



Standards and infrastructure for managing experimental metadata

BioInvestigation Index

(see also poster F4)

Susanna-Assunta Sansone, PhD

The European Bioinformatics Institute (EMBL-EBI)

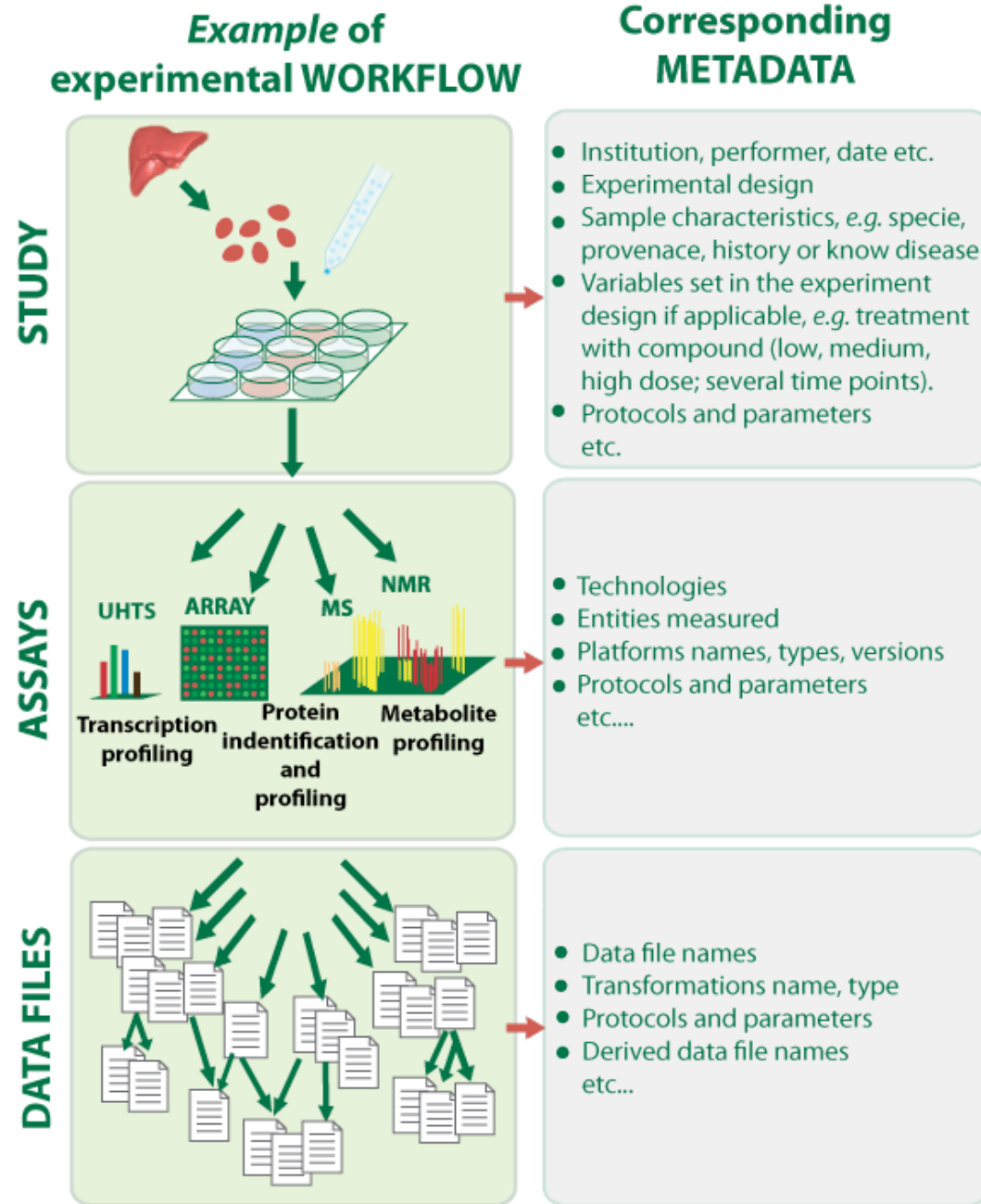
www.ebi.ac.uk/net-project

BioCurator Meeting – Berlin, April 16-19, 2009

Outline

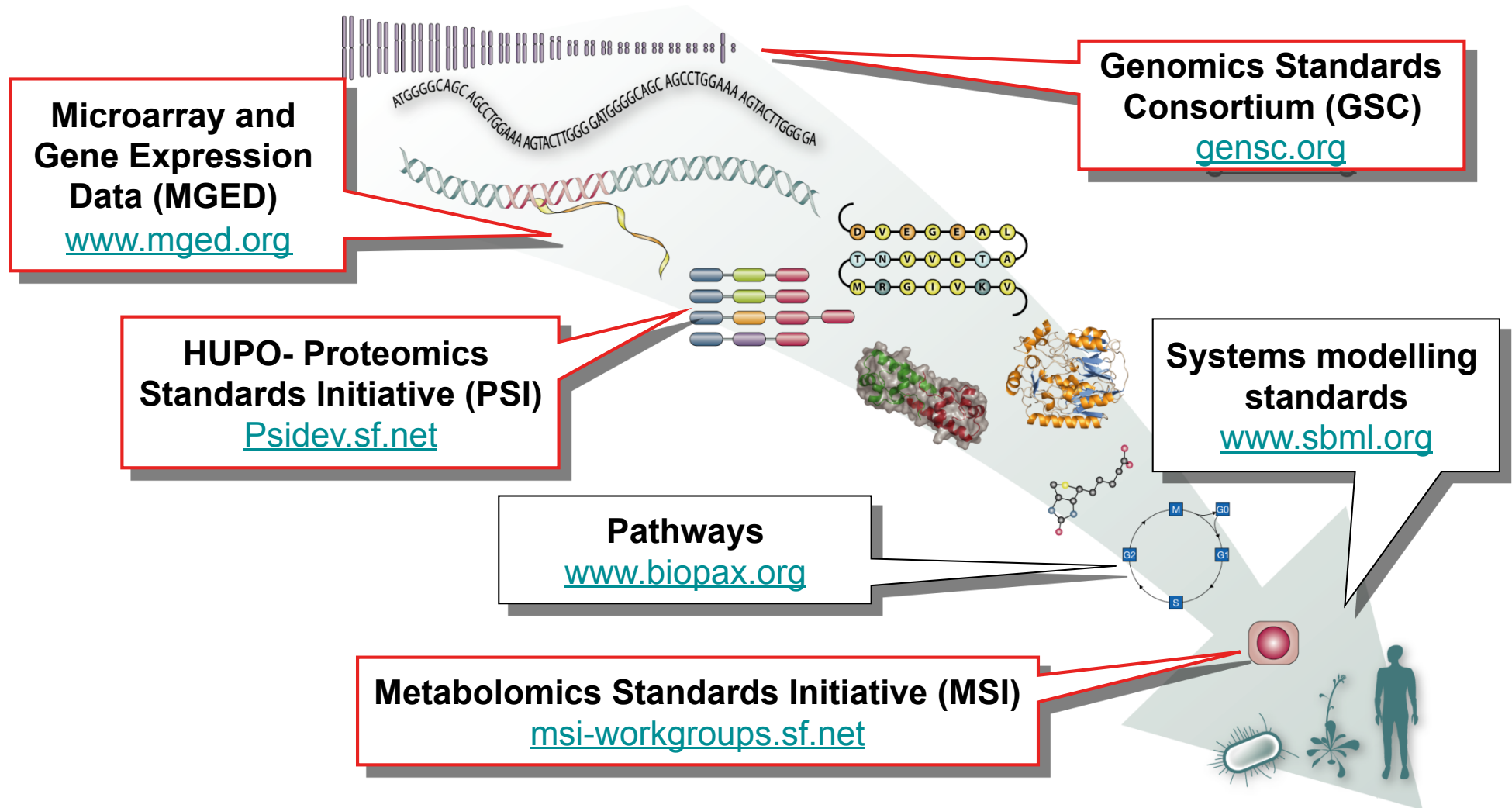
- Reporting standards
 - Hopes and hurdles
- Synergistic efforts
 - Overcome the fragmentation of standards
- Our standards-compliant implementation
 - Manage experimental metadata

