ISA infrastructure:

collecting & managing functional genomics datasets with rich semantics

Rencontres GenOuest, 2011, Rennes, 18th October 2011

Philippe Rocca-Serra (Ph. D)

ISA Team

philippe.rocca-serra@oerc.ox.ac.uk

http://www.isa-tools.org

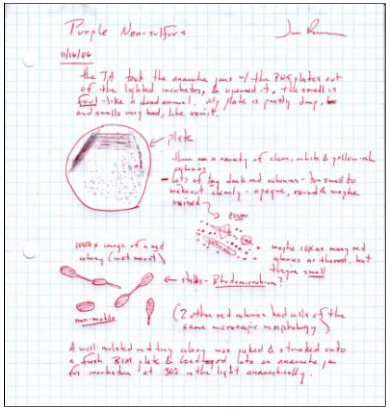
Presentation Outline

- Background information
- Rationale for developing ISA-tools
- ISA paradigm and interaction with ontologies
- Moving on: ISA future developments

MAIN THEME:

It is all about structuring experimental information to make it available to computer and software agents to enable mining.

But let's proceed gradually…



Purple Non-sulfurs

11/16/06

the TA took the anaerobe jars w/ the PNS plates out of the lighted incubator, & opened it. the smell is foul - like a dead animal. My plate is pretty damp, and smells very bad, like vomit.

plate

there are a variety of clear, white & yellow-ish colonies

- lots of tiny dark red colonies - too small to make out clearly - opaque, round & maybe raised

zoom

1000x image of a red colony (wet mount)

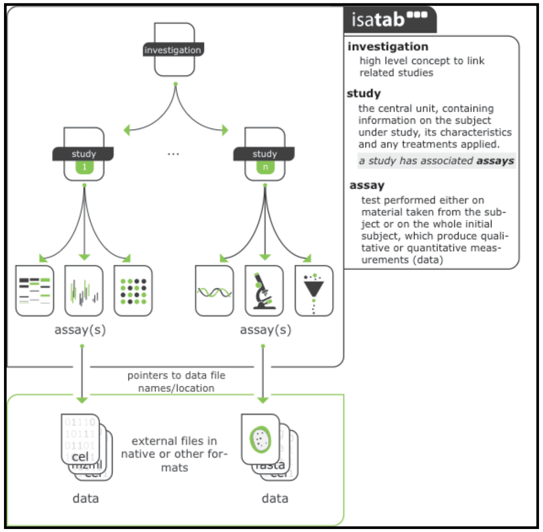
maybe 10x as many red colonies as the rest, but they're small

motile - Rhodomicrobium?

non-motile

(2 other red colonies had cells of the same microscopic morphology)

A well-isolated red tiny colony was picked & streaked onto a fresh BIM plate & loaded into an anaerobe jar for incubation at 30° in the light anaerobically.

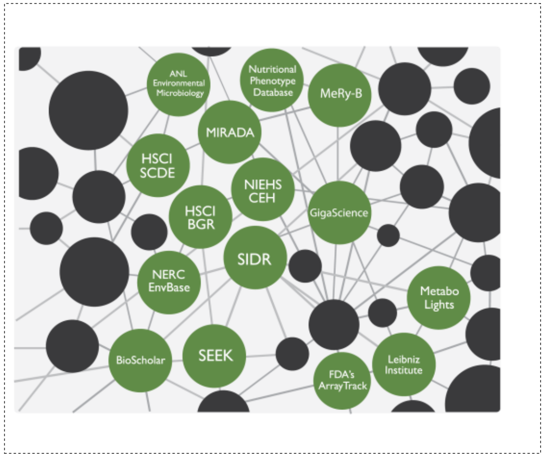

isatab
investigation
study
1
...
study
n
assay(s)
assay(s)
pointers to data file names/location
cel
mzml
cel
external files in native or other formats
fasta
cel
data
data

investigation
high level concept to link related studies

study
the central unit, containing information on the subject under study, its characteristics and any treatments applied.

a study has associated ***assays***

assay
test performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements (data)



Notes in Lab Books
(information for humans)

Spreadsheets and Tables
(the compromise)

Facts as RDF statements
(information for machines

What’s wrong with free text in electronic records?

An example: {WT, wildtype, wild-type, sauvage, wildtypo}

- Hard to index
- Hard to search
- Poor query recall

+>Unhappy users & missed opportunities

Limit free text by means of controlled terminologies and ontologies

Ontologies ?....it is about formalizing knowledge

- Organizing types into broad categories (e.g. `Objects`, `Subjects, Properties, Events` or `Material, Qualities, Processes`)

- Defining the properties of those types using sets of relations (e.g. `is_a, has_part/part_of, derives_from, located_in, participates_to`)

- Advanced software supporting the validation of those representations: Reasoners (Fact, Pellet, Hermit or ELK for OWL ontologies), .

How can this be useful?

Just one simple example:

•It makes possibly things like query expansion:

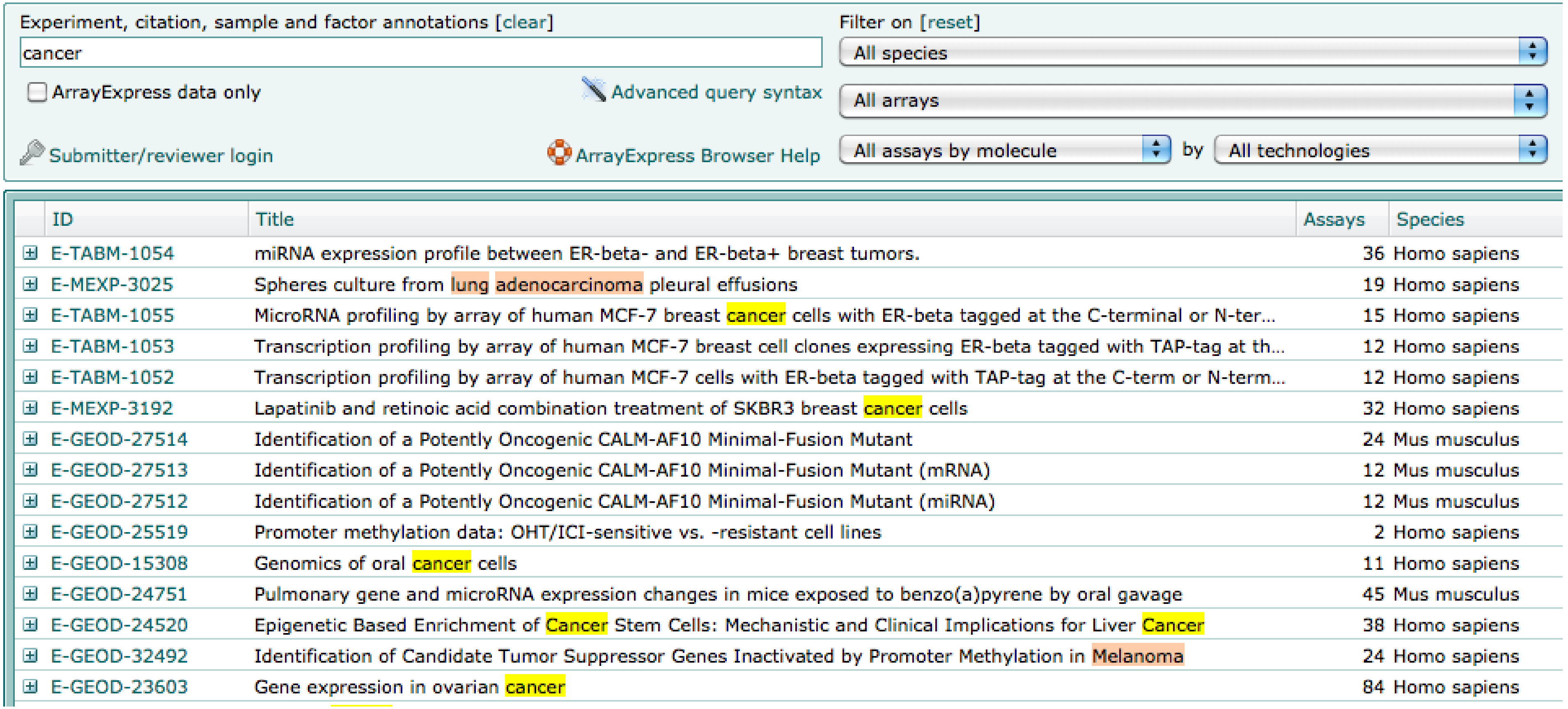
Searching for word ‘cancer’ should retrieve:

{‘carcinoma, adenocarcinoma, lipoma, sarcoma….}

•How does it work?

+> taking advantage of the ‘is_a’ relationship between those entities

An example of query expansion



Experiment, citation, sample and factor annotations [clear]

cancer

□ ArrayExpress data only

Filter on [reset]

All species

All arrays

All assays by molecule by All technologies

	ID	Title	Assays	Species
⊞	E-TABM-1054	miRNA expression profile between ER-beta- and ER-beta+ breast tumors.	36	Homo sapiens
⊞	E-MEXP-3025	Spheres culture from lung adenocarcinoma pleural effusions	19	Homo sapiens
⊞	E-TABM-1055	MicroRNA profiling by array of human MCF-7 breast cancer cells with ER-beta tagged at the C-terminal or N-ter...	15	Homo sapiens
⊞	E-TABM-1053	Transcription profiling by array of human MCF-7 breast cell clones expressing ER-beta tagged with TAP-tag at th...	12	Homo sapiens
⊞	E-TABM-1052	Transcription profiling by array of human MCF-7 cells with ER-beta tagged with TAP-tag at the C-term or N-term...	12	Homo sapiens
⊞	E-MEXP-3192	Lapatinib and retinoic acid combination treatment of SKBR3 breast cancer cells	32	Homo sapiens
⊞	E-GEOD-27514	Identification of a Potently Oncogenic CALM-AF10 Minimal-Fusion Mutant	24	Mus musculus
⊞	E-GEOD-27513	Identification of a Potently Oncogenic CALM-AF10 Minimal-Fusion Mutant (mRNA)	12	Mus musculus
⊞	E-GEOD-27512	Identification of a Potently Oncogenic CALM-AF10 Minimal-Fusion Mutant (miRNA)	12	Mus musculus
⊞	E-GEOD-25519	Promoter methylation data: OHT/ICI-sensitive vs. -resistant cell lines	2	Homo sapiens
⊞	E-GEOD-15308	Genomics of oral cancer cells	11	Homo sapiens
⊞	E-GEOD-24751	Pulmonary gene and microRNA expression changes in mice exposed to benzo(a)pyrene by oral gavage	45	Mus musculus
⊞	E-GEOD-24520	Epigenetic Based Enrichment of Cancer Stem Cells: Mechanistic and Clinical Implications for Liver Cancer	38	Homo sapiens
⊞	E-GEOD-32492	Identification of Candidate Tumor Suppressor Genes Inactivated by Promoter Methylation in Melanoma	24	Homo sapiens
⊞	E-GEOD-23603	Gene expression in ovarian cancer	84	Homo sapiens

[Source: http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=cancer&expandefo=on.]

