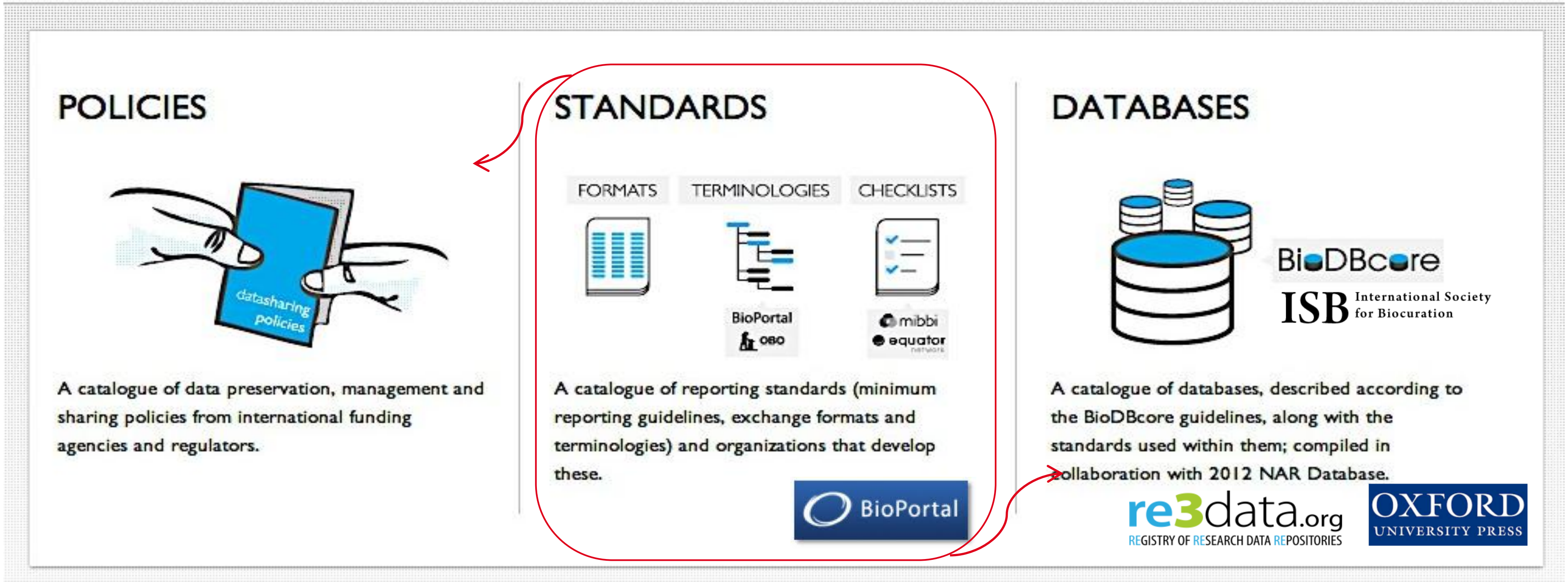
DATABASES

A catalogue of databases, described according to the BioDBcore guidelines, along with the standards used within them; compiled in collaboration with 2012 NAR Database.

- A *coherent, curated* and *searchable registry* of **standards** for describing and reporting experiments in life science, environmental, biomedical and biotechnological domains



biosharing.org



POLICIES

A catalogue of data preservation, management and sharing policies from international funding agencies and regulators.

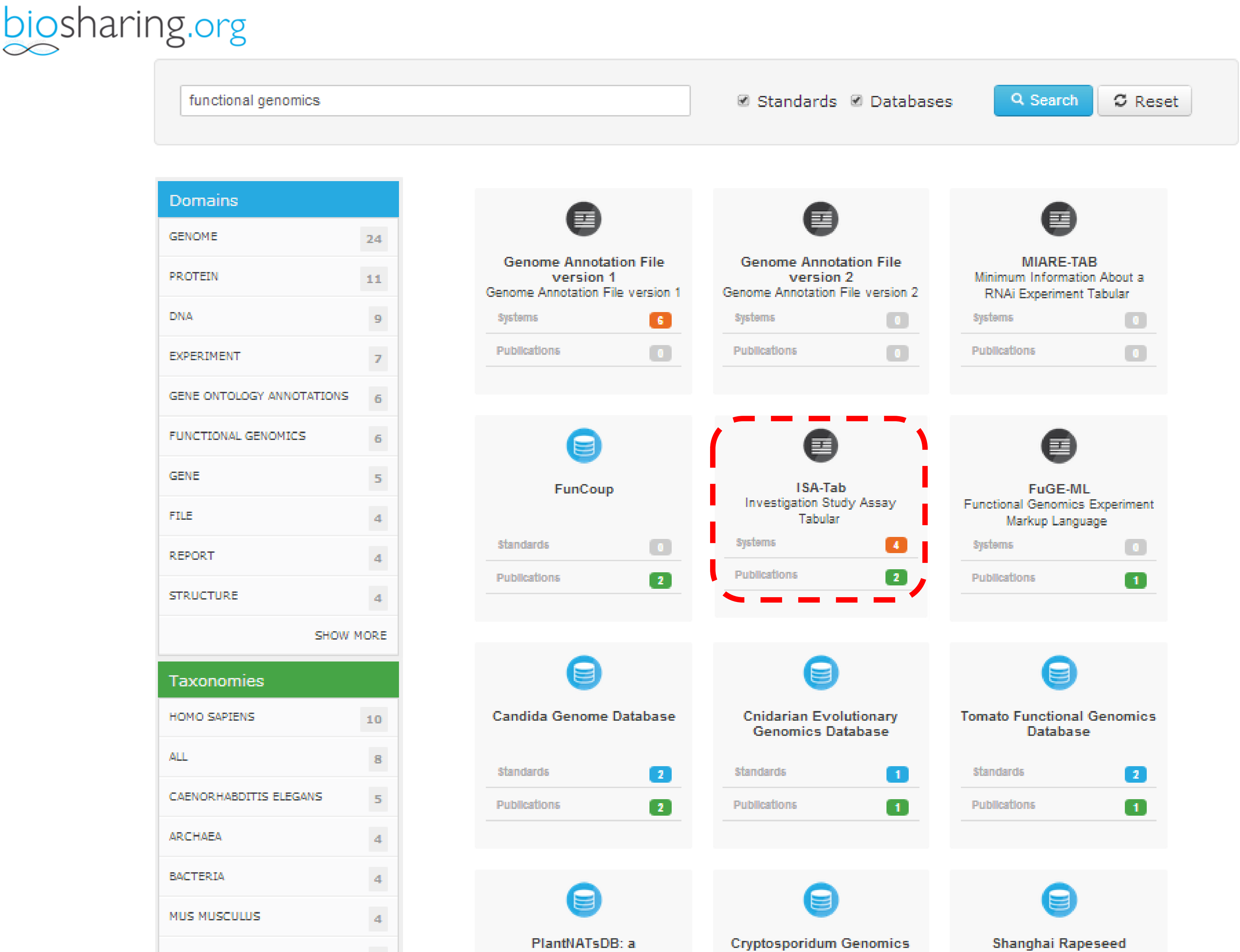
STANDARDS

A catalogue of reporting standards (minimum reporting guidelines, exchange formats and terminologies) and organizations that develop these.

DATABASES

A catalogue of databases, described according to the BioDBcore guidelines, along with the standards used within them; compiled in collaboration with 2012 NAR Database.