Growing complexity of the experiments



Consistent reporting of the experimental metadata - along with the associated data- has a positive and long-lasting impact on the value of collective scientific outputs

Grass root omics initiatives (*de facto* standards), e.g.:



Some are loosely connected to regulatory/healthcare-driven initiatives and accredited Standards Developing Organizations (SDOs), e.g. CDISC, SEND, HL7, developing *de jure* standards.

Three types of reporting standards



Content

minimal information to be reported

```

<ArrayDesign_package>
  <ReporterGroup_asanlist>
    <ReporterGroup identifier="ebi.ac.uk:MIAMEExpress:ReporterGroup.A-MEXP-123.1"
      name="Experimental">
      <Species_asan>
        <OntologyEntry category="Organism" value="Homo sapiens">
        <OntologyReference_asan>
          /DatabaseEntry_IDI="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/tax
  
```

XML

Dataset Variables for Vital Signs - Findings

Variable	Label	Type	Code	Origin	Role	Comment
USUBJID	Unique Subject Identifier	text		Sponsor Defined		Unique subject identifier within the submission.
USUBJID	VISIT		VSTESTCD	VSORRES		
0001	1		DIABP	70		
0001	1		SYSBP	110		
0001	1		BMI	25.3		

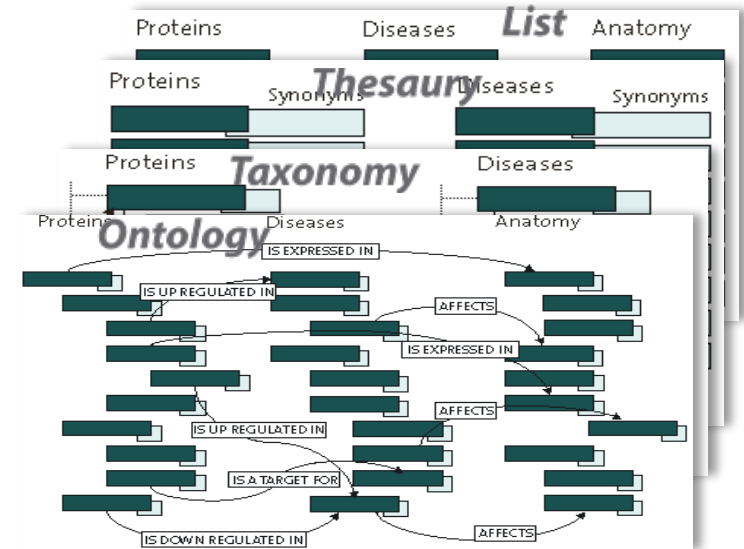
Tabular



domain experts

Syntax

format for the communication



Semantics

terminology for the description

Fragmented standards, fragmented systems, e.g.:

DIFFERENT

Access and exchange format

DIFFERENT

Core requirements captured

DIFFERENT

Terminologies

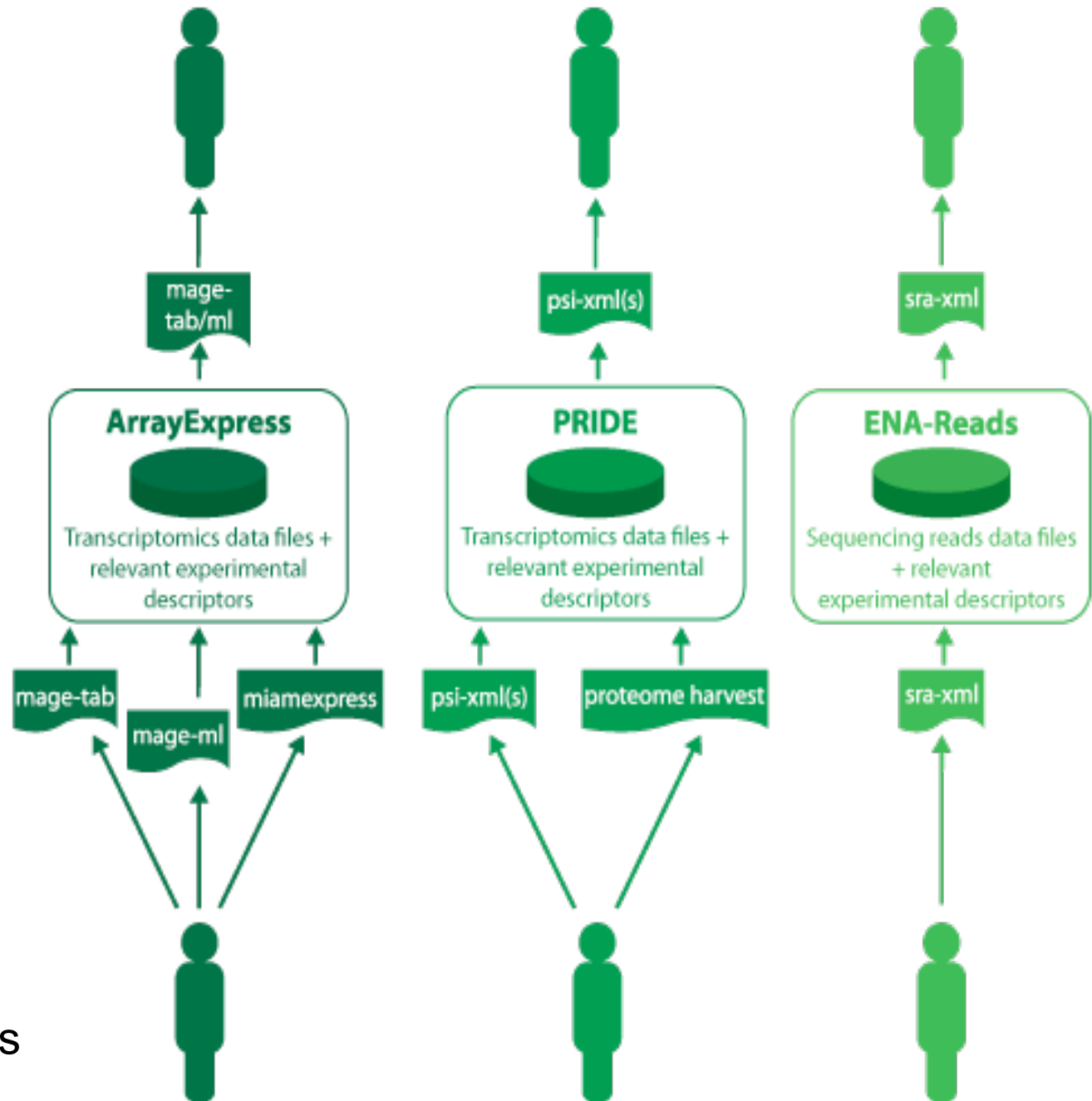
DIFFERENT

Deposition formats

DIFFERENT

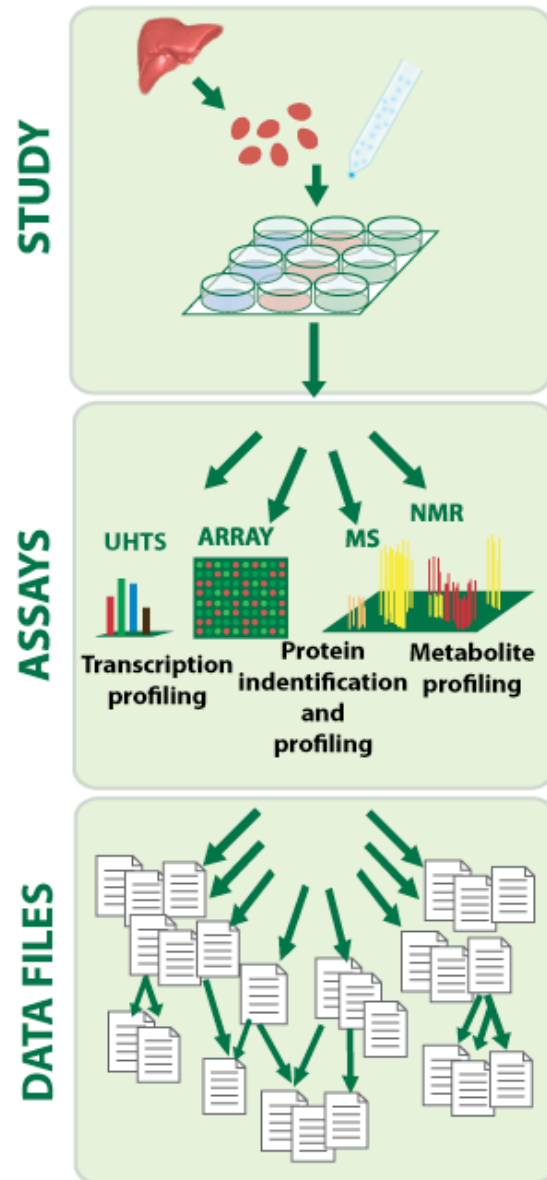
Curation practices and tools

Three EBI omics systems



But....how do we manage complex experiments?

Example of experimental WORKFLOW



How do we **encourage** submissions of experimental metadata and data and **enable** consistent reporting and curation in the current scenario?

We need to address the fragmentation of standards

- **Promote synergies among standards initiatives**
 - 'Limit' the range and variability of formats, in particular
- **Create interoperable reporting standards**
 - Fit neatly into a jigsaw, resolving inconsistency and filling gaps
- **Overcome several barriers**
 - Technical, funds and (overall) sociological.....

We need to address the fragmentation of standards

- **Promote synergies among standards initiatives**
 - 'Limit' the range and variability of formats, in particular
- **Create interoperable reporting standards**
 - Fit neatly into a jigsaw, resolving inconsistency and filling gaps
- **Overcome several barriers**
 - Technical, funds and (overall) sociological.....
- **Our* contribution to address these hurdles**
 - ✓ Risen funds to hold workshops, supporting synergistic efforts
 - ✓ Initiated new synergistic efforts, where missing
 - ✓ Work with our data producers and collaborators to implement standards-compliant systems

* *Sansone SA, Rocca-Serra P, Field D, Taylor C.*

Synergistic efforts we contribute to



Scope

minimal information to be reported

MIBBI: <http://mibbi.org>

XML

```
<arrayDesign_package>
  <ReporterGroup_asnlist>
    <ReporterGroup identifier="ebi.ac.uk:MIAMEExpress:ReporterGroup.A-MEXP-123.1"
      name="Experimental">
      <Species_asn>
        <OntologyEntry category="Organism" value="Homo sapiens">
          <OntologyReference_asn>
            <DatabaseEntry ID="http://tax.ncbi.nlm.nih.gov/Taxonomy/Browse.jsp"
              name="Homo sapiens" value="Homo sapiens" />
          </OntologyReference_asn>
        </OntologyEntry>
      </Species_asn>
    </ReporterGroup>
  </ReporterGroup_asnlist>
</arrayDesign_package>
```

Tabular

Variable	Label	Type	Code	Origin	Role	Comment
USUBJID	Unique Subject Identifier	text		Sponsor Defined		Unique subject identifier within the submission.
USUBJID	VISIT		VSTESTCD	VSORRES		
0001	1		DIABP	70		
0001	1		SYSBP	110		
0001	1		BMI	25.3		