The system may suggest unseen association, could help generate new hypothesis -> Happier users

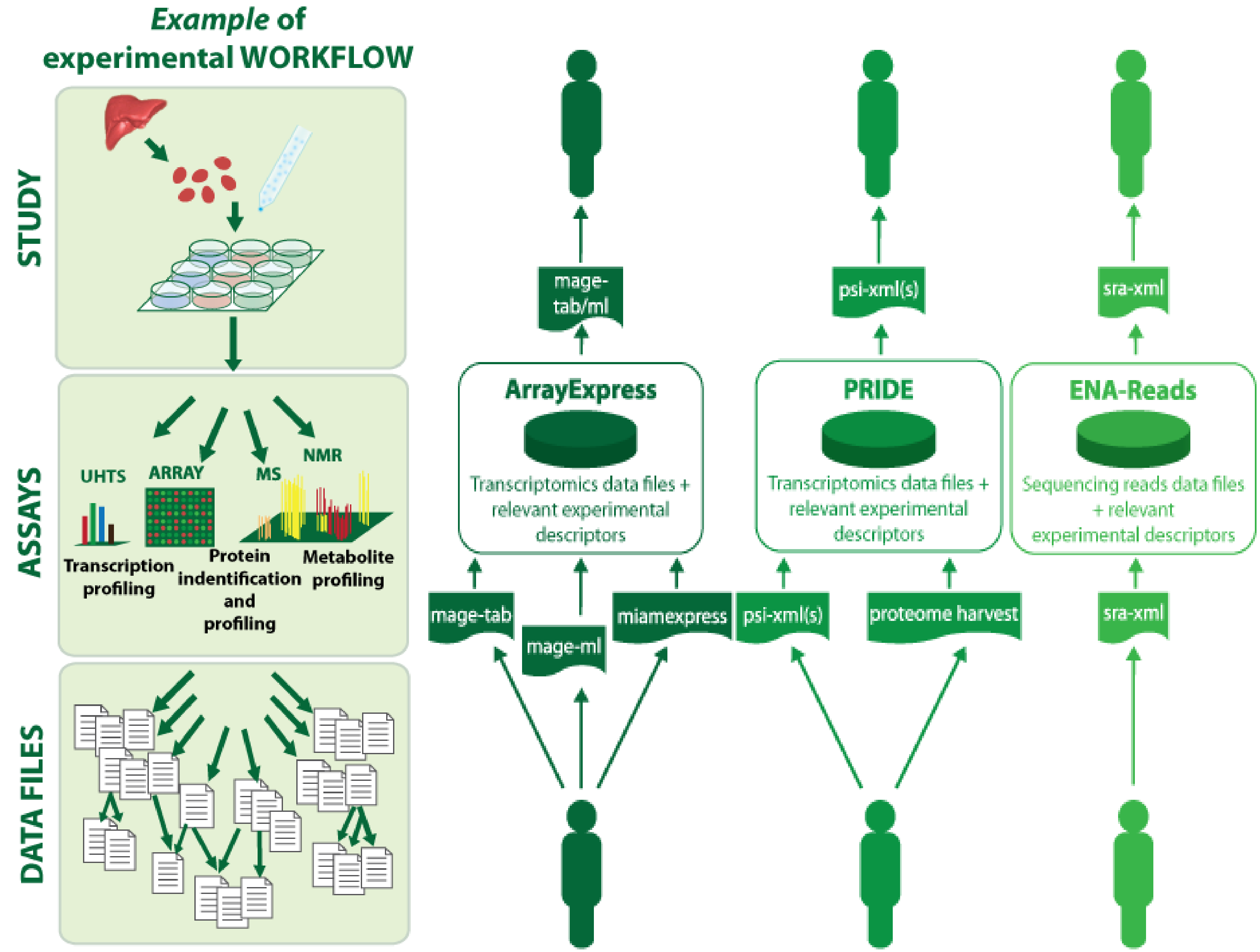
Presentation Outline

- Background information
- Rationale for developing ISA-tools
- ISA paradigm and interaction with ontologies
- Moving on: ISA future developments

Observations

- Experiments are expensive, often publicly funded, still many fail to see the light.
- Spreadsheets are the most common vehicle for so-called ‘omics’ (functional genomics) experimental metadata tracking
- technology centric repositories form de facto silos
- conversions are required to allow for deposition to public databases.
- submitting to common information across a series of repositories is inefficient

Case Study


Example of
experimental WORKFLOW
STUDY
ASSAYS
UHTS
ARRAY
MS
NMR
Transcription profiling
Protein indentification and profiling
Metabolite profiling
DATA FILES
mage-tab/ml
psi-xml(s)
sra-xml
ArrayExpress
Transcriptomics data files + relevant experimental descriptors
PRIDE
Transcriptomics data files + relevant experimental descriptors
ENA-Reads
Sequencing reads data files + relevant experimental descriptors
mage-tab
mage-ml
miamexpress
psi-xml(s)
proteome harvest
sra-xml

CORBIN BERNSEN
Sweet Dreams
THE DENTIST 2

Observations II

- A growing number to ‘annotations requirements’ (a.k.a MI checklists)
- Many different communities, many different needs.

mibbi

- Creation of the MIBBI portal to harmonize and identify a core of common descriptors, create extensions where necessary.

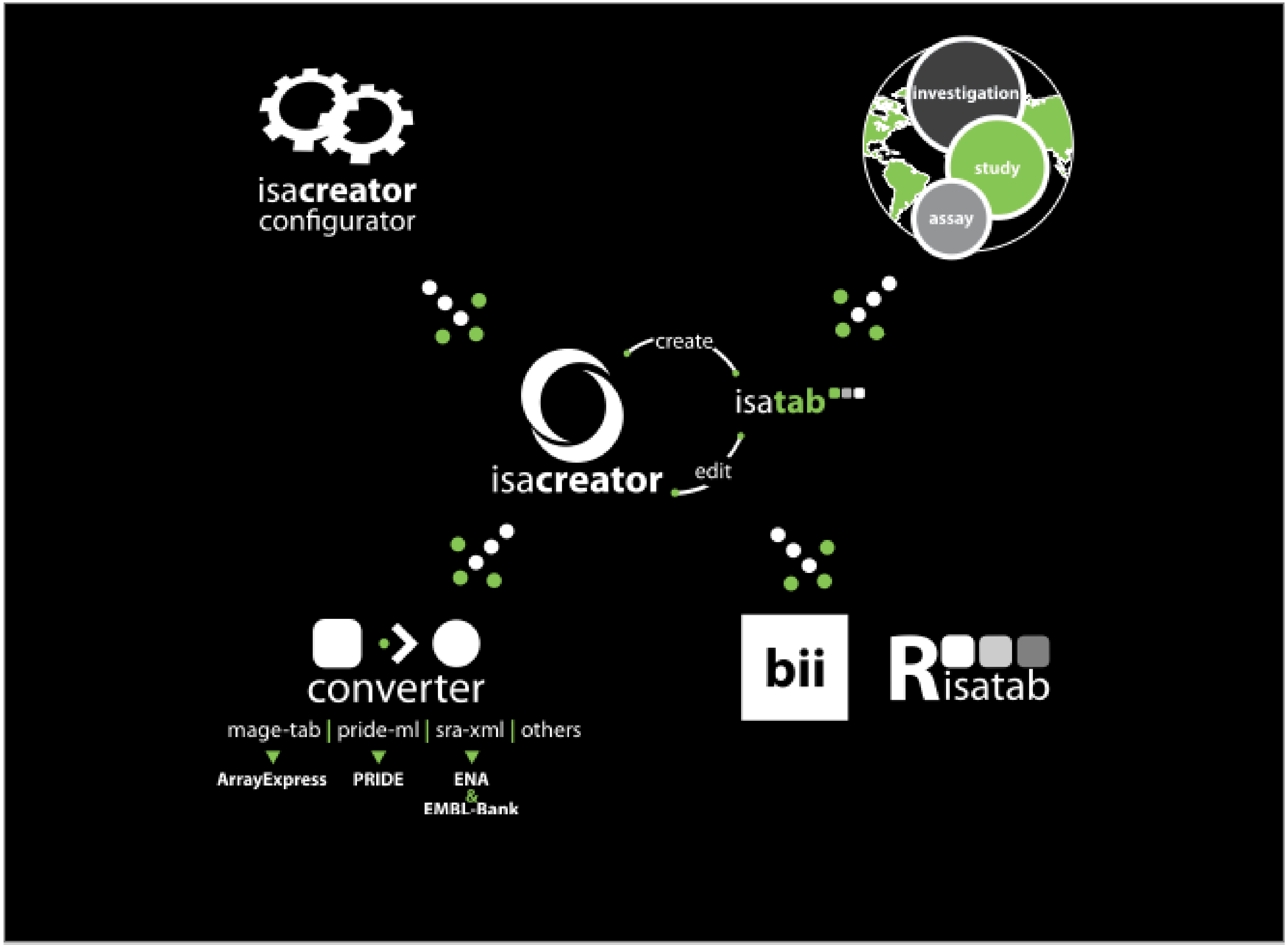
Many Requirements, Many Formats, Many ontologies …

- To support different fields of molecular biology:
  - Soil Metagenomics
  - Cancer genomics
  - Chromatin remodelling event and Stem Cell fate.
- To support various cases in data reporting & data management
  - Dealing with legacy data (spreadsheets hanging around)
    - Data Mapping and Import Function from files.
  - Dealing with de novo datasets:
    - Planning and Templating: reduce repetitive tasks by relying on patterns found in experimental designs