- A *coherent, curated* and *searchable registry* of **standards** for describing and reporting experiments in life science, environmental, biomedical and biotechnological domains
- Progressively ***associate*** standards **to data policies** and **databases**
- ***Develop*** assessment **criteria** for usability and popularity of standards
- Help stakeholders to make informed decisions on e.g. what standards or databases to use or recommend; identify efforts they have funded



Domains

GENOME	24
PROTEIN	11
DNA	9
EXPERIMENT	7
GENE ONTOLOGY ANNOTATIONS	6
FUNCTIONAL GENOMICS	6
GENE	5
FILE	4
REPORT	4
STRUCTURE	4

SHOW MORE

Taxonomies

HOMO SAPIENS	10
ALL	8
CAENORHABDITIS ELEGANS	5
ARCHAEA	4
BACTERIA	4
MUS MUSCULUS	4

Genome Annotation File version 1
Genome Annotation File version 1
Systems 6
Publications 0

Genome Annotation File version 2
Genome Annotation File version 2
Systems 0
Publications 0

MIARE-TAB
Minimum Information About a RNAi Experiment Tabular
Systems 0
Publications 0

FunCoup
Standards 0
Publications 2

ISA-Tab
Investigation Study Assay Tabular
Systems 4
Publications 2

FuGE-ML
Functional Genomics Experiment Markup Language
Systems 0
Publications 1

Candida Genome Database
Standards 2
Publications 2

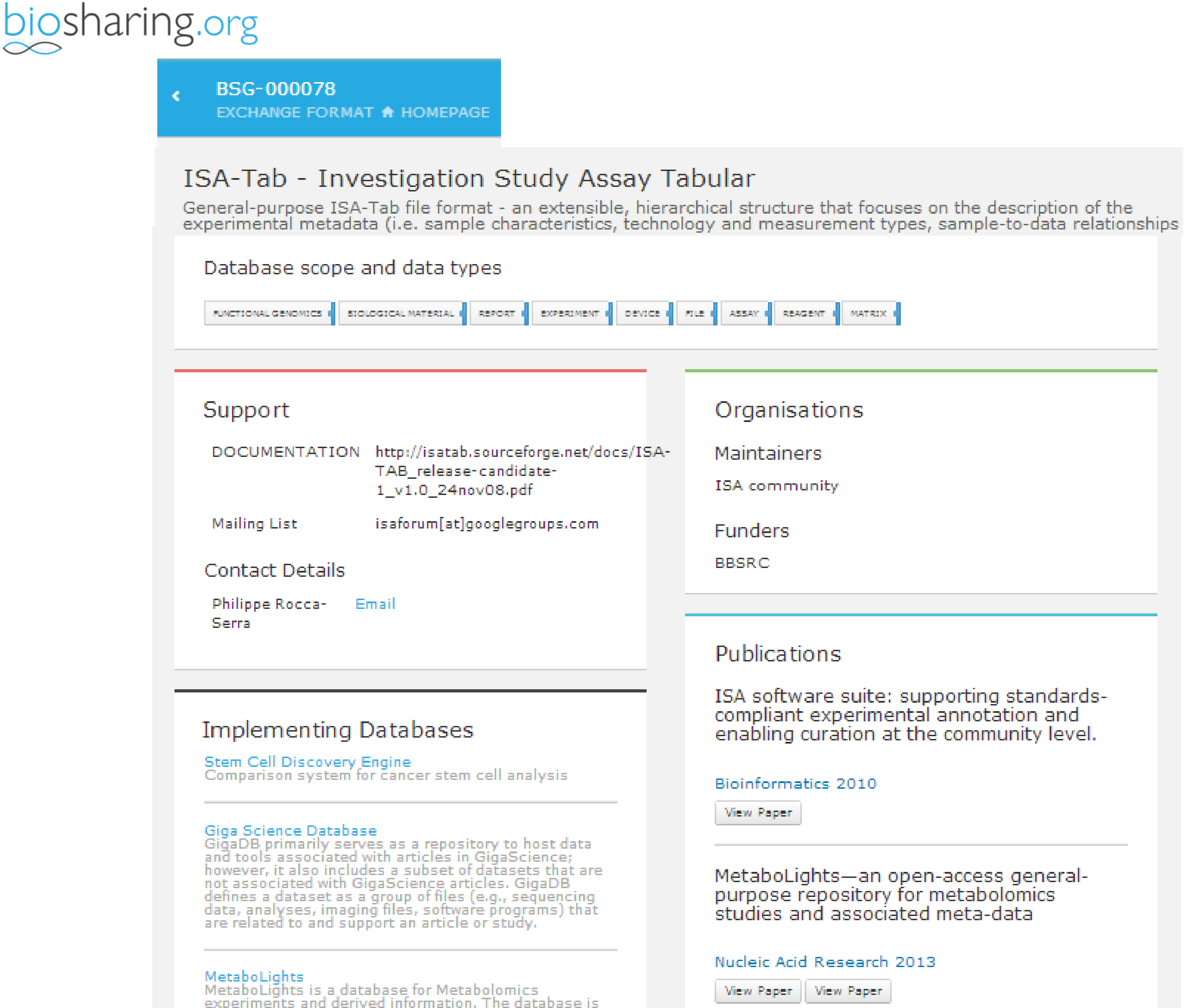
Cnidarian Evolutionary Genomics Database
Standards 1
Publications 1

Tomato Functional Genomics Database
Standards 2
Publications 1

PlantNATsDB: a

Cryptosporidum Genomics

Shanghai Rapeseed



ISA-Tab - Investigation Study Assay Tabular

General-purpose ISA-Tab file format - an extensible, hierarchical structure that focuses on the description of the experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships

Database scope and data types

Support

DOCUMENTATION	http://isatab.sourceforge.net/docs/ISA-TAB_release-candidate-1_v1.0_24nov08.pdf
Mailing List	isaforum[at]googlegroups.com

Contact Details

Philippe Rocca-Serra	Email

Implementing Databases

Stem Cell Discovery Engine
Comparison system for cancer stem cell analysis

Giga Science Database
GigaDB primarily serves as a repository to host data and tools associated with articles in GigaScience; however, it also includes a subset of datasets that are not associated with GigaScience articles. GigaDB defines a dataset as a group of files (e.g., sequencing data, analyses, imaging files, software programs) that are related to and support an article or study.

MetaboLights
MetaboLights is a database for Metabolomics experiments and derived information. The database is

Organisations

Maintainers

ISA community

Funders

BBSRC

Publications

ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level.