Syntax

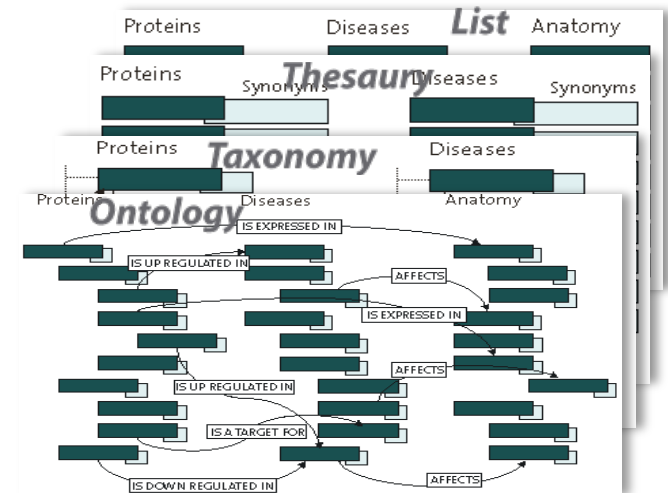
format(s) for the communication

ISA-tab: <http://isatab.sf.net>

FuGE: <http://fuge.sf.net>



domain experts



Semantics

terminology(s) for the description

OBO foundry: <http://obofoundry.org>

OBI: <http://obi-ontology.org>

Several stakeholders play pivotal role as enablers

EDITORIAL Volume 10, Number 10

October 2008

nature
cell biology

Standardizing data

Biological research is benefiting from an explosion of data. There is an urgent need to invest in bioinformatic infrastructure and education to interpret this data and guarantee its archiving.

High-throughput research has helped fuel scientific progress at an unprecedented pace and left vast amounts of digital data in its wake. Even traditional hypothesis-driven research is now published at a rate that prohibits individuals from retaining the necessary overview. Bibliographic databases, such as PubMed, are key tools to navigate the information, but do not provide access to the primary data. The value of

How then can we ensure that researchers record the appropriate metadata to allow for accurate interpretation and repetition of data, and that the data are appropriately annotated before being deposited into databases? Several communities have developed data standards, but there has been little effort to coordinate these to avoid redundancy and incompatibility. Recently, the European Bioinformatics Institute (EBI) spearheaded the development of a shared platform for such standardization initiatives. The Minimum Information for Biological and Biomedical Investigations (MIBBI) project currently encompasses a collection of 22 minimum information guidelines on techniques such as microarrays, RNAi, quantitative PCR or FACS analysis. MIBBI aims to be a 'one-stop shop' for so-called checklist projects including MIAME (Minimum Information About a Microarray Experiment) and MIAPE (Minimum Information About a Proteomics Experiment). The 'Portal' section of MIBBI contains a growing number of links to the checklist projects, whereas the 'Foundry' invites input aimed at creating new, non-redundant checklists. Active community participation in the MIBBI Foundry will help ensure that the minimum information checklists remain relevant and thus, high-throughput data adequately annotated.

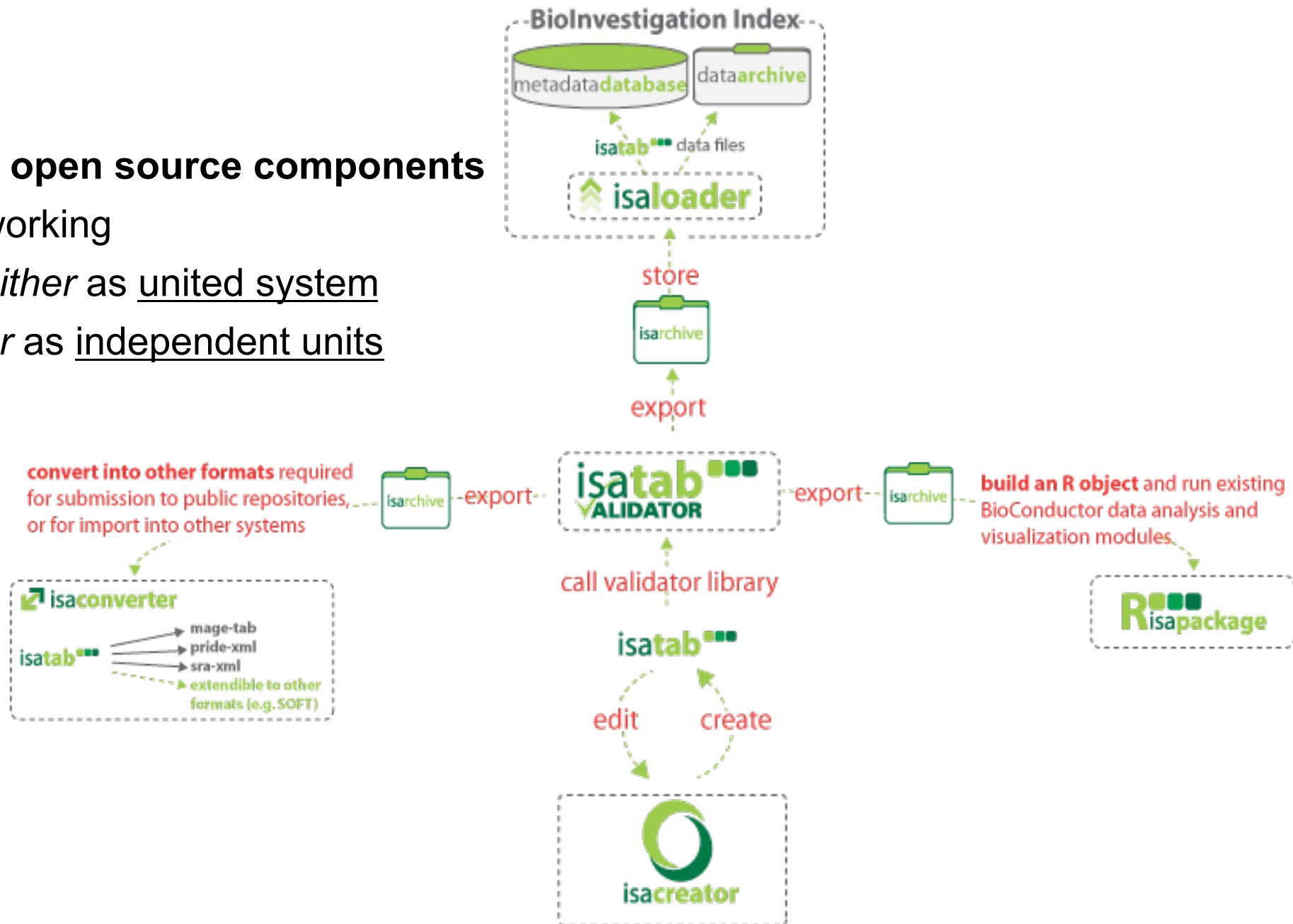
In addition to proper annotation, data must be described systematically in unambiguous language to make them machine-readable. To achieve this goal, communities must agree on ontologies (formal vocabularies for data and concepts). Ontologies allow semantic interoperability between various bioinformatics platforms and ensure that multiple repositories are compatible with each other. Although this remains a critical roadblock, the creation of the Open Biomedical Ontologies (OBO) Foundry, which acts as an umbrella organization for ontology projects, represents