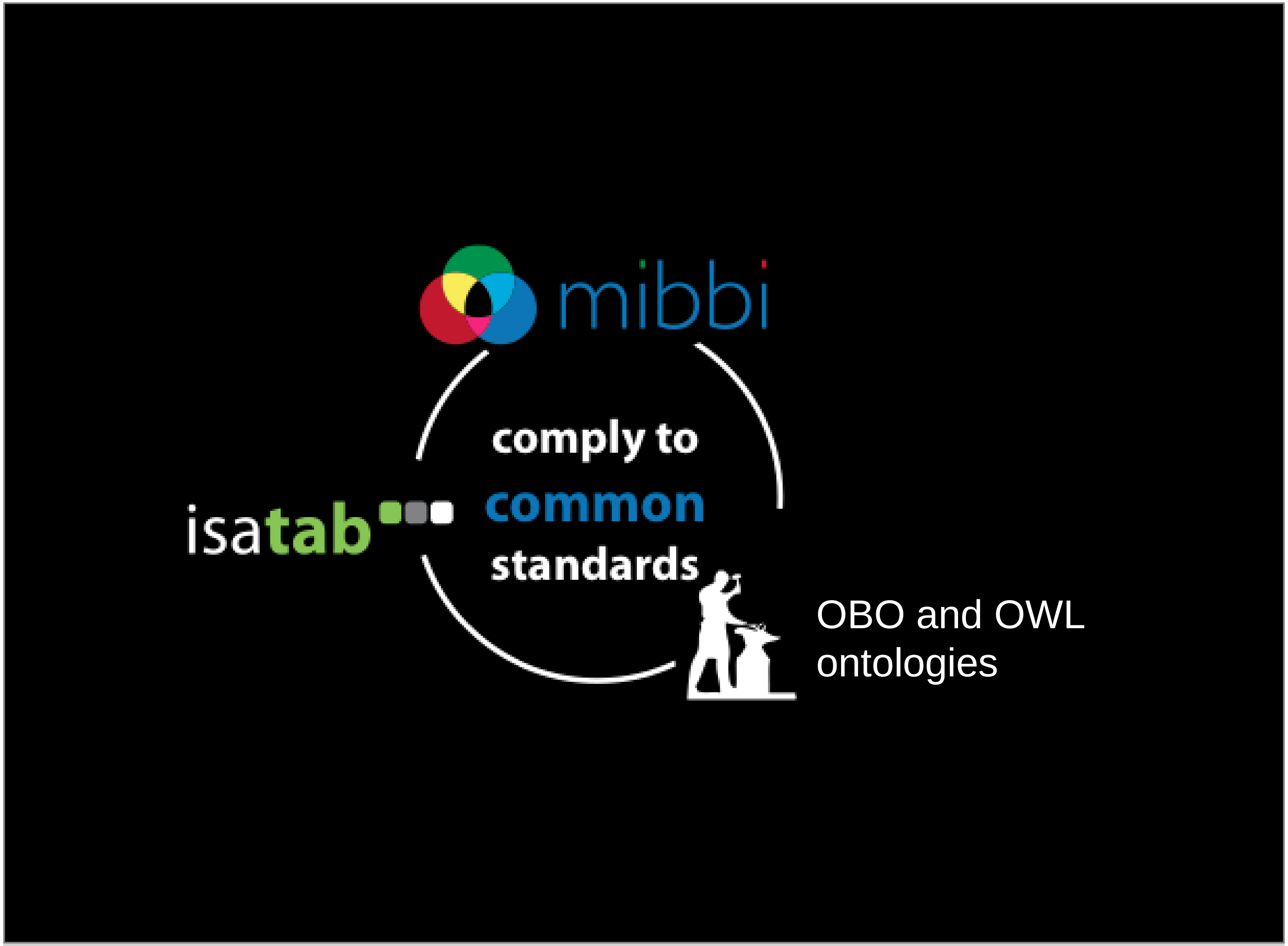
Many ontologies, Many Formats, Many Requirements…

Credits: http://liverpoolsolfed.wordpress.com/resources/image-bank/demonstration/

ISA framework overview



A focus on standards...



OBO and OWL ontologies

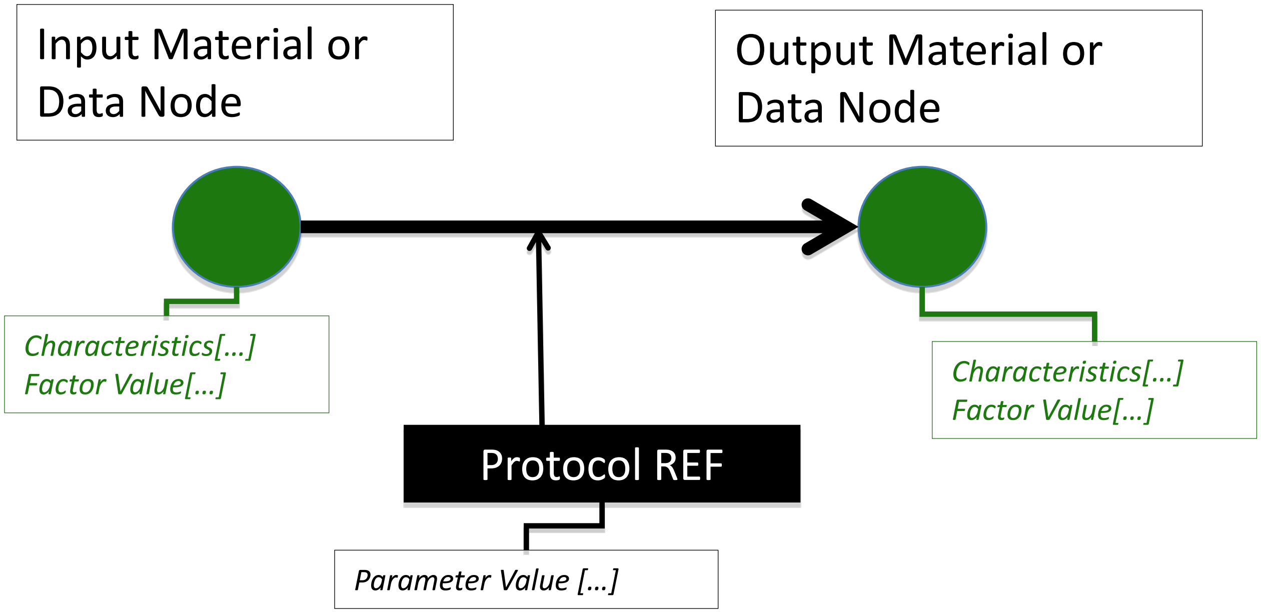
Presentation Outline

- Background information
- Rationale for developing ISA-tools
- ISA paradigm and interaction with ontologies
- Moving on: ISA future developments

ISA syntax and Table definition

- Configuration files broadly define Material /Data workflows

metaboliteprofiling_ms



Input Material or Data Node

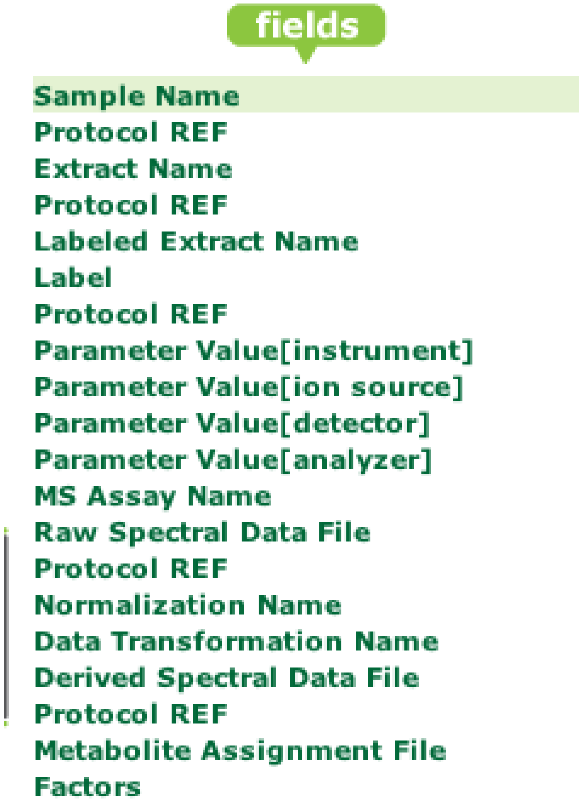
Output Material or Data Node

Characteristics[…]
Factor Value[…]

Characteristics[…]
Factor Value[…]

Protocol REF

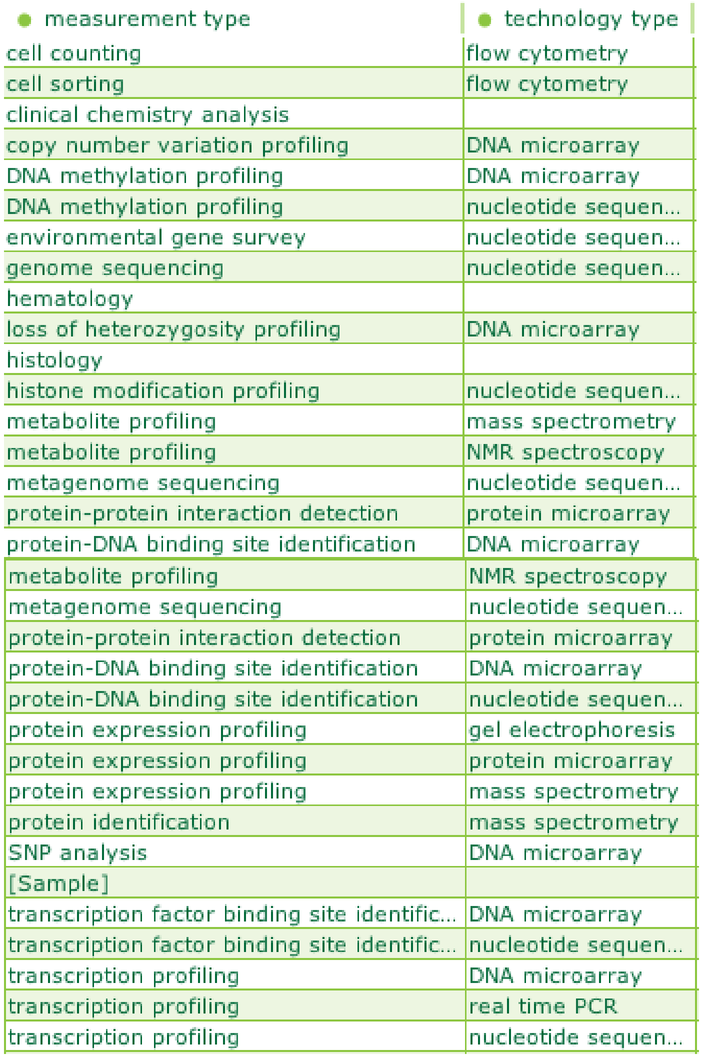
Parameter Value […]



fields

Sample Name
Protocol REF
Extract Name
Protocol REF
Labeled Extract Name
Label
Protocol REF
Parameter Value[instrument]
Parameter Value[ion source]
Parameter Value[detector]
Parameter Value[analyzer]
MS Assay Name
Raw Spectral Data File
Protocol REF
Normalization Name
Data Transformation Name
Derived Spectral Data File
Protocol REF
Metabolite Assignment File
Factors