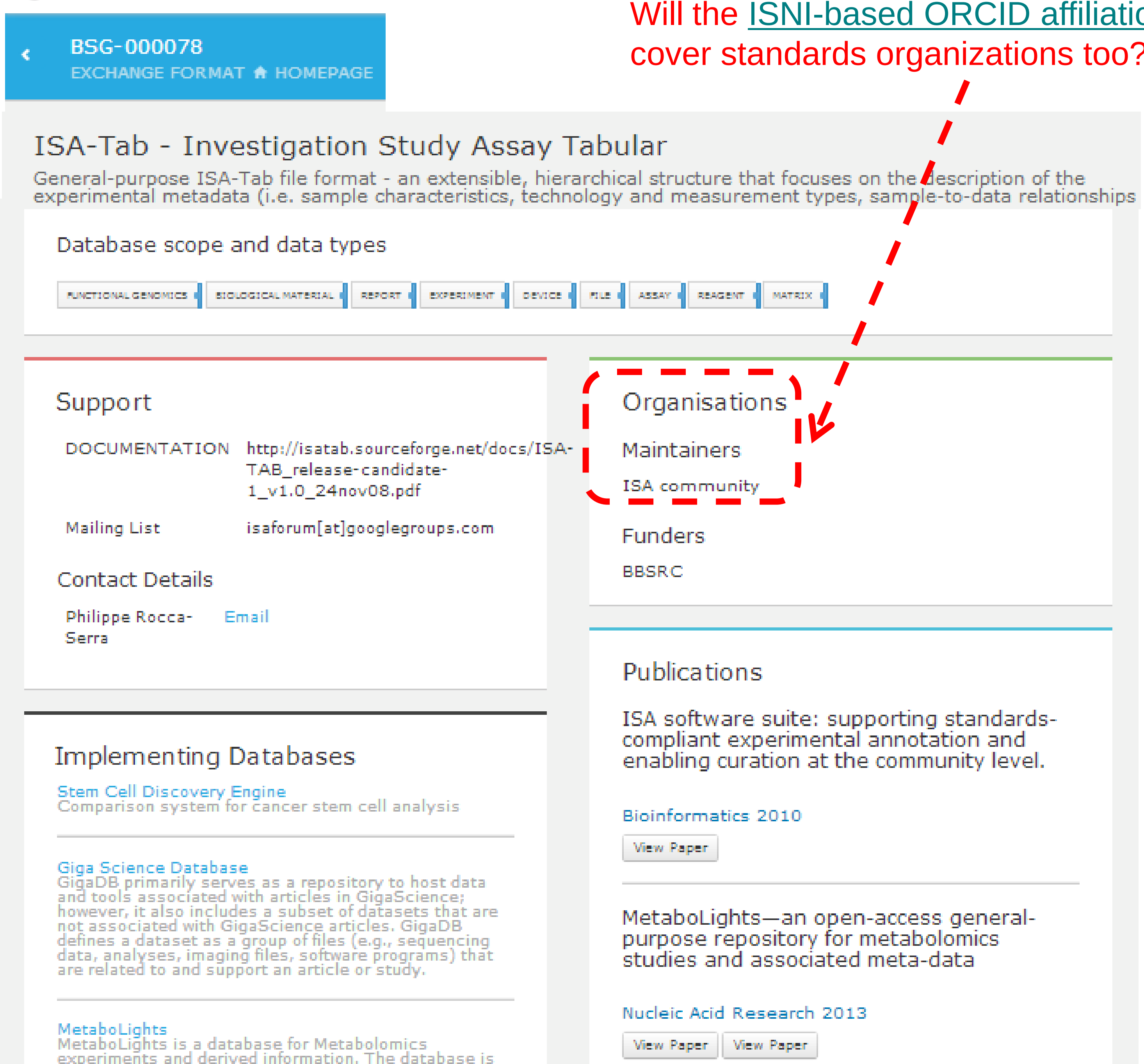
Bioinformatics 2010

MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data

Nucleic Acid Research 2013

biosharing.org

Will the ISNI-based ORCID affiliation module cover standards organizations too?



BSG-000078
EXCHANGE FORMAT HOMEPAGE

ISA-Tab - Investigation Study Assay Tabular

General-purpose ISA-Tab file format - an extensible, hierarchical structure that focuses on the description of the experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships

Database scope and data types

FUNCTIONAL GENOMICS | BIOLOGICAL MATERIAL | REPORT | EXPERIMENT | DEVICE | FILE | ASSAY | REAGENT | MATRIX

Support

DOCUMENTATION	http://isatab.sourceforge.net/docs/ISA-TAB_release-candidate-1_v1.0_24nov08.pdf
Mailing List	isaforum[at]googlegroups.com

Contact Details

Philippe Rocca-Serra	Email

Organisations

Maintainers

ISA community

Funders

BBSRC

Publications

ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level.

Bioinformatics 2010

View Paper

MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data

Nucleic Acid Research 2013

View Paper View Paper

Implementing Databases

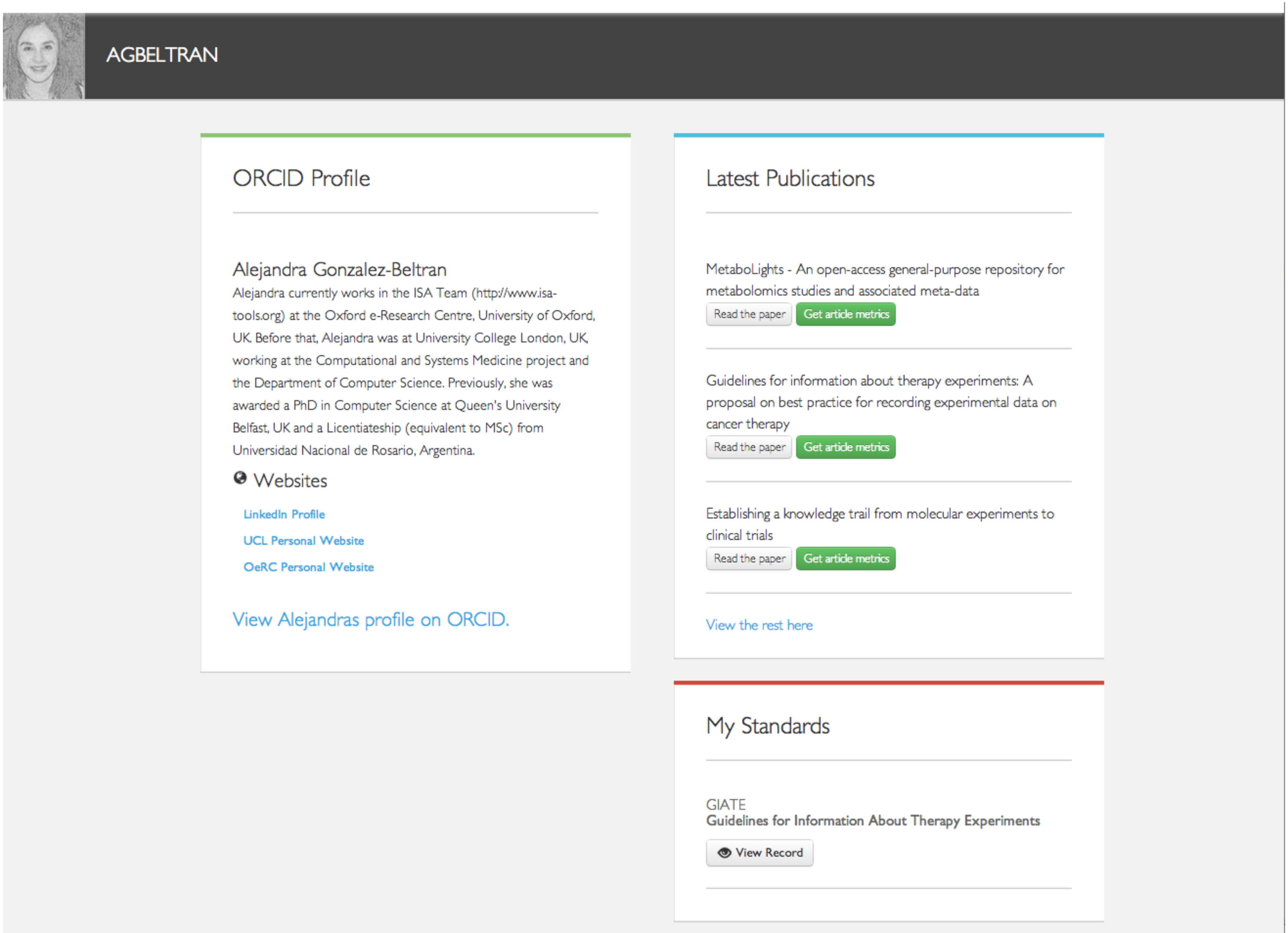
Stem Cell Discovery Engine
Comparison system for cancer stem cell analysis

Giga Science Database
GigaDB primarily serves as a repository to host data and tools associated with articles in GigaScience; however, it also includes a subset of datasets that are not associated with GigaScience articles. GigaDB defines a dataset as a group of files (e.g., sequencing data, analyses, imaging files, software programs) that are related to and support an article or study.