BioMed Central's journals - with clinical content and BMC

Bioinformatics - now include a link to the MIBBI in the instructions for authors and encourage data deposition

Our infrastructure based on synergistic standards

5 open source components
working
either as united system
or as independent units



Component 1: ISAcreeator, standalone editor tool



Component 1: ISAcreeator, standalone editor tool

The screenshot displays the ISAcreeator software interface. The main window is titled "Assay measuring protein expression profiling using mass spectrometry". The interface includes a menu bar (file, view, help), a toolbar with icons for grid operations and factor settings (fact, char, prot, para), and a sidebar with navigation options: OVERVIEW, STUDY, and INFORMATION.

The main workspace contains a table of data transformations:

Data Transformation Name	Derived Data File	Factor Value[limiting nutrient]	Factor Value[rate]	Unit
atatransformation1	PRIDE_Exp_Complete_Ac_8761.xml	CHEBI:sulphur	0.1	l/hr
atatransformation1	PRIDE_Exp_Complete_Ac_8761.xml	CHEBI:carbon	0.1	l/hr
atatransformation1	PRIDE_Exp_Complete_Ac_8761.xml	CHEBI:Nitrogen	0.1	l/hr
atatransformation1	PRIDE_Exp_Complete_Ac_8761.xml	CHEBI:sulphur		l/hr
atatransformation1	PRIDE_Exp_Complete_Ac_8761.xml	CHEBI:carbon		l/hr
atatransformation1	PRIDE_Exp_Complete_Ac_8761.xml	CHEBI:Nitrogen		l/hr
atatransformation2	PRIDE_Exp_Complete_Ac_8762.xml	CHEBI:sulphur	0.2	l/hr
atatransformation2	PRIDE_Exp_Complete_Ac_8762.xml	CHEBI:carbon	0.2	l/hr
atatransformation2	PRIDE_Exp_Complete_Ac_8762.xml	CHEBI:Nitrogen	0.1	l/hr
atatransformation2	PRIDE_Exp_Complete_Ac_8762.xml	CHEBI:sulphur		l/hr
atatransformation2	PRIDE_Exp_Complete_Ac_8762.xml	CHEBI:carbon		l/hr
atatransformation2	PRIDE_Exp_Complete_Ac_8762.xml	CHEBI:Nitrogen		l/hr
atatransformation3	PRIDE_Exp_Complete_Ac_8763.xml	CHEBI:Nitrogen	0.2	l/hr
atatransformation3	PRIDE_Exp			
atatransformation3	PRIDE_Exp			
atatransformation3	PRIDE_Exp			
atatransformation3	PRIDE_Exp			
atatransformation3	PRIDE_Exp			

An "ontologylookup" window is open, showing search results for the term "nitrogen". The window includes a search bar, a "recommended search" section, and a "recent history" list. The search results are as follows:

ontologylookup recenthistory

recommended search all ontologies

term: nitrogen search

— 279 results in 18 ontologies

- + CCO - Cell Cycle Ontology
 - CHEBI - Chemical Entities of Biological
 - 2'-carboxy-2-methylphenylazo nit
 - Nitrogen oxide cation << 29120 >>
 - Nitrogen << 17997 >>
 - boron-carbon-nitrogen nanotube << 508 >>
 - diatomic nitrogen << 33266 >>
 - elemental nitrogen << 33267 >>

recent history:

- + UO:microliter
- + EFO:genotype
- + NEWT:Saccharomyces cerevisia
- + CHEBI:carbon
- + CHEBI:phosphorous acid
- + EFO:_temp bin organism part
- + OBI:protein extraction
- + OBI:mass spectrometry
- + EFO:time

selected term(s) CHEBI:Nitrogen close ok



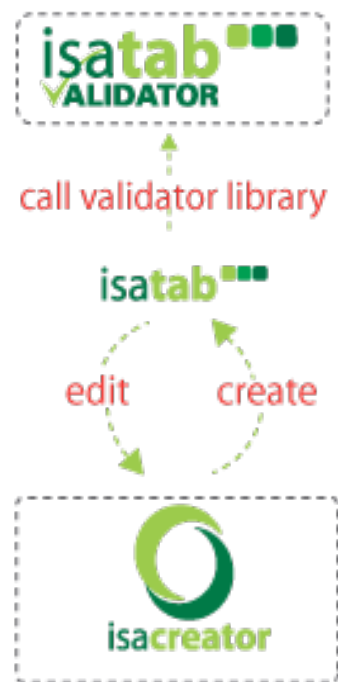
Component 1: ISAcreeator, standalone editor tool

The screenshot displays the ISAcreeator software interface. At the top, there is a menu bar with 'file', 'view', and 'help'. Below the menu is a green header with 'OVERVIEW' and a document icon. The main content area is divided into several sections:

- Overview Panel (Left):** Shows a project titled 'Growth control of the e...'. Underneath, there are two main sections: 'BII-S-1' and 'BII-S-2'. Under 'BII-S-1', there are four sub-items: 's_BII-S-1.txt', 'a_proteome.txt', 'a_metabolome.txt', and 'a_transcriptome.txt'. Under 'BII-S-2', there is one sub-item: 'BII-S-2'.
- Help Panel (Top Right):** Contains a 'help' icon and text: 'to recenter the graph on screen, right click on the graph area.', 'to move around the view area, hold the left mouse button and move the mouse in the direction you wish to navigate to.', 'to zoom, either hold the right mouse button down and move up or down to zoom in or out respectively, or use the scroller wheel on your mouse (if you have one).', and 'to get more information about a graph item, click on it.'
- Investigation Diagram (Center):** A tree diagram showing 'Investigation' branching into 'BII-S-1' and 'BII-S-2'. 'BII-S-1' further branches into 'a_proteome.txt (protein expression profiling using mass spectrometry)', 'a_metabolome.txt (metabolite profiling using mass spectrometry)', and 'a_transcriptome.txt (transcription profiling using DNA microarray)'. 'BII-S-2' branches into 'a_microarray.txt (transcription profiling using DNA microarray)'.
- Zoomed View (Bottom Left):** A detailed view of the 'a_metabolome.txt' dataset. It shows several circular nodes representing different metabolites. Each node is labeled with its name and experimental conditions, such as 'nitrogen 0.07 l/hr (10 samples)', 'nitrogen 0.2 l/hr (6 samples)', 'nitrogen 0.1 l/hr (6 samples)', and 'carbon 0.07 l/hr (10 samples)'. There are also labels for 'phosphate 0.1 l/hr (6 samples)' and 'nitrogen 0.1 l/hr (6 samples)'.
- Information Panel (Bottom Right):** A panel with the title 'a_metabolome.txt' and two buttons: 'view information' and 'view sample names'.