List of supported assays in ISA default configuration



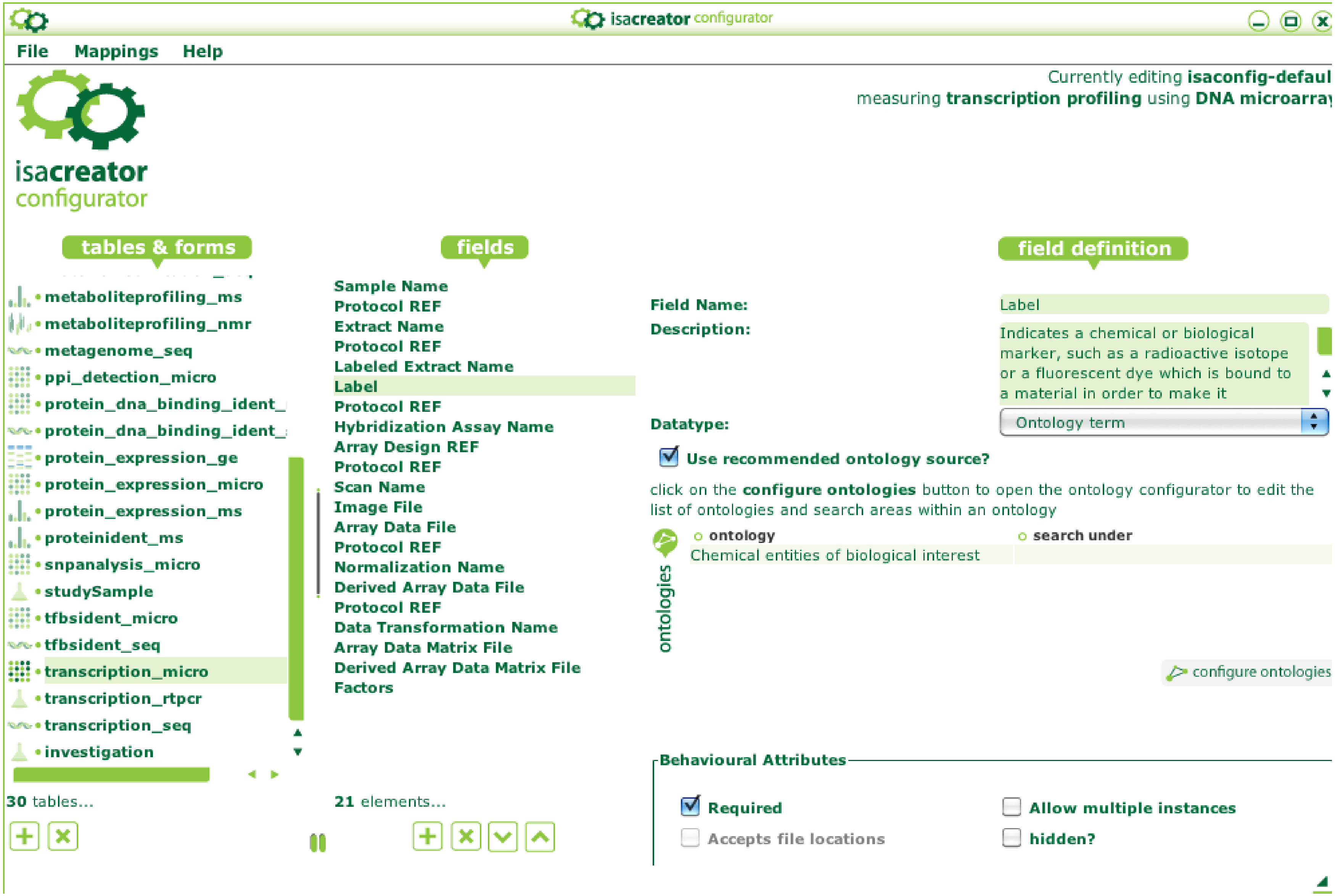
● measurement type	● technology type
cell counting	flow cytometry
cell sorting	flow cytometry
clinical chemistry analysis	
copy number variation profiling	DNA microarray
DNA methylation profiling	DNA microarray
DNA methylation profiling	nucleotide sequen...
environmental gene survey	nucleotide sequen...
genome sequencing	nucleotide sequen...
hematology	
loss of heterozygosity profiling	DNA microarray
histology	
histone modification profiling	nucleotide sequen...
metabolite profiling	mass spectrometry
metabolite profiling	NMR spectroscopy
metagenome sequencing	nucleotide sequen...
protein-protein interaction detection	protein microarray
protein-DNA binding site identification	DNA microarray
metabolite profiling	NMR spectroscopy
metagenome sequencing	nucleotide sequen...
protein-protein interaction detection	protein microarray
protein-DNA binding site identification	DNA microarray
protein-DNA binding site identification	nucleotide sequen...
protein expression profiling	gel electrophoresis
protein expression profiling	protein microarray
protein expression profiling	mass spectrometry
protein identification	mass spectrometry
SNP analysis	DNA microarray
[Sample]	
transcription factor binding site identific...	DNA microarray
transcription factor binding site identific...	nucleotide sequen...
transcription profiling	DNA microarray
transcription profiling	real time PCR
transcription profiling	nucleotide sequen...

Potential for Compliance with:
MIGS
MIMARKS
MIAME

Expanding the number of ISAconfigurations

Available from:
https://github.com/ISA-tools/Configuration-Files

ISAconfigurator Tables



Currently editing **isaconfig-defaul**
measuring **transcription profiling** using **DNA microarray**

tables & forms

- metaboliteprofiling_ms
- metaboliteprofiling_nmr
- metagenome_seq
- ppi_detection_micro
- protein_dna_binding_ident_
- protein_dna_binding_ident_
- protein_expression_ge
- protein_expression_micro
- protein_expression_ms
- proteinident_ms
- snpanalysis_micro
- studySample
- tfbsident_micro
- tfbsident_seq
- transcription_micro
- transcription_rtpcr
- transcription_seq
- investigation

30 tables...

fields

Sample Name
Protocol REF
Extract Name
Protocol REF
Labeled Extract Name
Label
Protocol REF
Hybridization Assay Name
Array Design REF
Protocol REF
Scan Name
Image File
Array Data File
Protocol REF
Normalization Name
Derived Array Data File
Protocol REF
Data Transformation Name
Array Data Matrix File
Derived Array Data Matrix File
Factors

21 elements...

field definition

Field Name: Label

Description: Indicates a chemical or biological marker, such as a radioactive isotope or a fluorescent dye which is bound to a material in order to make it

Datatype: Ontology term

☑ **Use recommended ontology source?**

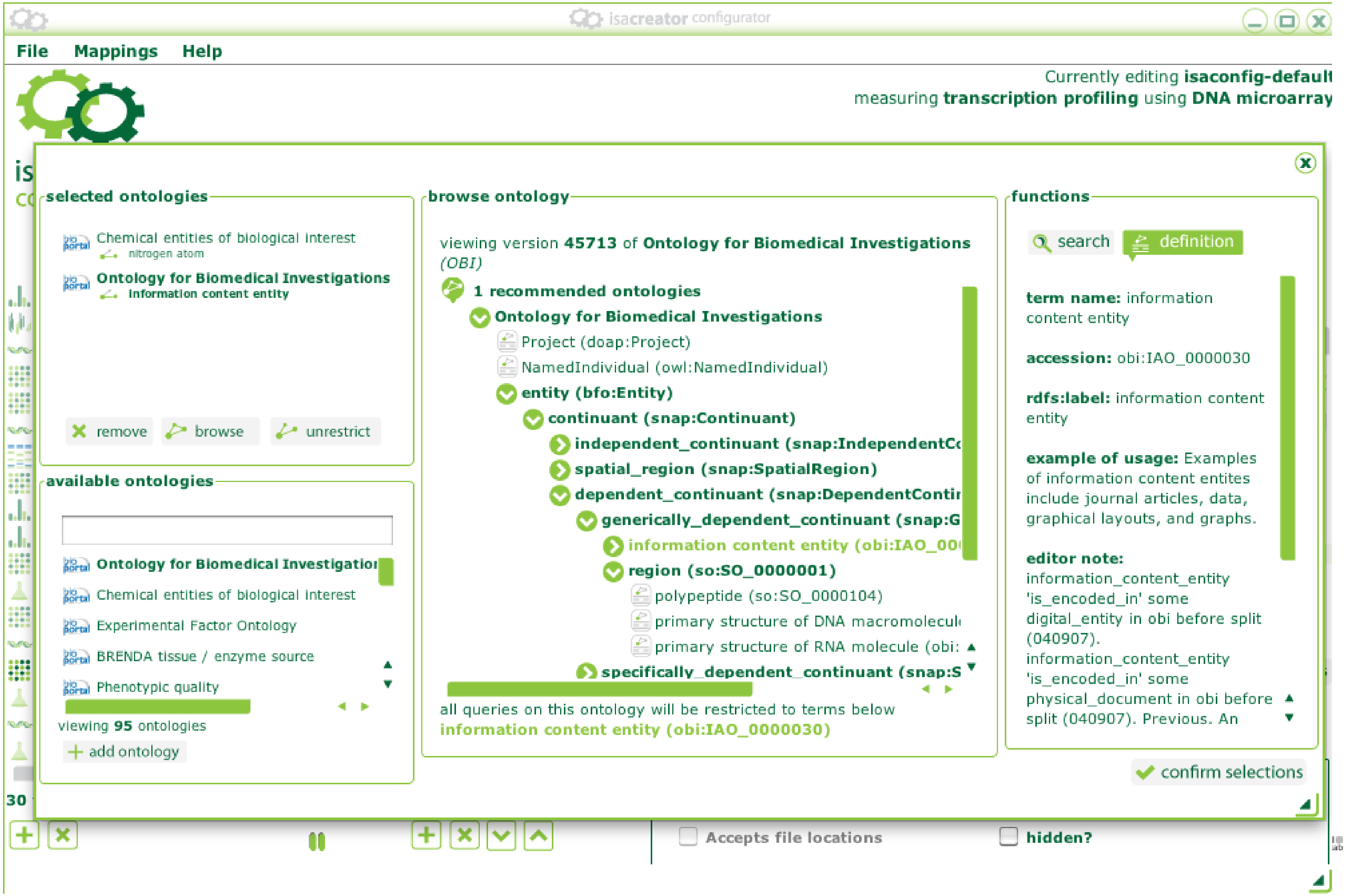
click on the **configure ontologies** button to open the ontology configurator to edit the list of ontologies and search areas within an ontology