MetaboLights
MetaboLights is a database for Metabolomics experiments and derived information. The database is

biosharing.org

User profiles populated from ORCID...



AGBELTRAN

ORCID Profile

Alejandra Gonzalez-Beltran

Alejandra currently works in the ISA Team (http://www.isa-tools.org) at the Oxford e-Research Centre, University of Oxford, UK. Before that, Alejandra was at University College London, UK, working at the Computational and Systems Medicine project and the Department of Computer Science. Previously, she was awarded a PhD in Computer Science at Queen's University Belfast, UK and a Licentiateship (equivalent to MSc) from Universidad Nacional de Rosario, Argentina.

Websites

- LinkedIn Profile
- UCL Personal Website
- OeRC Personal Website

View Alejandras profile on ORCID.

Latest Publications

MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data

Guidelines for information about therapy experiments: A proposal on best practice for recording experimental data on cancer therapy

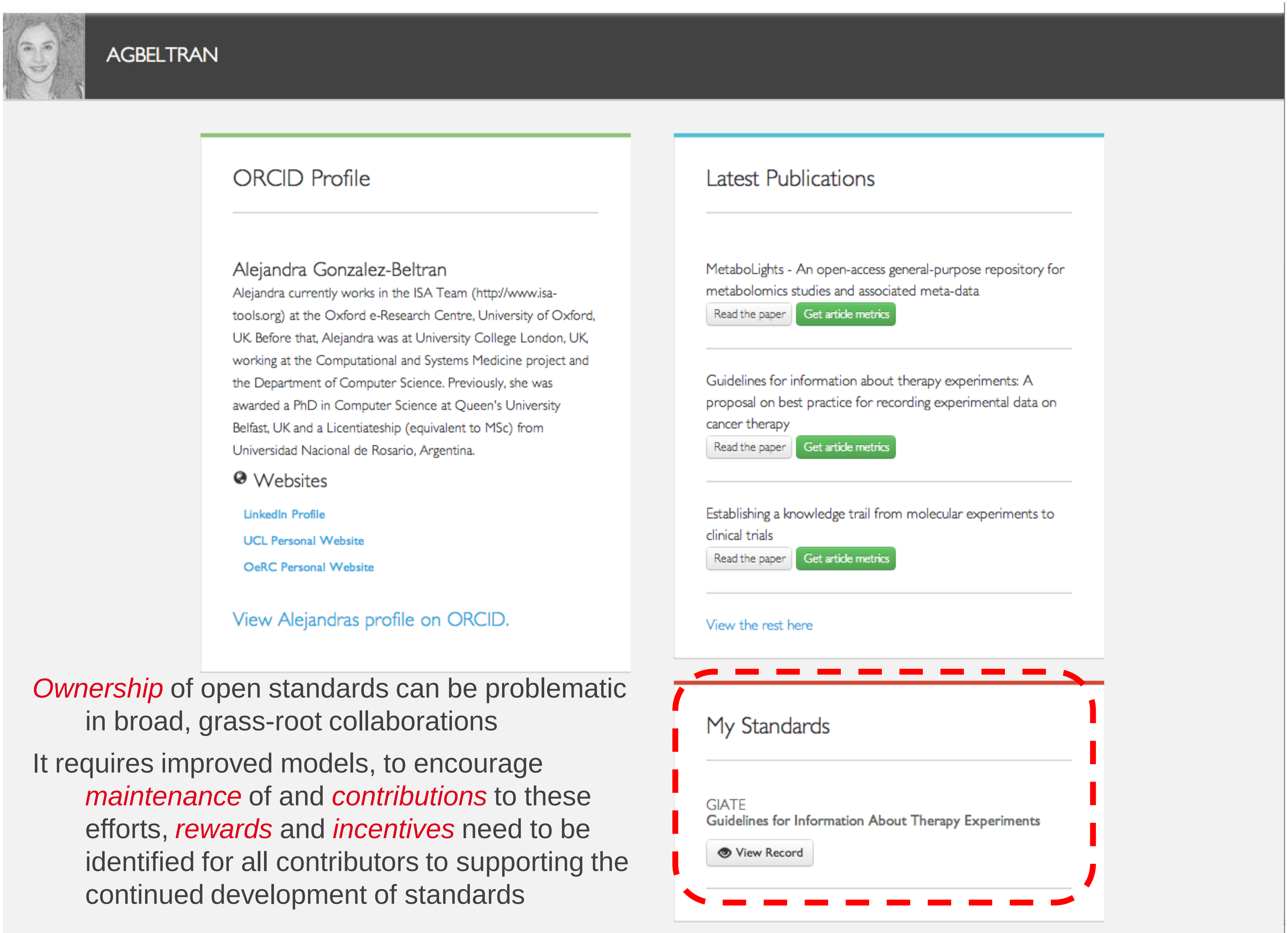
Establishing a knowledge trail from molecular experiments to clinical trials

View the rest here

My Standards

GIATE
Guidelines for Information About Therapy Experiments

... credit for creating, contributing to, maintaining standards



AGBELTRAN

ORCID Profile

Alejandra Gonzalez-Beltran

Alejandra currently works in the ISA Team (http://www.isa-tools.org) at the Oxford e-Research Centre, University of Oxford, UK. Before that, Alejandra was at University College London, UK, working at the Computational and Systems Medicine project and the Department of Computer Science. Previously, she was awarded a PhD in Computer Science at Queen's University Belfast, UK and a Licentiateship (equivalent to MSc) from Universidad Nacional de Rosario, Argentina.

Websites

- LinkedIn Profile
- UCL Personal Website
- OeRC Personal Website

View Alejandras profile on ORCID.