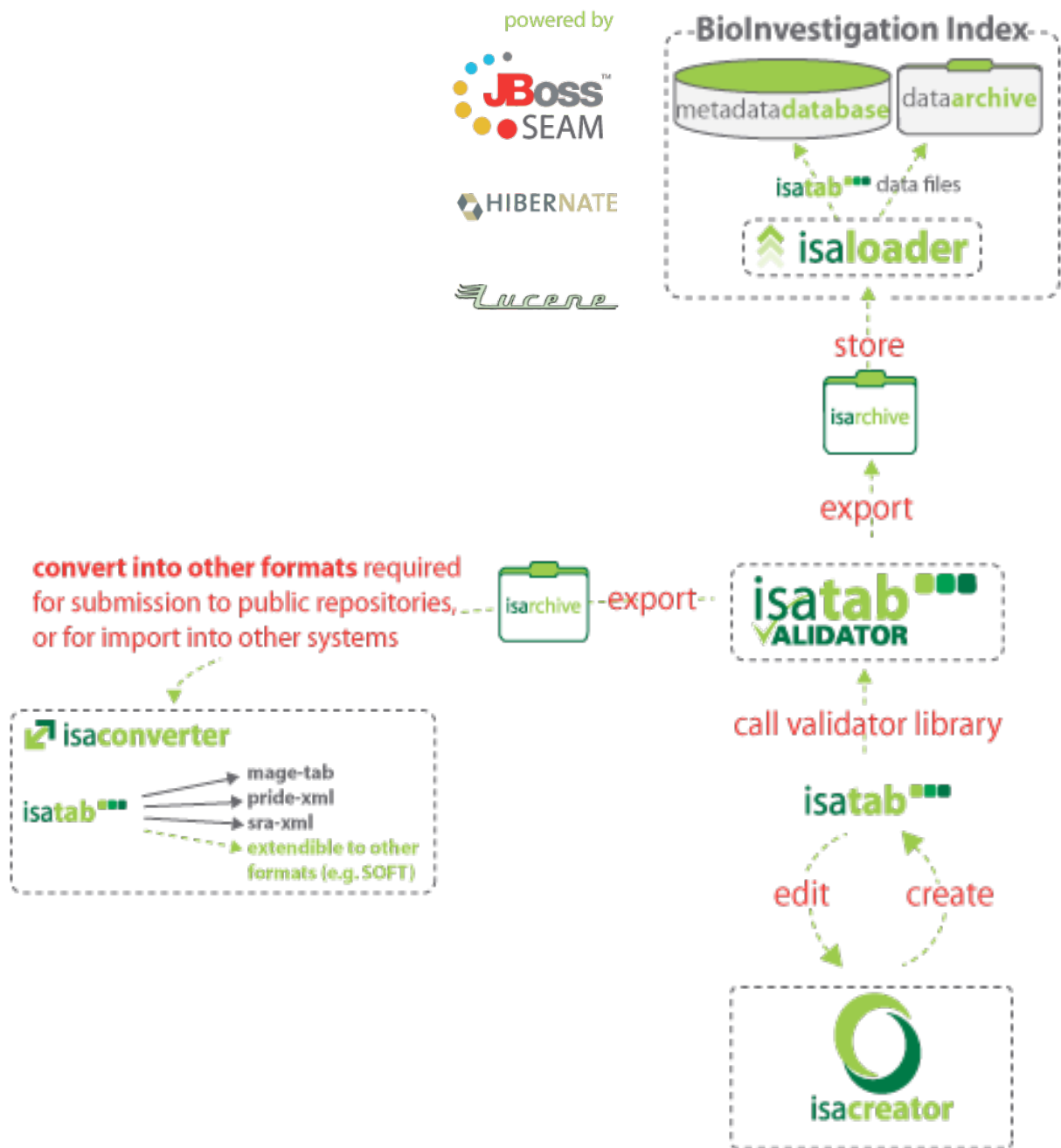
Component 2: ISAvalidator



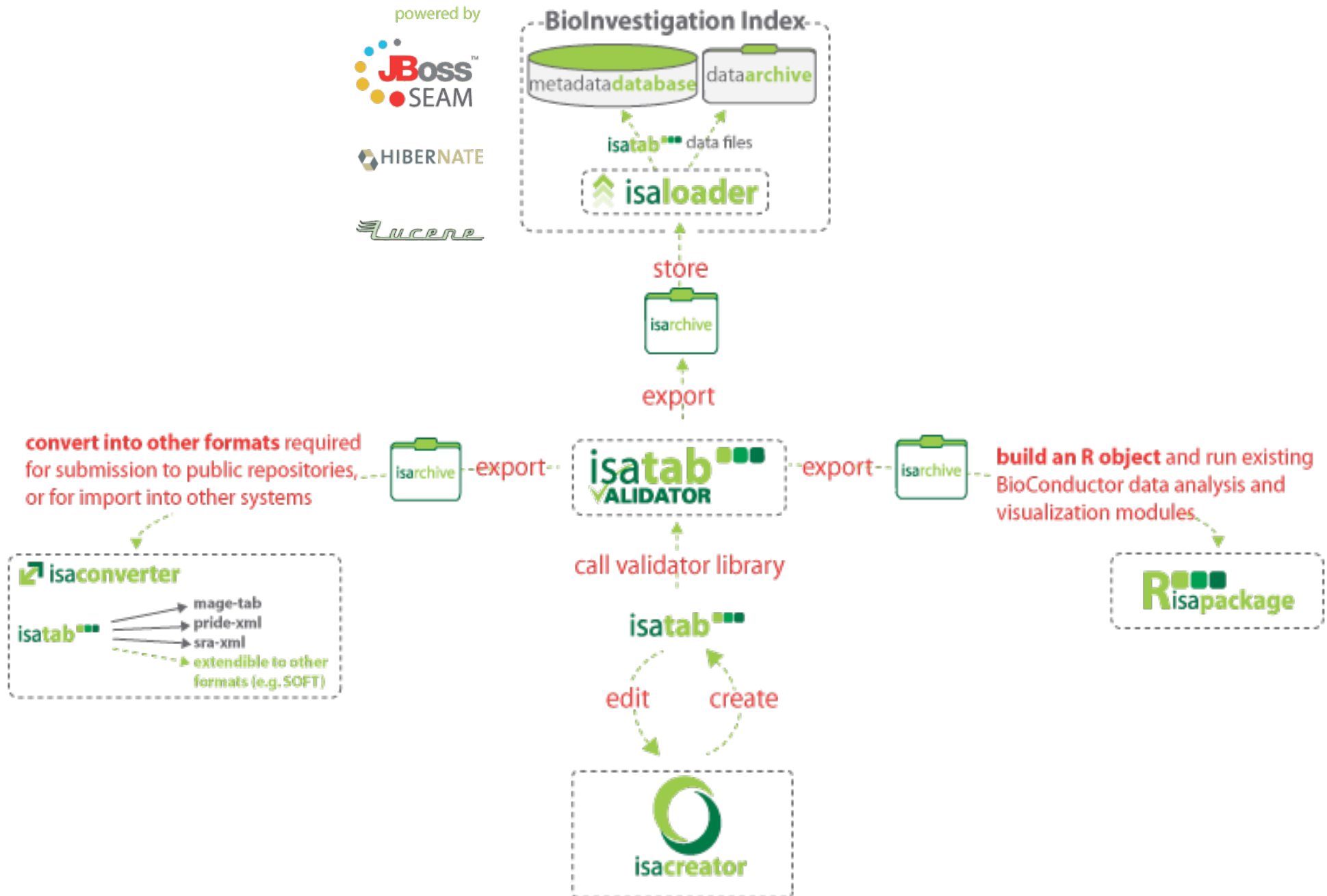
Component 3: BioInvestigation Index database