○ ontology	○ search under
Chemical entities of biological interest	

Behavioural Attributes

☑ **Required**	□ **Allow multiple instances**
□ Accepts file locations	□ **hidden?**

ISAconfigurator Tables



Currently editing **isaconfig-default**
measuring **transcription profiling** using **DNA microarray**

selected ontologies

- Chemical entities of biological interest
  - nitrogen atom
- **Ontology for Biomedical Investigations**
  - **information content entity**

available ontologies

- **Ontology for Biomedical Investigatio**
- Chemical entities of biological interest
- Experimental Factor Ontology
- BRENDA tissue / enzyme source
- Phenotypic quality

viewing **95** ontologies

browse ontology

viewing version **45713** of **Ontology for Biomedical Investigations** *(OBI)*

1 recommended ontologies

- **Ontology for Biomedical Investigations**
  - Project (doap:Project)
  - NamedIndividual (owl:NamedIndividual)
  - **entity (bfo:Entity)**
    - **continuant (snap:Continuant)**
      - **independent_continuant (snap:IndependentC**
      - **spatial_region (snap:SpatialRegion)**
      - **dependent_continuant (snap:DependentContir**
        - **generically_dependent_continuant (snap:G**
          - **information content entity (obi:IAO_00**
          - **region (so:SO_0000001)**
            - polypeptide (so:SO_0000104)
            - primary structure of DNA macromolecul
            - primary structure of RNA molecule (obi:
        - **specifically_dependent_continuant (snap:S**

all queries on this ontology will be restricted to terms below
information content entity (obi:IAO_0000030)

functions

term name: information content entity

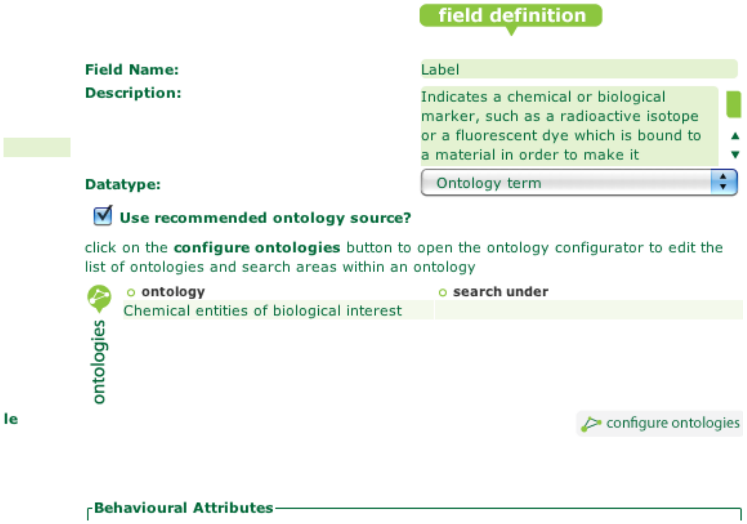
accession: obi:IAO_0000030

rdfs:label: information content entity

example of usage: Examples of information content entites include journal articles, data, graphical layouts, and graphs.

editor note: information_content_entity 'is_encoded_in' some digital_entity in obi before split (040907). information_content_entity 'is_encoded_in' some physical_document in obi before split (040907). Previous. An

ISAconfigurator Tables



field definition

Field Name: Label

Description: Indicates a chemical or biological marker, such as a radioactive isotope or a fluorescent dye which is bound to a material in order to make it

Datatype: Ontology term

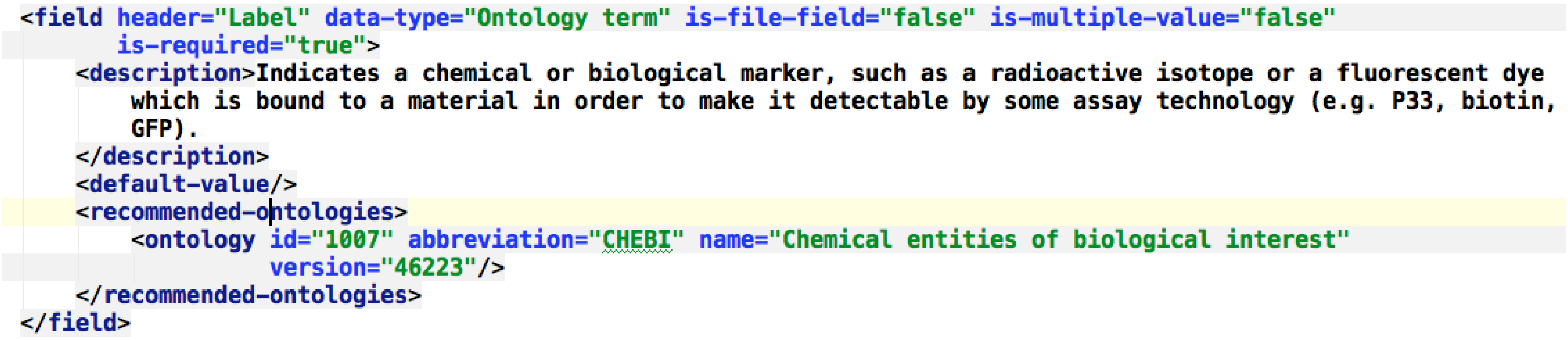
☑ Use recommended ontology source?

click on the **configure ontologies** button to open the ontology configurator to edit the list of ontologies and search areas within an ontology

○ ontology	○ search under
Chemical entities of biological interest	

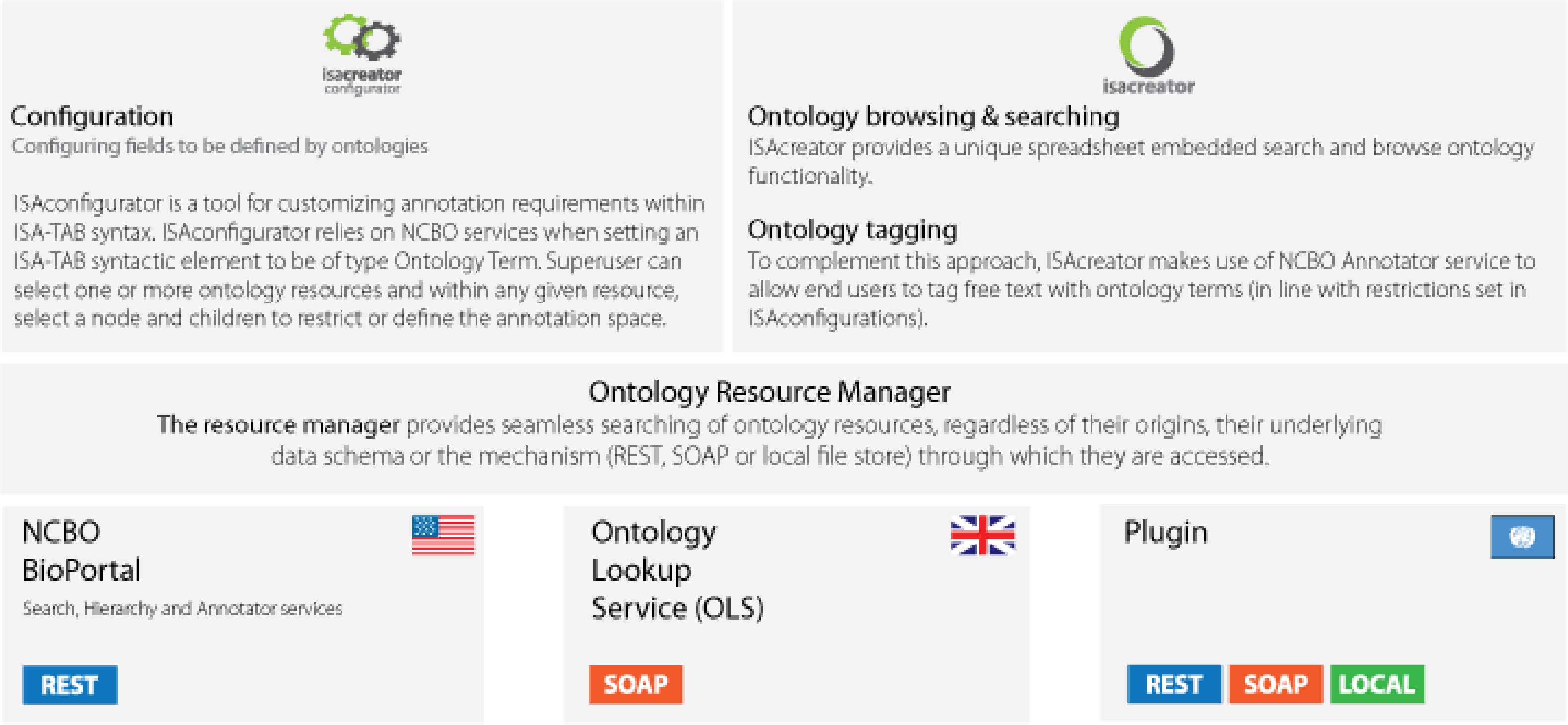
This is an example of a field definition created by the configurator. In this instance we are describing a label field, in particular, one used to describe the label used in a microarray experiment.

We have defined it to come from an ontology, and we recommend the ChEBI ontology. It is also required.



```
<field header="Label" data-type="Ontology term" is-file-field="false" is-multiple-value="false"
       is-required="true">
    <description>Indicates a chemical or biological marker, such as a radioactive isotope or a fluorescent dye
        which is bound to a material in order to make it detectable by some assay technology (e.g. P33, biotin,
        GFP).
    </description>
    <default-value/>
    <recommended-ontologies>
        <ontology id="1007" abbreviation="CHEBI" name="Chemical entities of biological interest"
                  version="46223"/>
    </recommended-ontologies>
</field>
```

How do ISA tools access Ontology servers?



Configuration

Configuring fields to be defined by ontologies

ISAconfigurator is a tool for customizing annotation requirements within ISA-TAB syntax. ISAconfigurator relies on NCBO services when setting an ISA-TAB syntactic element to be of type Ontology Term. Superuser can select one or more ontology resources and within any given resource, select a node and children to restrict or define the annotation space.

Ontology browsing & searching

ISAcreator provides a unique spreadsheet embedded search and browse ontology functionality.

Ontology tagging

To complement this approach, ISAcreator makes use of NCBO Annotator service to allow end users to tag free text with ontology terms (in line with restrictions set in ISAconfigurations).