Latest Publications

MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data

Read the paper | Get article metrics

Guidelines for information about therapy experiments: A proposal on best practice for recording experimental data on cancer therapy

Read the paper | Get article metrics

Establishing a knowledge trail from molecular experiments to clinical trials

Read the paper | Get article metrics

View the rest here

My Standards

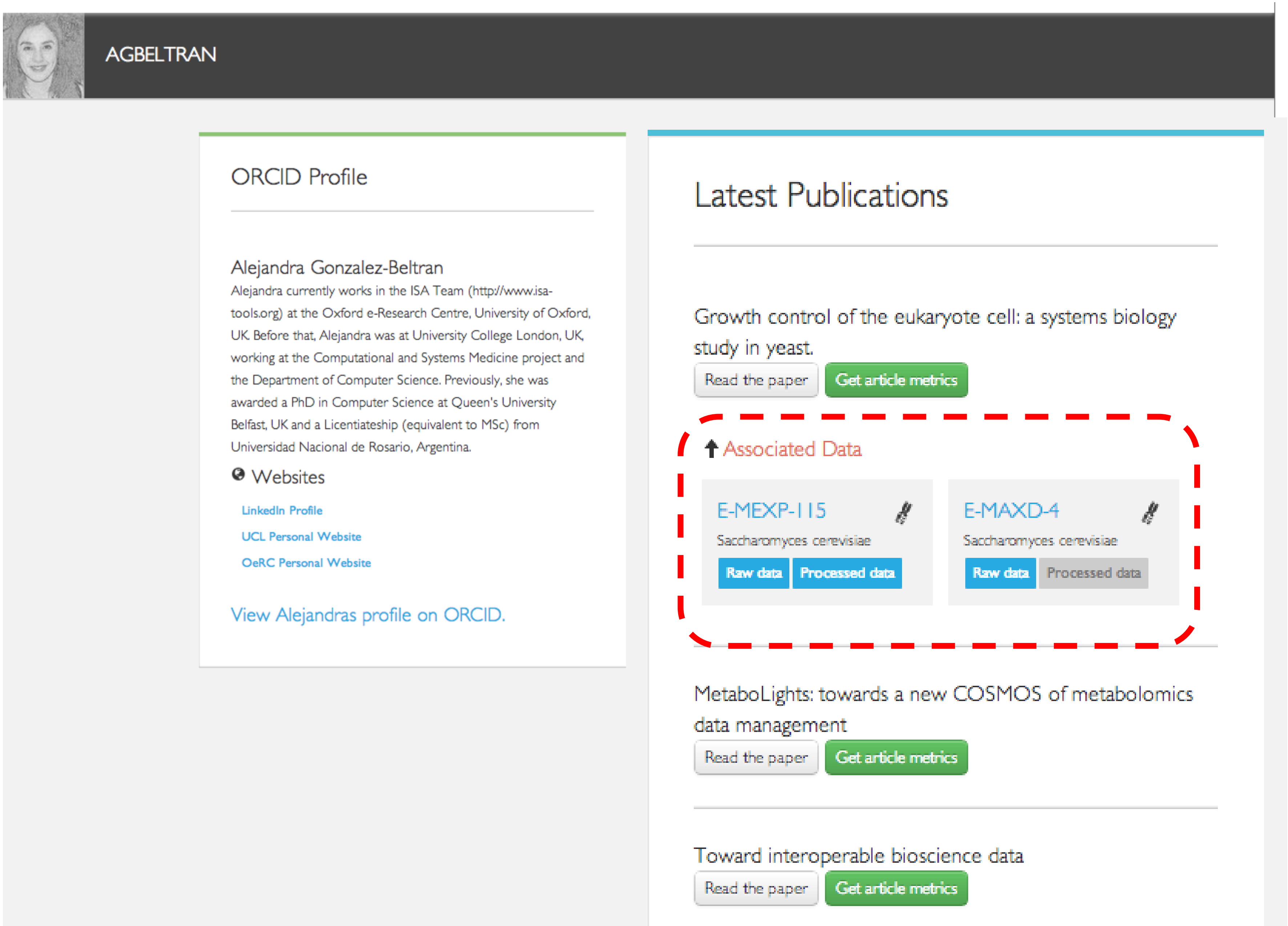
GIATE
Guidelines for Information About Therapy Experiments

View Record

Ownership of open standards can be problematic in broad, grass-root collaborations

It requires improved models, to encourage *maintenance* of and *contributions* to these efforts, *rewards* and *incentives* need to be identified for all contributors to supporting the continued development of standards

... link to data records associated to publications



AGBELTRAN

ORCID Profile

Alejandra Gonzalez-Beltran

Alejandra currently works in the ISA Team (http://www.isa-tools.org) at the Oxford e-Research Centre, University of Oxford, UK. Before that, Alejandra was at University College London, UK, working at the Computational and Systems Medicine project and the Department of Computer Science. Previously, she was awarded a PhD in Computer Science at Queen's University Belfast, UK and a Licentiateship (equivalent to MSc) from Universidad Nacional de Rosario, Argentina.

Websites

- LinkedIn Profile
- UCL Personal Website
- OeRC Personal Website

View Alejandras profile on ORCID.

Latest Publications

Growth control of the eukaryote cell: a systems biology study in yeast.

Associated Data

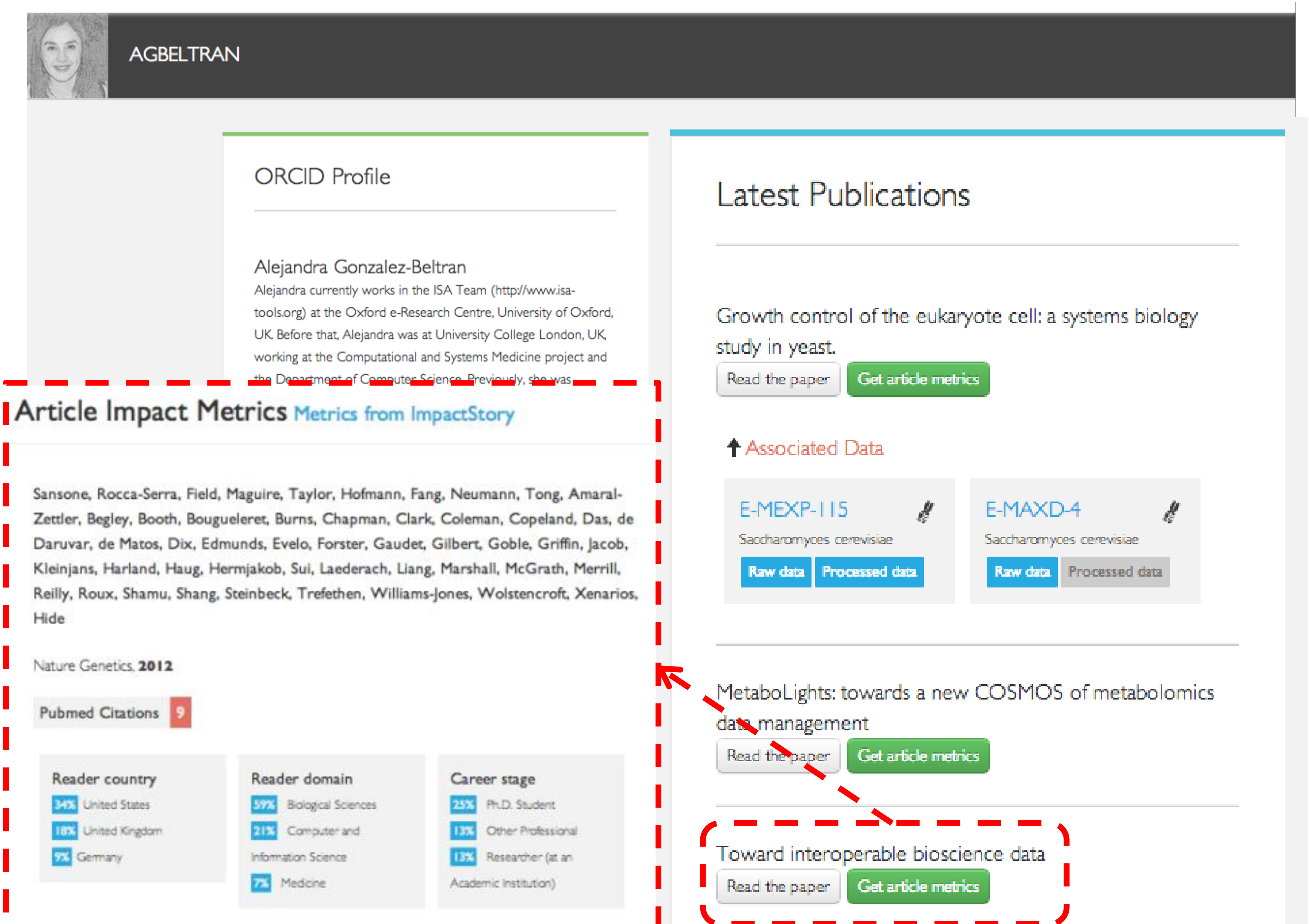
E-MEXP-115 Saccharomyces cerevisiae Raw data Processed data	E-MAXD-4 Saccharomyces cerevisiae Raw data Processed data