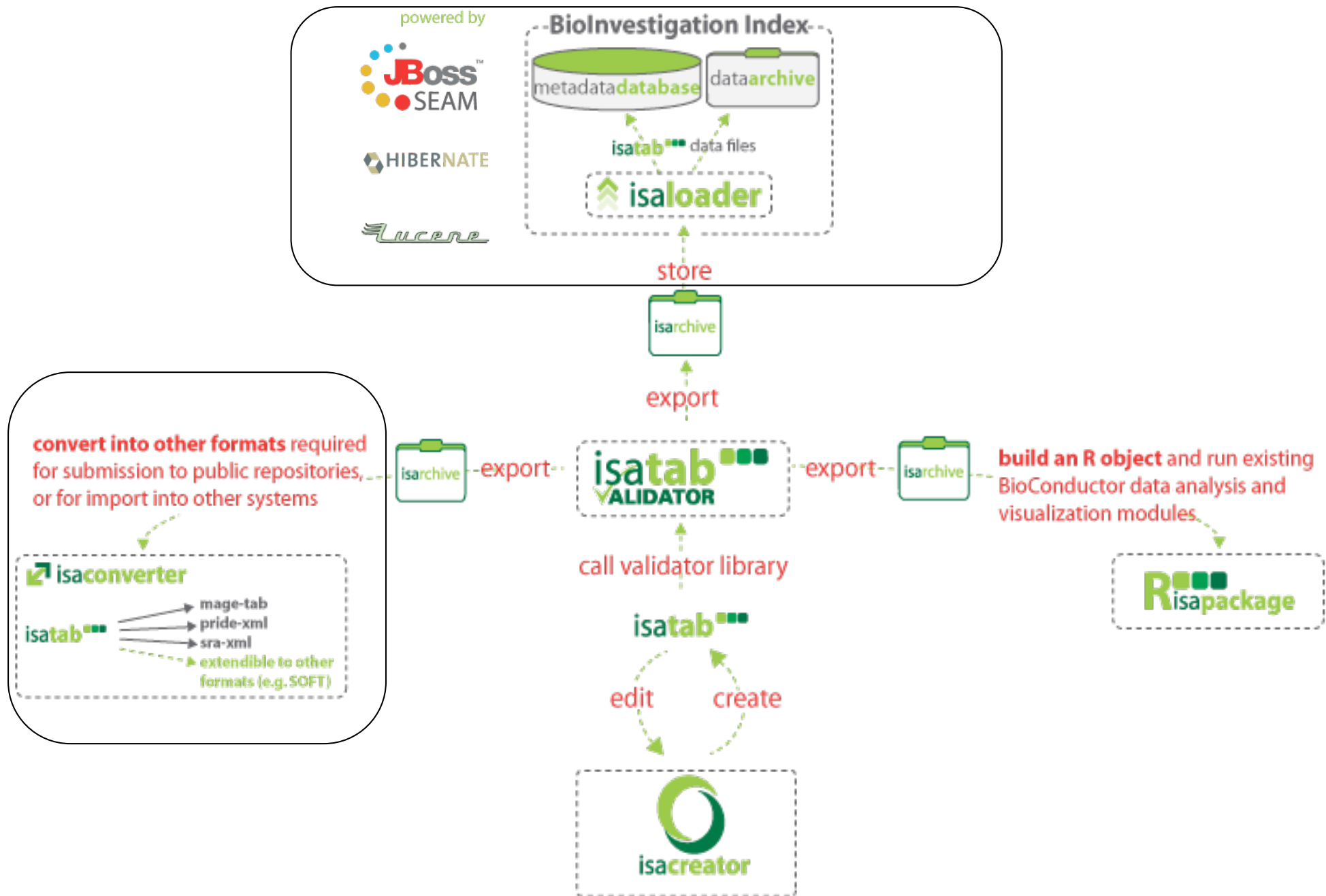
Component 4: ISAconverter



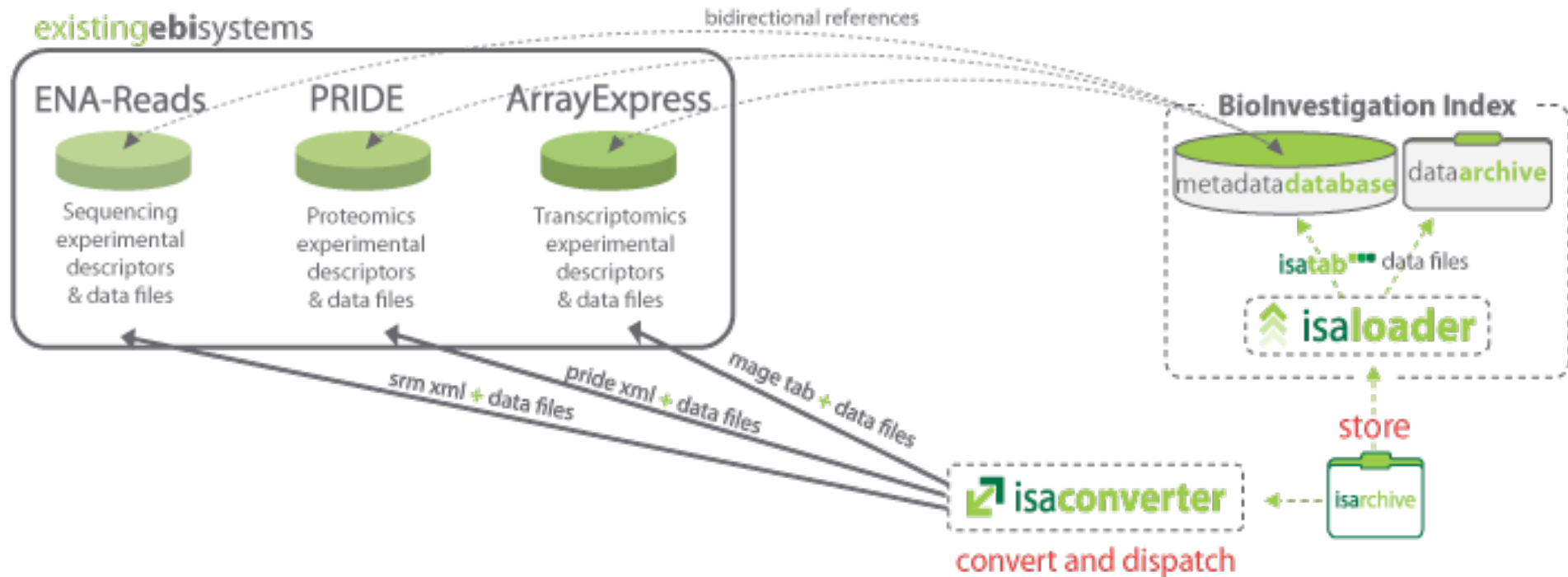
Component 5: R package for ISA-TAB (ongoing)



Instance deployed at EBI, as prototype



Instance deployed at EBI, as prototype



<http://www.ebi.ac.uk/bioinvindex>

Public studies are visible and searchable

bii bio investigation index

- [\(b\) browse](#)
- [\(s\) submit](#)
- [\(e\) contact](#)
- [\(c\) credits](#)

search

Filter on organisms | Filter on measurement | Filter on technology | Filter on Platform

clearfields searchIndex

views

COMPOUND-CENTRIC [VIEW](#)

browsestudies

6 public studies containing 282 assays

Investigation		Study			Assay		
Acc	Acc	Title	Organism	Factor	Measurement	Technology	#
	BII-S-5	Determination of the complete genome sequence of Salmonella paratyphi A str. AKU_12601	Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601		genome sequencing	nucleotide sequencing	1
	BII-S-6	The Influence of Pharmacogenetics on Fatty Liver Disease in the Wistar and Kyoto Rats: A Combined Transcriptomic and Metabonomic	Rattus norvegicus (Rat)	time, strain, compound	transcription profiling metabolite profiling	DNA microarray NMR spectroscopy	17 79
	BII-S-3	Metagenomes and Metatranscriptomes of phytoplankton blooms from an ocean acidification mesocosm experiment	marine metagenome	dose, collection time, compound	metagenome sequencing transcription profiling	nucleotide sequencing nucleotide sequencing	4 4
	BII-S-4	An initial characterisation of the Fasciola hepatica transcriptome using 454-FLX sequencing	Fasciola hepatica (Liver fluke)		transcription profiling	nucleotide sequencing	1
BII-I-1	BII-S-2	A time course analysis of transcription response in yeast treated with rapamycin, a specific inhibitor of the TORC1 complex: impact on yeast growth	Saccharomyces cerevisiae (Baker's yeast)	exposure time, dose, compound	transcription profiling	DNA microarray	14
BII-I-1	BII-S-1	Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast Saccharomyces cerevisiae under different nutrient limitations	Saccharomyces cerevisiae (Baker's yeast)	limiting nutrient, rate	protein expression profiling transcription profiling metabolite profiling	mass spectrometry DNA microarray mass spectrometry	3 48 111