Ontology Resource Manager

The resource manager provides seamless searching of ontology resources, regardless of their origins, their underlying data schema or the mechanism (REST, SOAP or local file store) through which they are accessed.

NCBO BioPortal

Search, Hierarchy and Annotator services

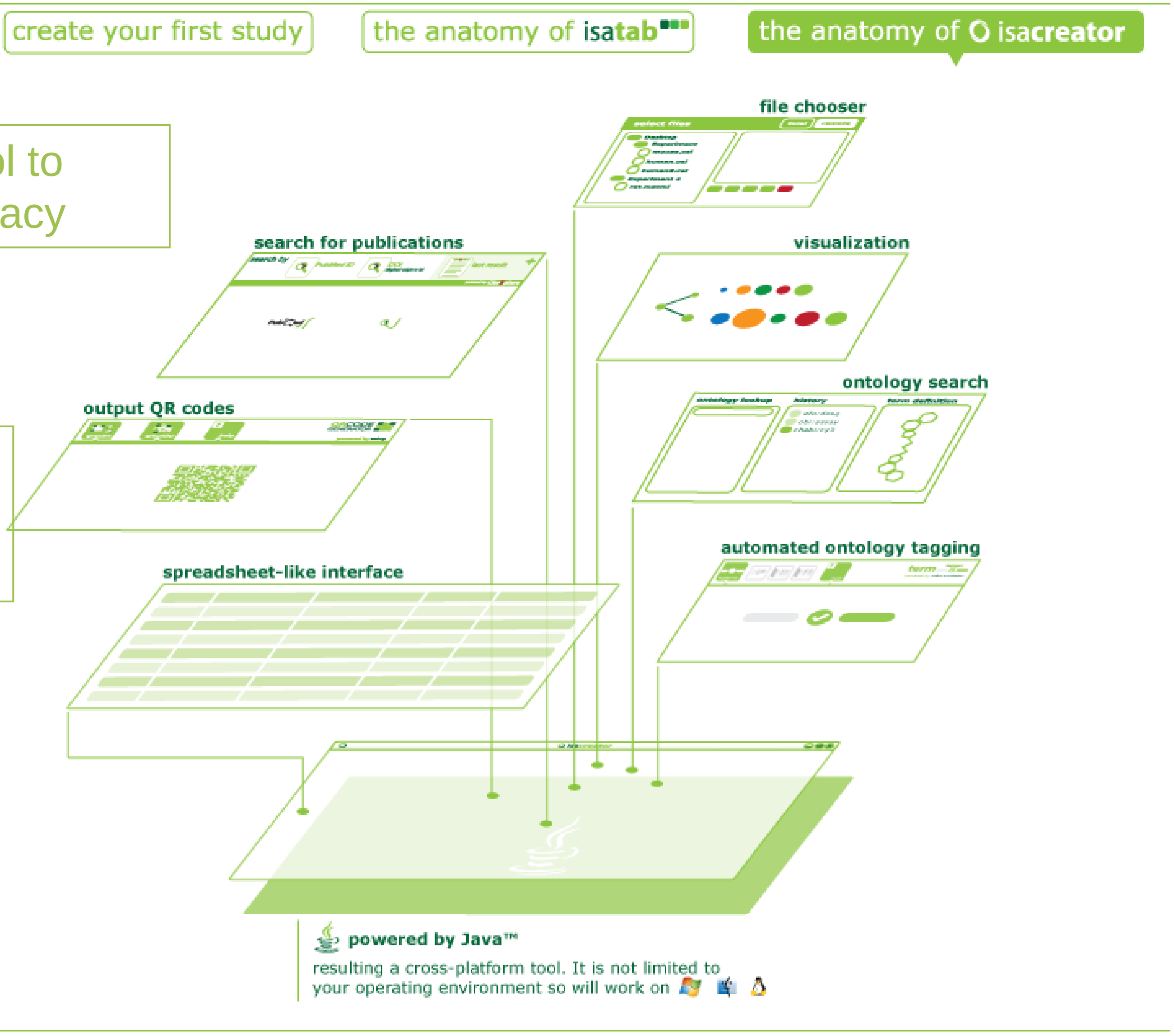
Ontology Lookup Service (OLS)

Plugin

Anatomy of ISAcreator

Mapping Tool to deal with legacy data

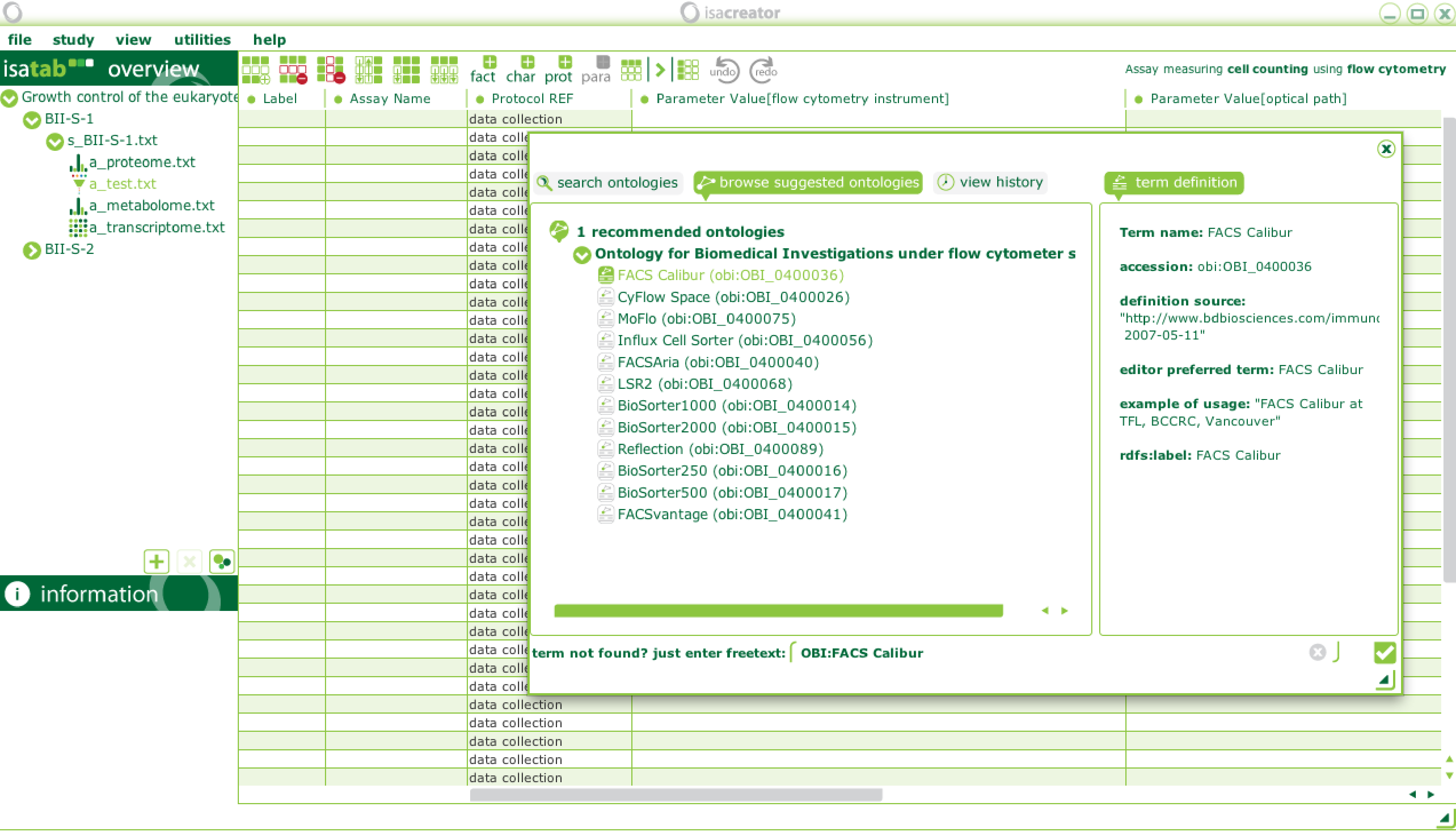
Experiment design wizard for templating



powered by Java™

resulting a cross-platform tool. It is not limited to your operating environment so will work on

Select and Annotate in ISAcreator



Assay measuring **cell counting** using **flow cytometry**

Growth control of the eukaryote
- BII-S-1
  - s_BII-S-1.txt
    - a_proteome.txt
    - a_test.txt
    - a_metabolome.txt
    - a_transcriptome.txt
- BII-S-2

Label	Assay Name	Protocol REF	Parameter Value[flow cytometry instrument]	Parameter Value[optical path]
		data collection		
		data colle		
		data collection		
		data collection		
		data collection		
		data collection		
		data collection		

1 recommended ontologies

Ontology for Biomedical Investigations under flow cytometer s

- FACS Calibur (obi:OBI_0400036)
- CyFlow Space (obi:OBI_0400026)
- MoFlo (obi:OBI_0400075)
- Influx Cell Sorter (obi:OBI_0400056)
- FACSAria (obi:OBI_0400040)
- LSR2 (obi:OBI_0400068)
- BioSorter1000 (obi:OBI_0400014)
- BioSorter2000 (obi:OBI_0400015)
- Reflection (obi:OBI_0400089)
- BioSorter250 (obi:OBI_0400016)
- BioSorter500 (obi:OBI_0400017)
- FACSvantage (obi:OBI_0400041)

Term name: FACS Calibur

accession: obi:OBI_0400036

definition source: "http://www.bdbiosciences.com/immun
2007-05-11"