MetaboLights: towards a new COSMOS of metabolomics data management

Toward interoperable bioscience data

biosharing.org

...and associated article-level metrics



AGBELTRAN

ORCID Profile

Alejandra Gonzalez-Beltran

Alejandra currently works in the ISA Team (http://www.isa-tools.org) at the Oxford e-Research Centre, University of Oxford, UK. Before that, Alejandra was at University College London, UK, working at the Computational and Systems Medicine project and the Department of Computer Science. Previously, she was

Latest Publications

Growth control of the eukaryote cell: a systems biology study in yeast.

Associated Data

E-MEXP-115	E-MAXD-4
Saccharomyces cerevisiae	Saccharomyces cerevisiae
Raw data · Processed data	Raw data · Processed data

MetaboLights: towards a new COSMOS of metabolomics data management

Toward interoperable bioscience data

Article Impact Metrics Metrics from ImpactStory

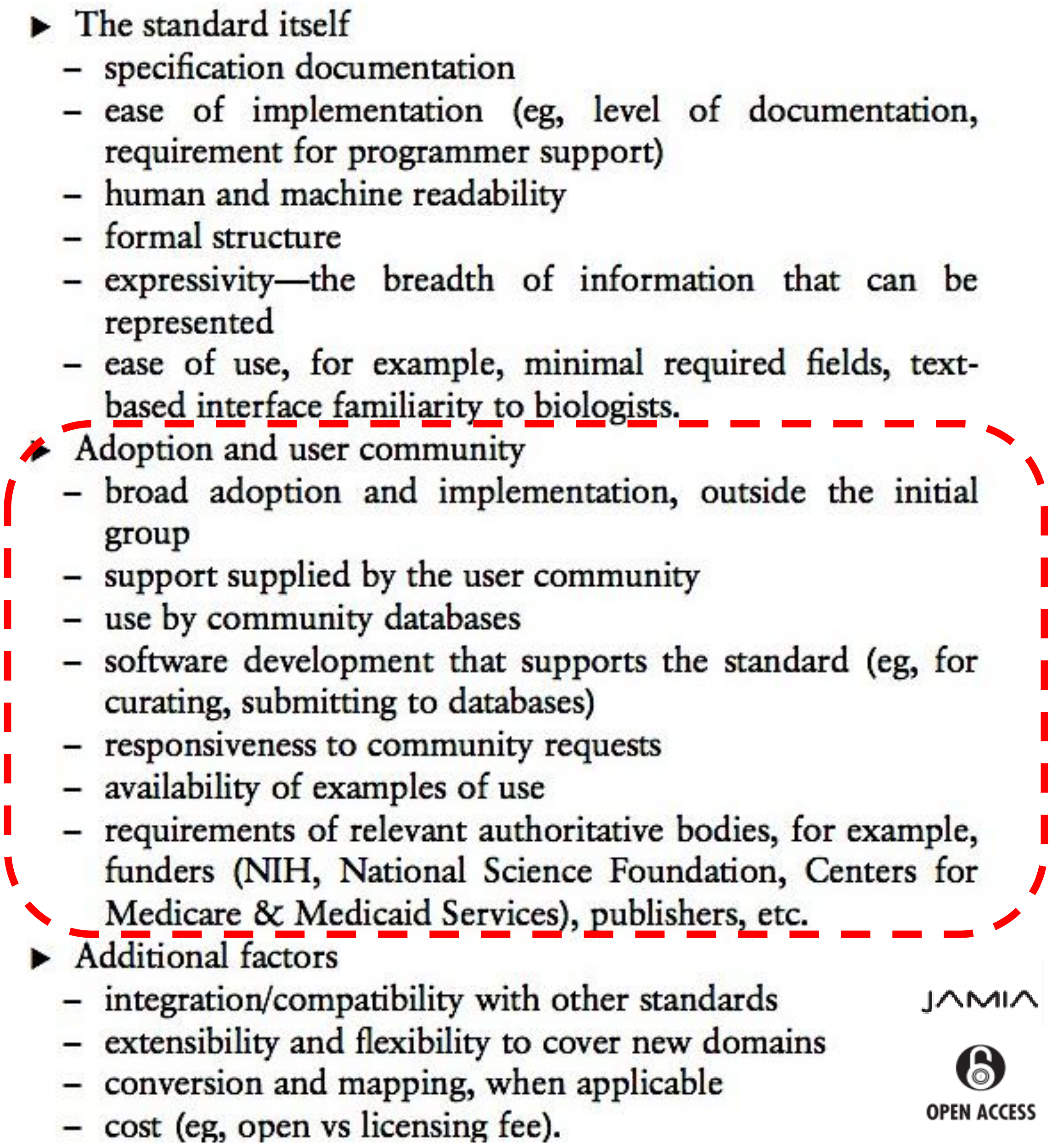
Sansone, Rocca-Serra, Field, Maguire, Taylor, Hofmann, Fang, Neumann, Tong, Amaral-Zettler, Begley, Booth, Bougueleret, Burns, Chapman, Clark, Coleman, Copeland, Das, de Daruvar, de Matos, Dix, Edmunds, Evelo, Forster, Gaudet, Gilbert, Goble, Griffin, Jacob, Kleinjans, Harland, Haug, Hermjakob, Sui, Laederach, Liang, Marshall, McGrath, Merrill, Reilly, Roux, Shamu, Shang, Steinbeck, Trefethen, Williams-Jones, Wolstencroft, Xenarios, Hide

Nature Genetics, **2012**

Pubmed Citations 9

Reader country	Reader domain	Career stage
34% United States	59% Biological Sciences	25% Ph.D. Student
18% United Kingdom	21% Computer and Information Science	13% Other Professional
9% Germany	7% Medicine	13% Researcher (at an Academic Institution)

We need “standards impact metrics” to evaluate use/usability



- The standard itself
  - specification documentation
  - ease of implementation (eg, level of documentation, requirement for programmer support)
  - human and machine readability
  - formal structure
  - expressivity—the breadth of information that can be represented
  - ease of use, for example, minimal required fields, text-based interface familiarity to biologists.
- Adoption and user community
  - broad adoption and implementation, outside the initial group
  - support supplied by the user community
  - use by community databases
  - software development that supports the standard (eg, for curating, submitting to databases)
  - responsiveness to community requests
  - availability of examples of use
  - requirements of relevant authoritative bodies, for example, funders (NIH, National Science Foundation, Centers for Medicare & Medicaid Services), publishers, etc.
- Additional factors
  - integration/compatibility with other standards
  - extensibility and flexibility to cover new domains
  - conversion and mapping, when applicable
  - cost (eg, open vs licensing fee).