studyinformation

Investigation: This Study is part of an Investigation, which also includes: [BII-S-2](#)
Study ID: BII-S-1
Title: Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast *Saccharomyces*

Organism(s):
Description: **Guideline(s) followed:** [CIMR](#) [MIAME](#) [MIAPE](#)
Download:



more information about the study including protocols | open isatab in spreadsheet software or download, import and sample processing steps... and view it in the **isacreator**

Design(s):
Publication(s):

ASSAY DATA FILES & RECORDS

the assays associated with this study are listed below with links to their raw and processed data files (if available) as well as links to submission records in other repositories (where applicable)...

Sample attribute(s):

assaytype

Measurement: **transcription profiling**
Technology: **DNA microarray**
Platform: **Affymetrix**



RawData



ProcessedData

View **ArrayExpress** Entry For **E-MEXP-115**

Experimental factor(s):

assaytype

Measurement: **metabolite profiling**
Technology: **mass spectrometry**
Platform: **LC-MS/MS**



RawData



ProcessedData

assaytype

Measurement: **protein expression profiling**
Technology: **mass spectrometry**
Platform: **iTRAQ**



ProcessedData

View **PRIDE** Entry For **8761**



ProcessedData

View **PRIDE** Entry For **8763**



ProcessedData

View **PRIDE** Entry For **8762**

Contact(s): Castrillo I Juan, Stephen G Oliver, Zeef A Leo

Acknowledgements and references

Marco Brandizi (*Software Engineer*)
Eamonn Maguire (*Software Engineer*)
Nataliya Sklyar (*Software Engineer*)
Chris Taylor (*Bioinformatician*)

Open source codes,
soon: <http://isatab.sf.net>

Posters: F4, E27, E3

Manon Delahaye (*Trainees -Software Engineer*)
Richard Evans (*Trainees -Software Engineer*)

Technical Coordinator: **Philippe Rocca-Serra**

Coordinator: **Susanna-Assunta Sansone**



European
Nutrigenomics
Organisation



The National Center for
Toxicological Research (NCTR)
Center for Toxicoinformatics