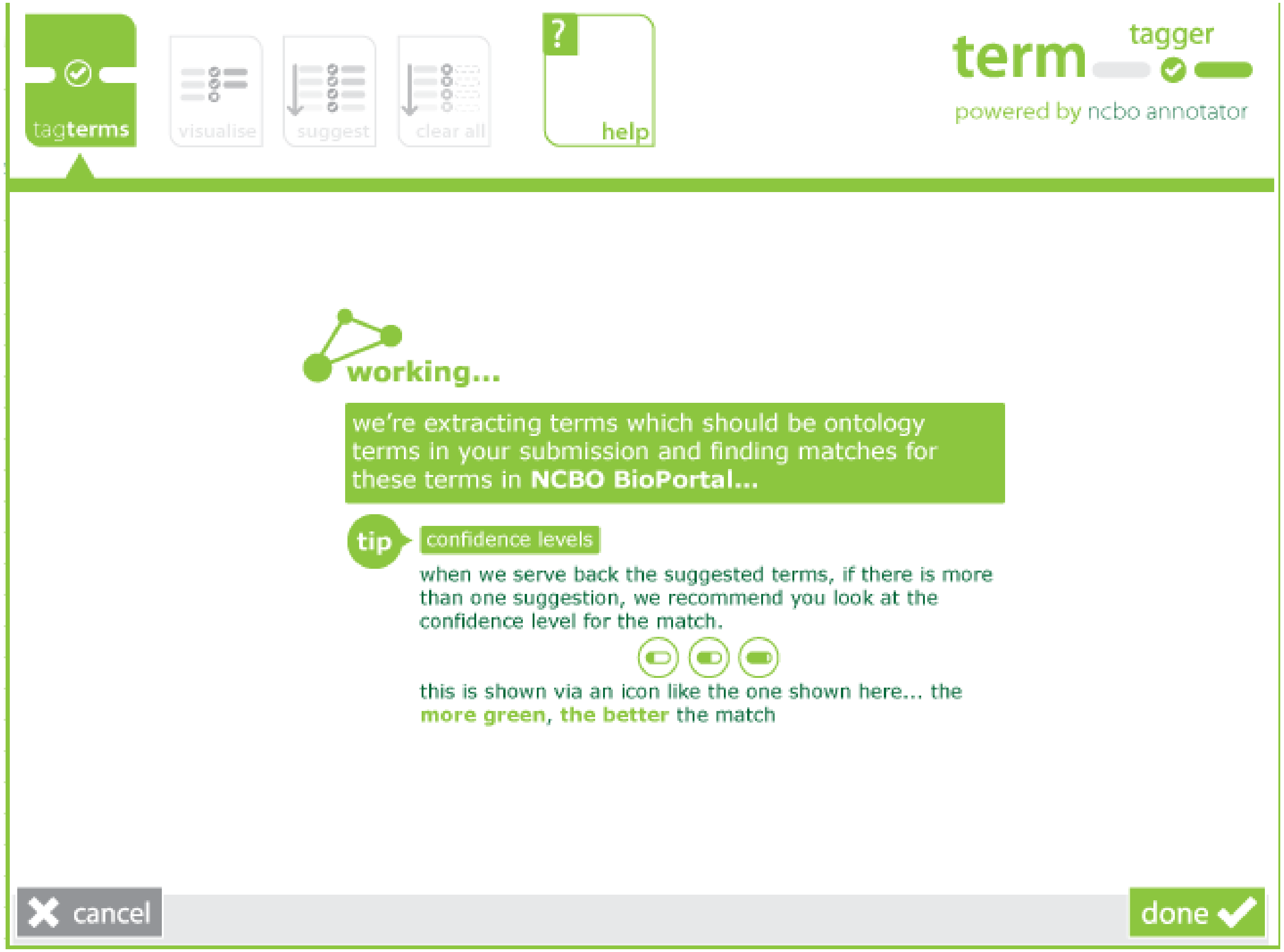
editor preferred term: FACS Calibur

example of usage: "FACS Calibur at TFL, BCCRC, Vancouver"

rdfs:label: FACS Calibur

term not found? just enter freetext: **OBI:FACS Calibur**

Calling NCBO Annotator service from ISAcreator: Tagging free text



working...

we're extracting terms which should be ontology terms in your submission and finding matches for these terms in **NCBO BioPortal...**

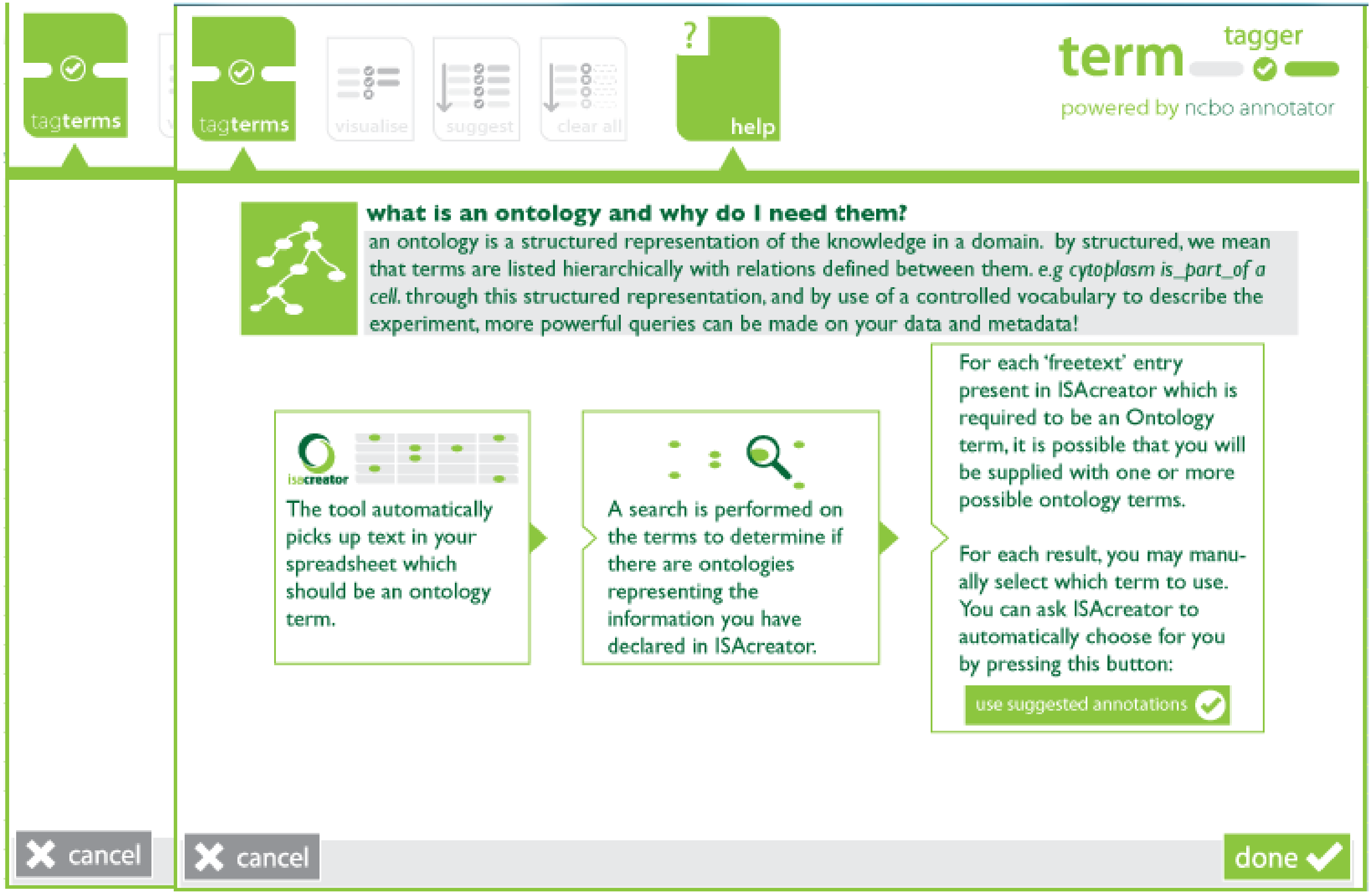
tip

confidence levels

when we serve back the suggested terms, if there is more than one suggestion, we recommend you look at the confidence level for the match.

this is shown via an icon like the one shown here... the **more green**, **the better** the match

Calling NCBO Annotator service from ISAcreator: Tagging free text



what is an ontology and why do I need them?

an ontology is a structured representation of the knowledge in a domain. by structured, we mean that terms are listed hierarchically with relations defined between them. e.g *cytoplasm is_part_of a cell.* through this structured representation, and by use of a controlled vocabulary to describe the experiment, more powerful queries can be made on your data and metadata!

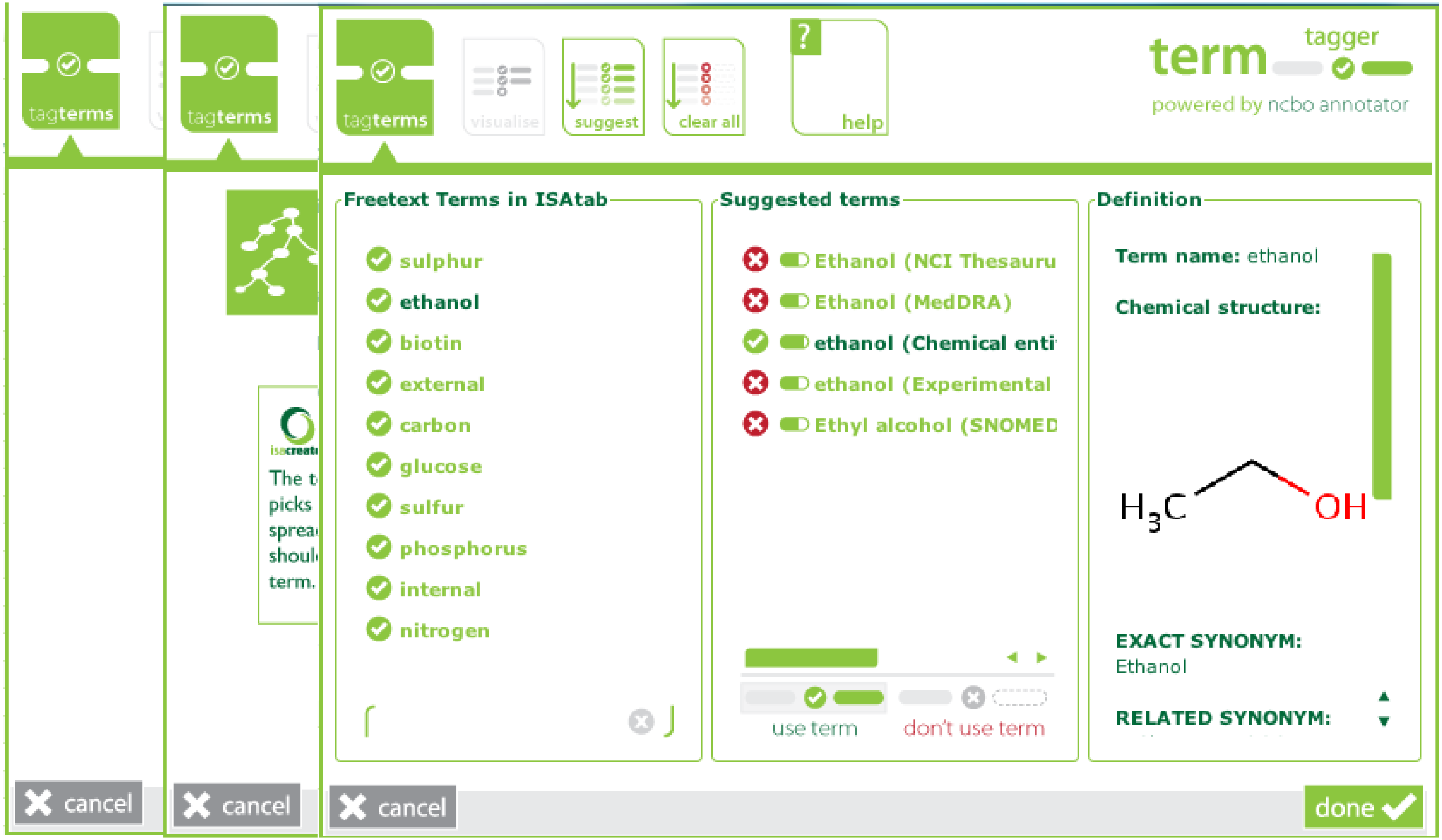
The tool automatically picks up text in your spreadsheet which should be an ontology term.