JAMIA

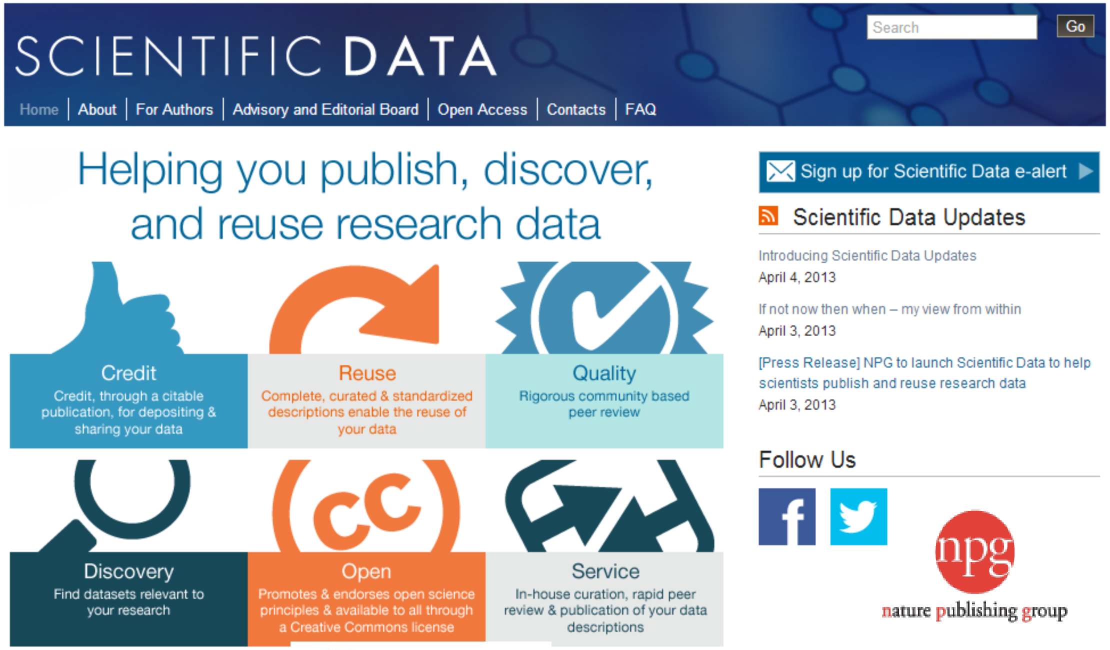
OPEN ACCESS

A sea of standards for omics data: sink or swim?

Jessica D Tenenbaum, Susanna-Assunta Sansone and Melissa Haendel

J Am Med Inform Assoc published online September 27, 2013
doi: 10.1136/amiajnl-2013-002066

biosharing.org working with data publication platforms:



SCIENTIFIC DATA

Search
Go
Home | About | For Authors | Advisory and Editorial Board | Open Access | Contacts | FAQ

Helping you publish, discover, and reuse research data

Credit
Credit, through a citable publication, for depositing & sharing your data

Reuse
Complete, curated & standardized descriptions enable the reuse of your data

Quality
Rigorous community based peer review

Discovery
Find datasets relevant to your research

Open
Promotes & endorses open science principles & available to all through a Creative Commons license

Service
In-house curation, rapid peer review & publication of your data descriptions

Sign up for Scientific Data e-alert

Scientific Data Updates

Introducing Scientific Data Updates
April 4, 2013

If not now then when – my view from within
April 3, 2013

[Press Release] NPG to launch Scientific Data to help scientists publish and reuse research data
April 3, 2013

Follow Us

npg nature publishing group

DRYAD

figshare
credit for all your research

isatools
COLLECT | CURATE | ANALYSE | SHARE | PUBLISH

F1000Research

(GIGA)^n DB

Promoting reproducible research

GigaScience's integrated "big-data" repository containing citable data available for public download and use

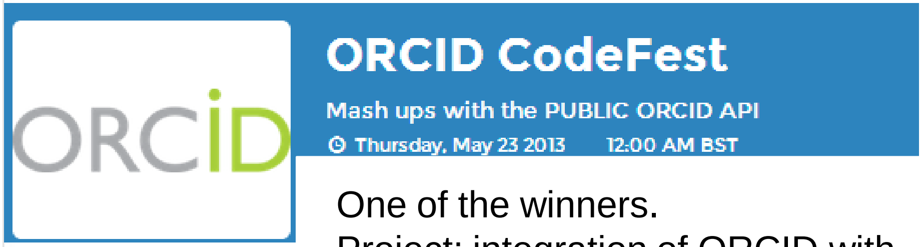
(GIGA)^n SCIENCE

华大基因 BGI

BioMed Central
The Open Access Publisher

“Invisible” use of standards in data reporting tools

Alejandra Gonzalez-Beltran, PhD

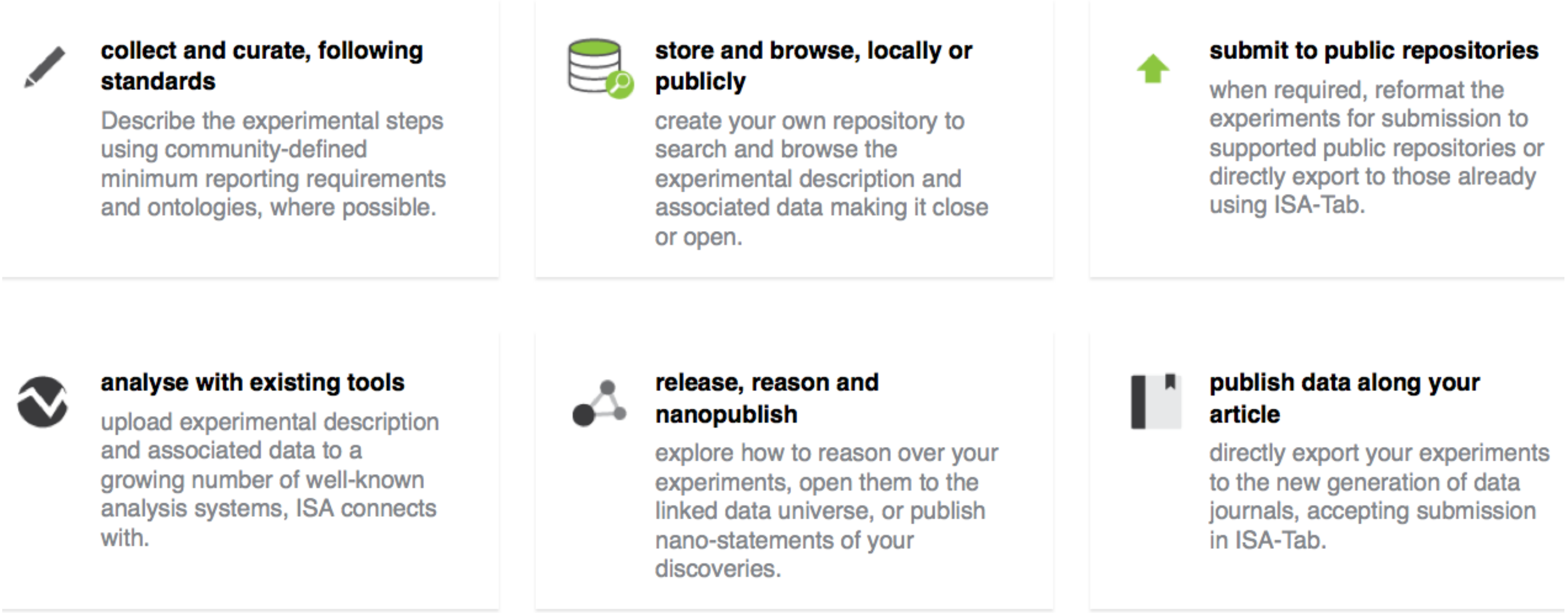


ORCID CodeFest

Mash ups with the PUBLIC ORCID API

Thursday, May 23 2013 12:00 AM BST

One of the winners.
Project: integration of ORCID with the ISAcreator, the editor tool, helping curators and researchers to describe experiments following community standards.



collect and curate, following standards

Describe the experimental steps using community-defined minimum reporting requirements and ontologies, where possible.

store and browse, locally or publicly

create your own repository to search and browse the experimental description and associated data making it close or open.

submit to public repositories

when required, reformat the experiments for submission to supported public repositories or directly export to those already using ISA-Tab.

analyse with existing tools

upload experimental description and associated data to a growing number of well-known analysis systems, ISA connects with.

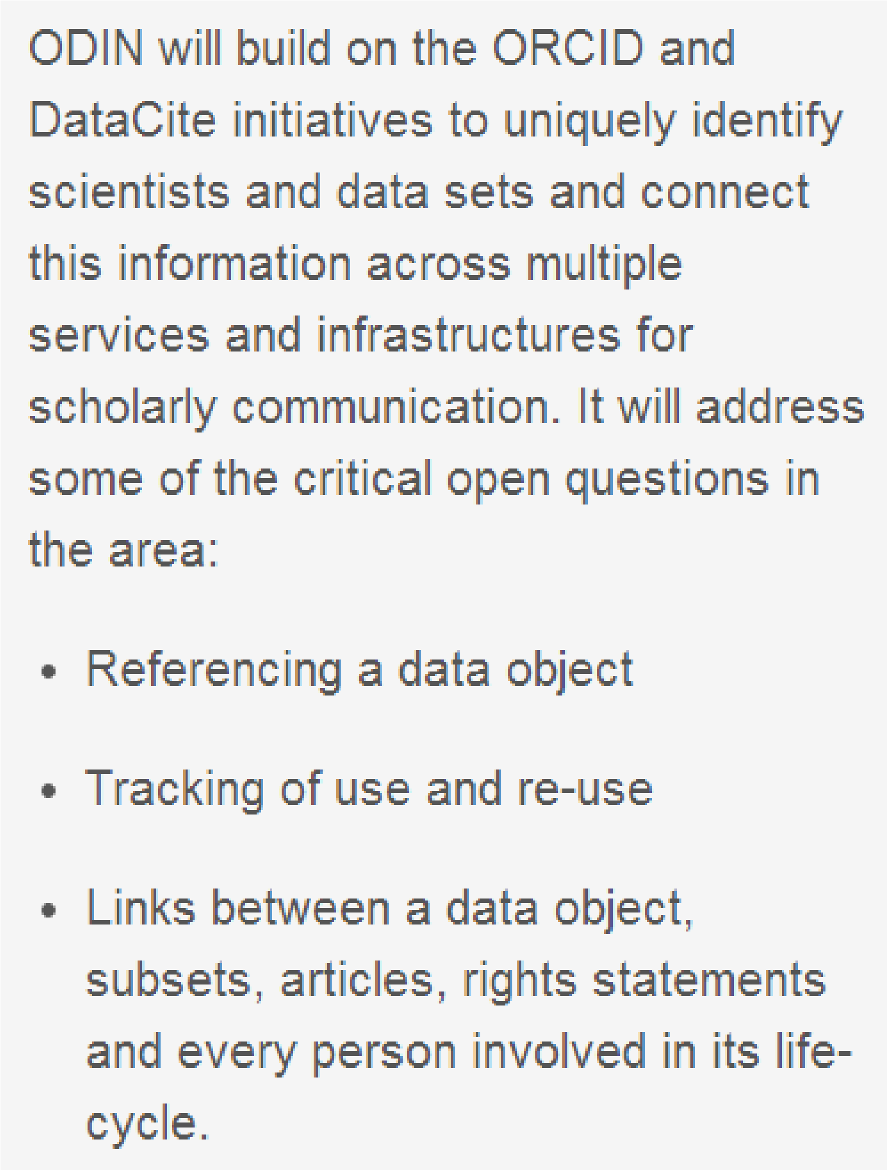
release, reason and nanopublish

explore how to reason over your experiments, open them to the linked data universe, or publish nano-statements of your discoveries.

publish data along your article

directly export your experiments to the new generation of data journals, accepting submission in ISA-Tab.

ODIN mission



ODIN will build on the ORCID and DataCite initiatives to uniquely identify scientists and data sets and connect this information across multiple services and infrastructures for scholarly communication. It will address some of the critical open questions in the area:

- Referencing a data object
- Tracking of use and re-use
- Links between a data object, subsets, articles, rights statements and every person involved in its life-cycle.

Summarizing my talk

***Problem**:*
Identification of datasets in pivotal. But meaningful sharing and (re)use also depend on how well described the datasets are.

Status quo:
In the life sciences there is a wealth of ‘reporting standards’ set to enhance and facilitate the experimental descriptions.

Challenges addressed by biosharing
Identify ‘reporting standards’ and their organizations, track their use, usability and impact (e.g. linking them to datasets), credit their developers, users (e.g. curators)...

Acknowledgements

Philippe Rocca-Serra
Alejandra Gonzalez-Beltran
Eamonn Maguire

Collaborators:

- OBO Foundry
- COSMOS
- GSC
- Metabolomics Society
- Data Dryad
- Pistoia Alliance
- Elixir UK
- NPG's Scientific Data
- and many more….

SEVENTH FRAMEWORK PROGRAMME

elixir UNITED KINGDOM

bbsrc biotechnology and biological sciences research council

NATURAL ENVIRONMENT RESEARCH COUNCIL

OXFORD e-Research CENTRE

UNIVERSITY OF OXFORD

isacommons isacommons.org

HARVARD School of Public Health · HARVARD MEDICAL SCHOOL · EMBL-EBI · DUTCH TECHCENTRE FOR LIFE SCIENCES · cgfb BIOINFORMATIQUE · HSCI HARVARD STEM CELL INSTITUTE · Argonne NATIONAL LABORATORY · nbic · Janssen Research & Development, LLC · SysMO Systems Biology of Microorganisms · CSIRO · The Novartis Institutes for BioMedical Research · FDA Food and Drug Administration NCTR National Center for Toxicological Research · NATURAL ENVIRONMENT RESEARCH COUNCIL · MARINE BIOLOGICAL LABORATORY WOODS HOLE · Bioplatforms Australia · cnrs · INRA · inist · UNIVERSITY OF OXFORD · diXa · SEURAT-1

isatools isa-tools.org

biosharing.org

BioMed Central The Open Access Publisher · science commons · OBO · mibbi · UK PUBMED CENTRAL · BioPortal · combine · DRYAD · isatools · FuGE · NIF · PeerJ · NITRC · Proteomics Standards Initiative · Sage BIONETWORKS · BRITISH LIBRARY · Pistoia Alliance · genomic STANDARDS consortium · F1000 FACULTY of 1000 · ISB International Society for Biocuration · F1000 FACULTY of 1000 · equator network · THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY · genomic observatories · DCC · BioCatalogue · (GIGA)n SCIENCE · DataCite · Science EXCHANGE · SCIENTIFIC DATA