A search is performed on the terms to determine if there are ontologies representing the information you have declared in ISAcreator.

For each 'freetext' entry present in ISAcreator which is required to be an Ontology term, it is possible that you will be supplied with one or more possible ontology terms.

For each result, you may manually select which term to use. You can ask ISAcreator to automatically choose for you by pressing this button:

Calling NCBO Annotator service from ISAcreator: Tagging free text



Freetext Terms in ISAtab

- ✓ sulphur
- ✓ ethanol
- ✓ biotin
- ✓ external
- ✓ carbon
- ✓ glucose
- ✓ sulfur
- ✓ phosphorus
- ✓ internal
- ✓ nitrogen

Suggested terms

- ☒ Ethanol (NCI Thesauru
- ☒ Ethanol (MedDRA)
- ✓ ethanol (Chemical enti
- ☒ ethanol (Experimental
- ☒ Ethyl alcohol (SNOMED

use term don't use term

Definition

Term name: ethanol

Chemical structure:

EXACT SYNONYM:
Ethanol

RELATED SYNONYM:

Presentation Outline

- Background information
- Rationale for developing ISA-tools
- ISA paradigm and interaction with ontologies
- Moving on: ISA future developments

Expand ISA community and welcome new members

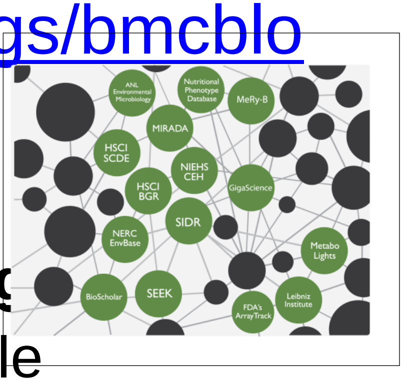
- Metabolights, EBI’s metabolomic data repository
- BBSRC funding for a Metagenomic Portal (EBI collaboration)
- Expand ISAconverter to support additional XML formats
  - GEO MiniML (ongoing development)
  - FuGE-ML (carried out at INIST [Magali Roux et al])
  - SRA XML maintenance (regular schema updates)
- Scout for new domains of application

Exposing Experimental Metadata on the semantic web

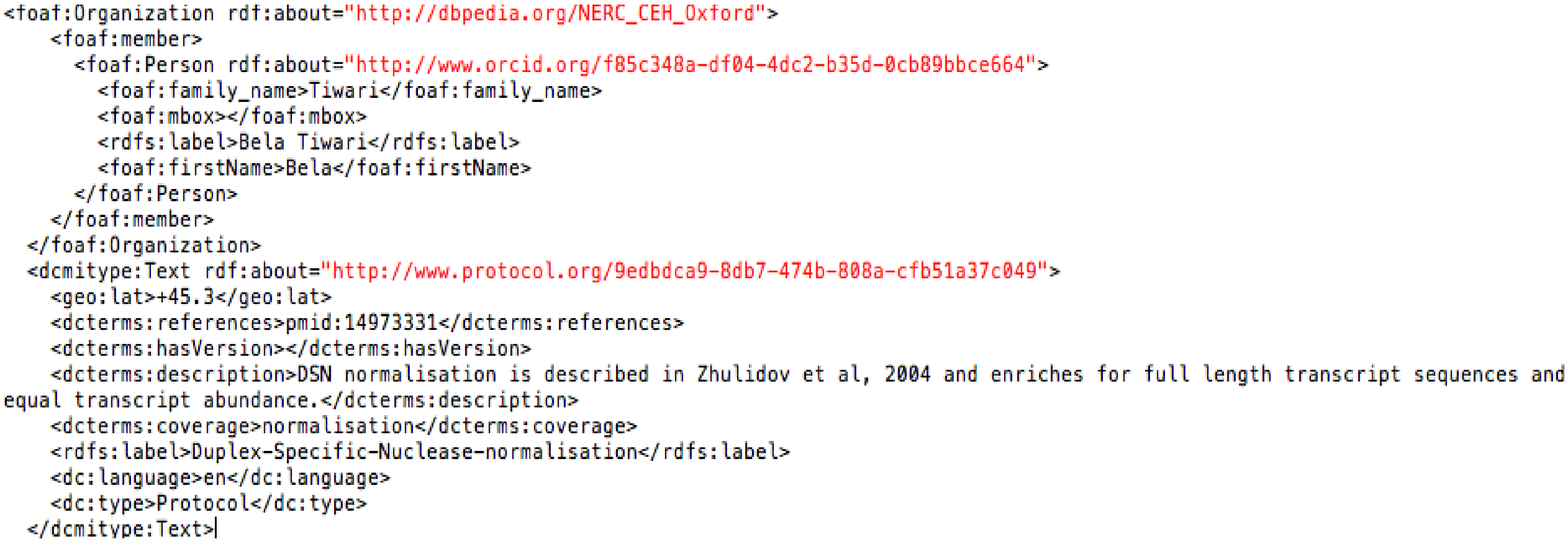
- Expansion of **ISAconverter** to provide RDF/OWL representation of experimental data
- Ontologies or Vocabularies matter again.
- Mapping of ISA elements to resources such as: OBI ontology classes, FOAF, Dublin Core.
- Expand ISAconfigurator to enable recording of Mapping to Vocabulary

ISA2RDF work in progress

- Use case on W3C HCLS scientific discourse list
  - deciding on the granularity of representation
  - building on previous experience
  - Evaluating alternative representations.
- Participitation to the Biohackathon 2011
  - http://blogs.openaccesscentral.com/blogs/bmcblog/entry/biohackathon_2011_number_1
  - Discussing best practices
    - Use of URI supplied by **www.identifiers.org**
    - Avoid use of blank node as much as possible



ISA2RDF: work in progress

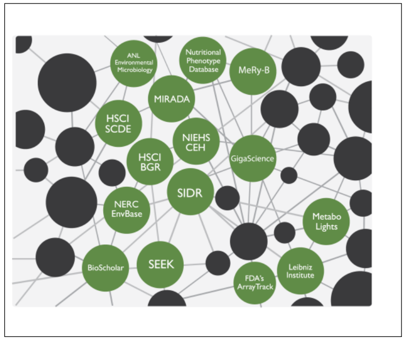


```
<foaf:Organization rdf:about="http://dbpedia.org/NERC_CEH_Oxford">
    <foaf:member>
      <foaf:Person rdf:about="http://www.orcid.org/f85c348a-df04-4dc2-b35d-0cb89bbce664">
        <foaf:family_name>Tiwari</foaf:family_name>
        <foaf:mbox></foaf:mbox>
        <rdfs:label>Bela Tiwari</rdfs:label>
        <foaf:firstName>Bela</foaf:firstName>
      </foaf:Person>
    </foaf:member>
  </foaf:Organization>
  <dcmitype:Text rdf:about="http://www.protocol.org/9edbdca9-8db7-474b-808a-cfb51a37c049">
    <geo:lat>+45.3</geo:lat>
    <dcterms:references>pmid:14973331</dcterms:references>
    <dcterms:hasVersion></dcterms:hasVersion>
    <dcterms:description>DSN normalisation is described in Zhulidov et al, 2004 and enriches for full length transcript sequences and
equal transcript abundance.</dcterms:description>
    <dcterms:coverage>normalisation</dcterms:coverage>
    <rdfs:label>Duplex-Specific-Nuclease-normalisation</rdfs:label>
    <dc:language>en</dc:language>
    <dc:type>Protocol</dc:type>
  </dcmitype:Text>
```

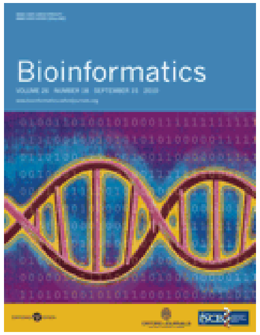
rdf:subject - rdf:predicate - rdf:object:

<lipoprotein>-<affects><inflammatory_cell>

<PRO:212342352>-<RO:543636><CL:84872762>



Publication...



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

Philippe Rocca-Serra; Marco Brandizi; Eamonn Maguire; Nataliya Sklyar; Chris Taylor; Kimberly Begley; Dawn Field; Stephen Harris; Winston Hide; Oliver Hofmann; Steffen Neumann; Peter Sterk; Weida Tong; Susanna-Assunta Sansone

Bioinformatics 2010 26: 2354-2356

MERCI!

Groups and individuals participating in:

MIBBI http://mibbi.org

ISA-Tab format http://isatab.sf.net

OBO Foundry http://obofoundry.org

OBI: http://obi-ontology.org/page/Main_Page

ISA Infrastructure Team:

Eamonn Maguire (Oxford)
Philippe Rocca-Serra (Oxford)
Susanna-Assunta Sansone (Oxford)
Chris Taylor (EMBL-EBI)

Alumni:

Marco Brandizi (EMBL-EBI)
Nataliya Sklyar (EMBL-EBI)

collaborators at:

Cambridge University
EuNuGO
Harvard School for Public Health
FDAs NCTR
Leibniz Plant Institute
NERCs NEBC
SIDR, INIST
Metabolights, EMBL-EBI

Funders:

EU Carcinogenomics Project
UK BBSRC

MERCI!

Groups and individuals participating in:

Dawn Field: NERC CEH Oxford
Winston Hide: HSPH
Oliver Hoffman: HSPH
Shannan Ho Sui : HSPH
Brad Chapman: HSPH
Christoph Steinbeck: Metabolights
Kenneth Haug: Metabolights
Paula de Matos: Metabolights
Magali Roux: INIST
Florian Mazur: INIST
Alain Zasadzinki: INIST
Marie Christine Jacquemot: INIST
And many more who have to forgive us!

Questions: