

# A personal experience developing systems requiring search functionality

Open Source search for Bioinformatics  
3 & 4 Feb 2016, EMBL-EBI

Rafael C Jimenez

[rafael.jimenez@elixir-europe.org](mailto:rafael.jimenez@elixir-europe.org)

ELIXIR

# Lessons learnt

**Biobank data federation**

**MIABIS Connect**

*Elasticsearch*

**Life Science event distribution**

**iAnn**

*SOLR*

# MIABIS Connect

A federated lightweight data sharing solution for biobanks based on MIABIS



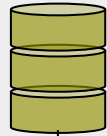
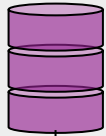
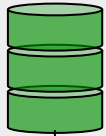
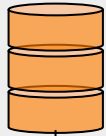
**BBMRI-ERIC**.SE



MAX-PLANCK-GESELLSCHAFT



# Problem



## Spread **data resources**

- Different data model
- Not easy to find them



## Different **query interfaces**



## Variable **results**

- Formats / schemas
- Controlled vocabularies
- Minimum information guidelines

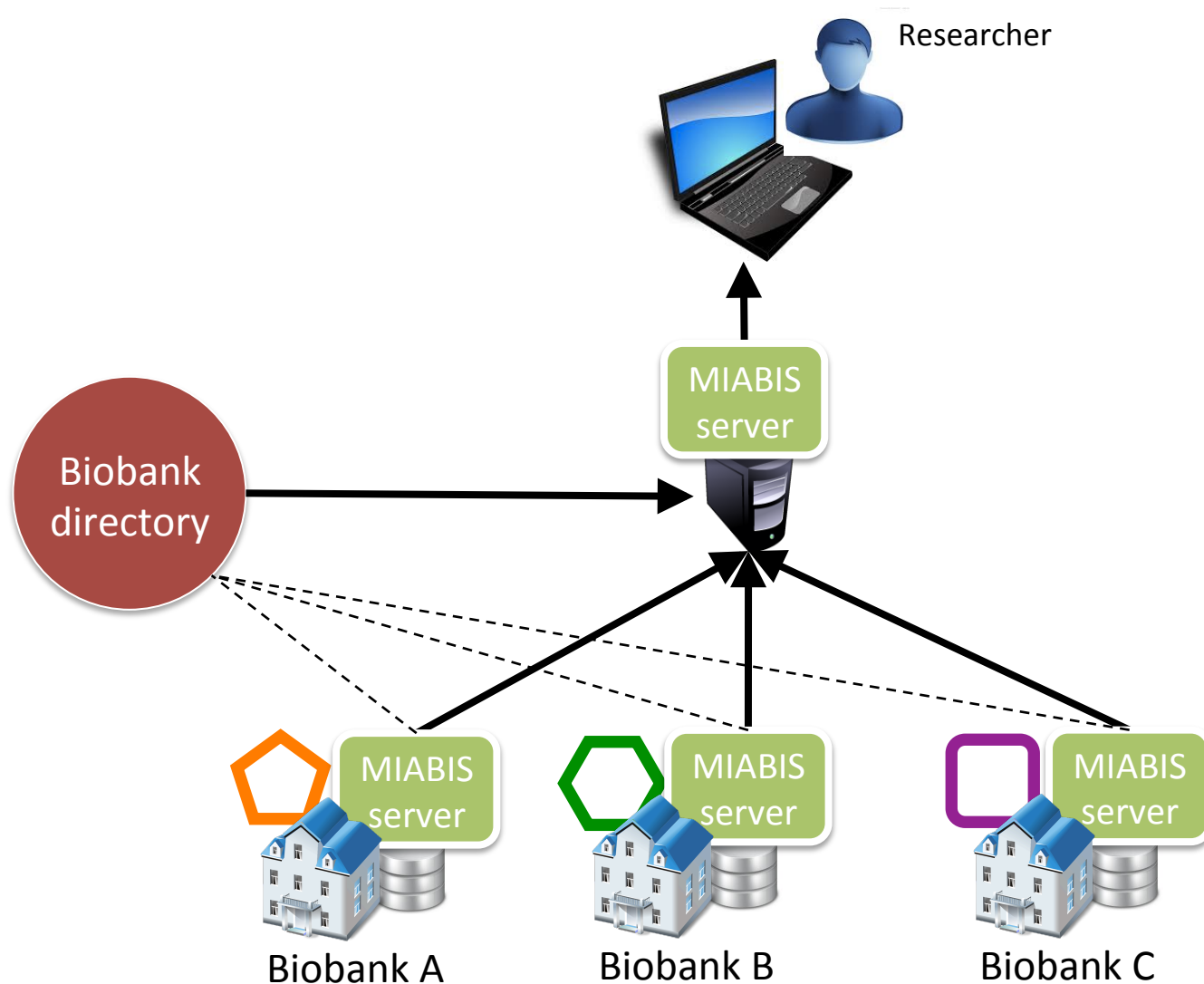


Discovery? Search? Integration?

# Requirements

- Facilitate search, integration and discovery
- Data stays at the BioBank
- Common query interfaces
- Simple adoption
- Respect existing models and interfaces
- Modular tools useful for biobanks

# Solution



# Reuse

- Format, CVs & information guidelines
  - **MIABIS** metadata standard
- Search & query interface
  - **Elastic Search**



# MIABIS

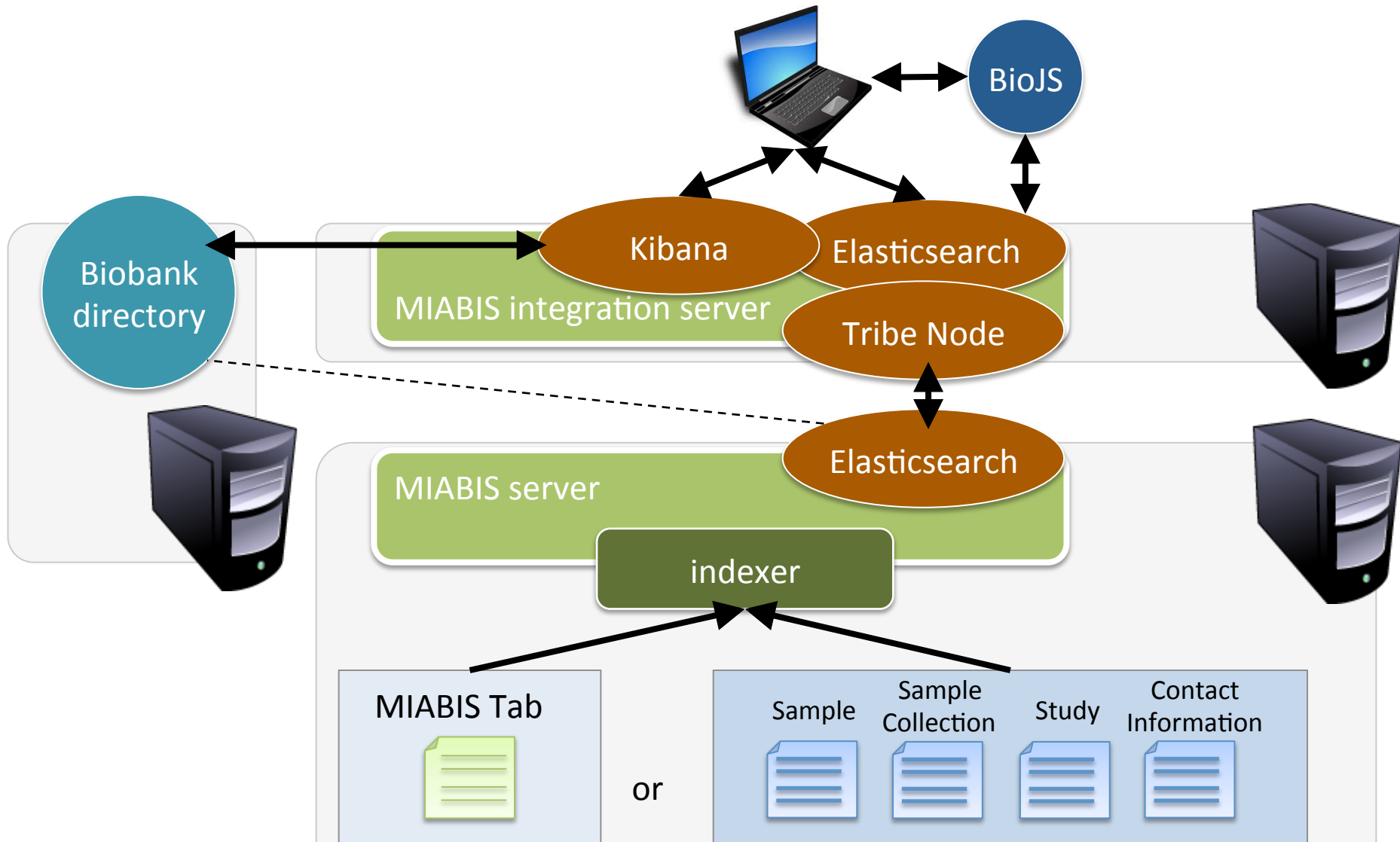
Exchange metadata format  
for biobank sample data

- **Biobank**
- **Sample Collection**
- **Study**

- **Sample**
- **Sample Quality**
- **Biological Experiment**
- **Participant**
- **Rare diseases**

Under review

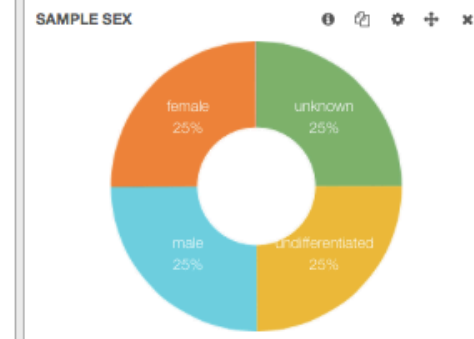
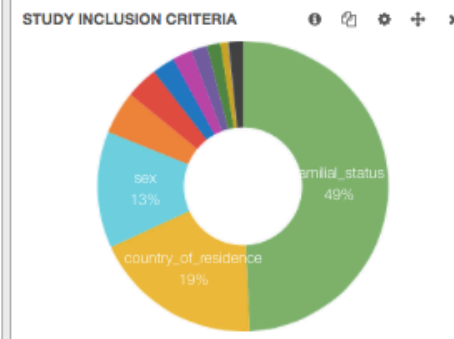
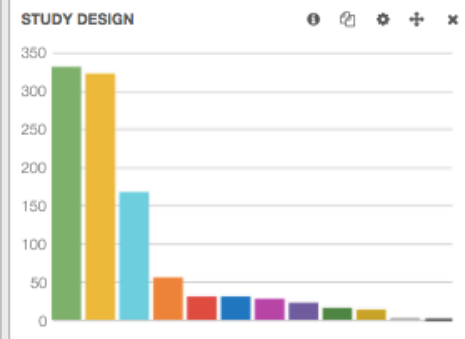
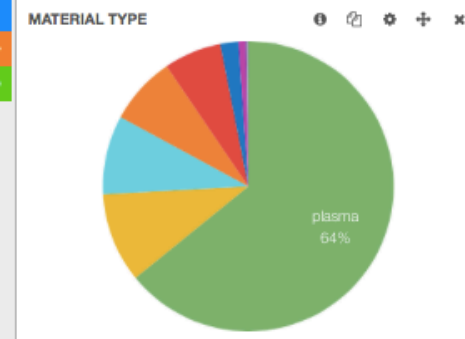
# Architecture









QUERY 






FILTERING 

Add **panels** to empty row



**DOCUMENTS**     

0 to 100 of 500 available for paging 

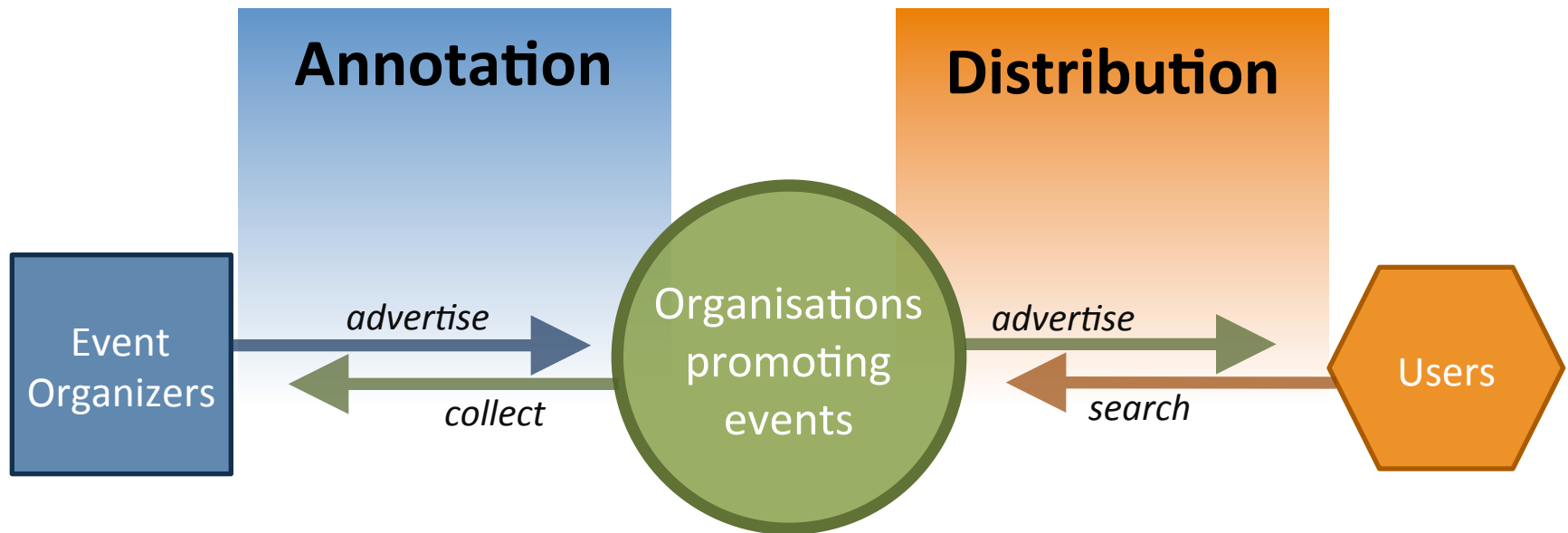
Fields 	id 	materialType 	storageTemperature 	sex 
All (1) / Current (70)	00231-00678978	PLASMA	22	UNKNOWN
Type to filter...	00231-00658146	PLASMA	13	UNDIFFERENTIATED
<input type="checkbox"/> _id	00231-00567986	PLASMA	4	MALE
<input type="checkbox"/> _index	00231-00551612	WHOLE_BLOOD	5	UNDIFFERENTIATED
<input type="checkbox"/> _type	00231-00539586	PLASMA	-26	MALE
<input type="checkbox"/> ageHigh	00231-00533812	PLASMA	23	UNKNOWN
<input type="checkbox"/> ageLow	00231-00485498	WHOLE_BLOOD	-27	MALE
<input type="checkbox"/> ageUnit	00231-00485412	WHOLE_BLOOD	16	UNKNOWN
<input type="checkbox"/> anatomicalSite.code	00231-00478614	PLASMA	6	FEMALE
<input type="checkbox"/> anatomicalSite.description	00231-00475514	PLASMA	-29	MALE
<input type="checkbox"/> anatomicalSite.ontology	00231-00456071	PLASMA	-8	FEMALE
<input type="checkbox"/> anatomicalSite.version	00231-00307510	WHOLE_BLOOD	30	UNDIFFERENTIATED
<input type="checkbox"/> biobank.acronym	00231-00289679	PLASMA	19	UNDIFFERENTIATED
<input type="checkbox"/> biobank.contactInformation.address				
<input type="checkbox"/> biobank.contactInformation.email				

<http://dachstein.biochem.mpg.de/kibana/>

# iAnn

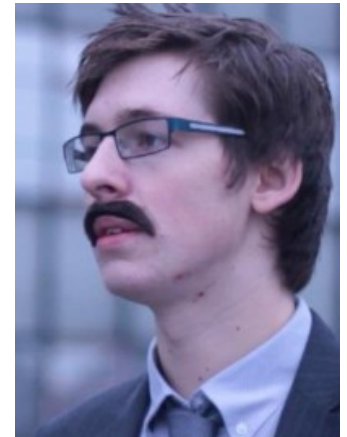
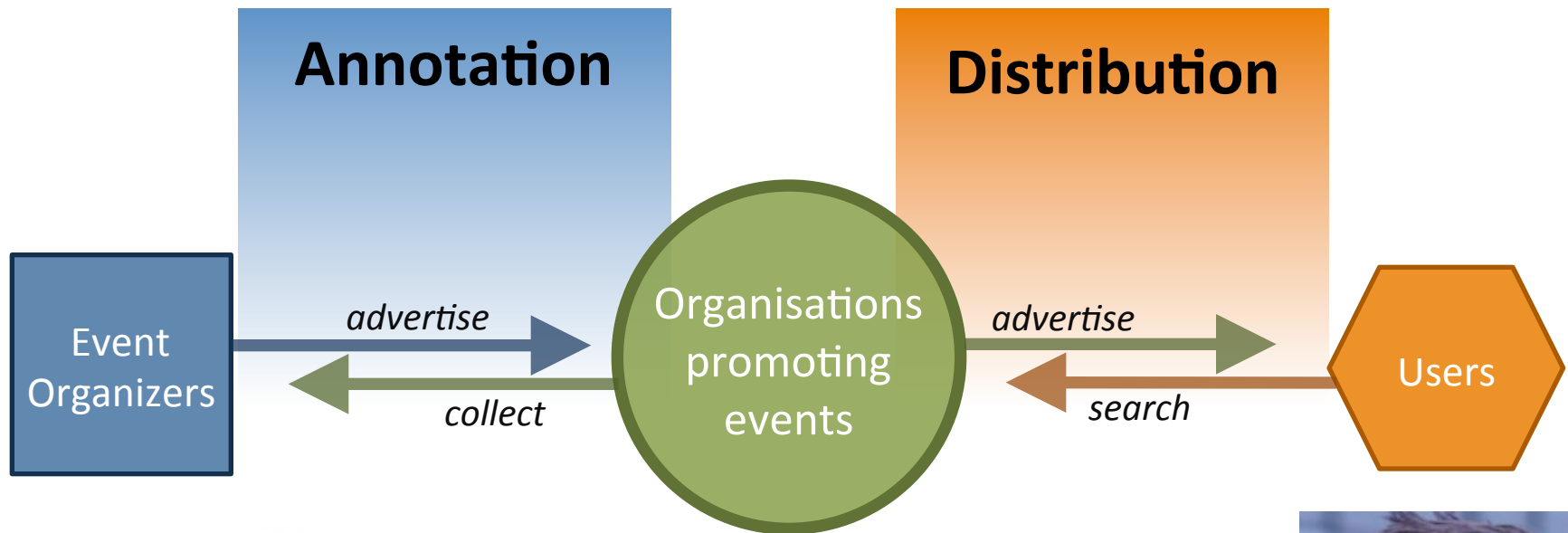
Event sharing platform  
for the life sciences

# Announcements dissemination



**Courses, workshops,** meetings, seminars,  
conferences, symposiums, webinars ...

# Announcements dissemination



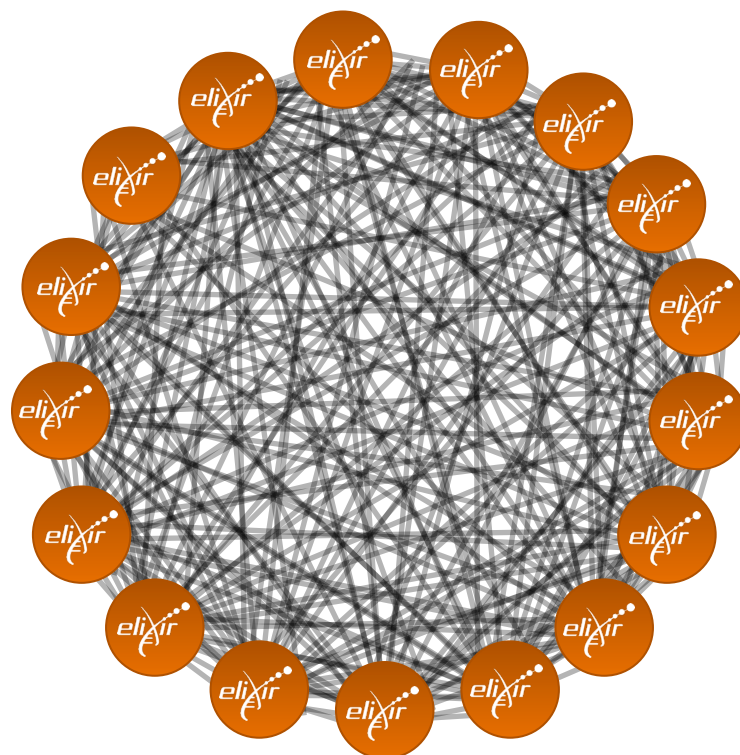
# Current dissemination

1 event per Node

Total of 306 events displayed

**289 manual copies to do**

17 Nodes



# Bioinformatics

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS

CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals > Life Sciences & Mathematics & Physical Sciences > Bioinformatics > Volume 29, Issue 15 > Pp. 1919-1921.

## iAnn: an event sharing platform for the life sciences



[« Previous](#) | [Next Article »](#)  
[Table of Contents](#)

### This Article

Bioinformatics (2013) 29 (15):  
 1919-1921.  
 doi: 10.1093/bioinformatics/btt306  
 First published online: June 5,  
 2013

This article is Open Access

» Abstract **Free**

Full Text (HTML) **Free**

Full Text (PDF) **Free**

All Versions of this Article:  
[btt306v1](#)

Rafael C. Jimenez<sup>1</sup>, Juan P. Albar<sup>2</sup>, Jong Bhak<sup>3</sup>, Marie-Claude Blatter<sup>4</sup>, Thomas Blicher<sup>5</sup>,  
 Michelle D. Brazas<sup>6</sup>, Cath Brooksbank<sup>1</sup>, Aidan Budd<sup>7</sup>, Javier De Las Rivas<sup>8</sup>,  
 Jacqueline Dreyer<sup>7</sup>, Marc A. van Driel<sup>9</sup>, Michael J. Dunn<sup>10</sup>, Pedro L. Fernandes<sup>11</sup>,  
 Celia W. G. van Gelder<sup>9</sup>, Henning Hermjakob<sup>1</sup>, Vassilios Ioannidis<sup>12</sup>, David P. Judge<sup>13</sup>,  
 Pascal Kahlem<sup>1</sup>, Eija Korpelainen<sup>14</sup>, Hans-Joachim Kraus<sup>15</sup>, Jane Loveland<sup>16</sup>,  
 Christine Mayer<sup>15</sup>, Jennifer McDowall<sup>1</sup>, Federico Moran<sup>17</sup>, Nicola Mulder<sup>18</sup>,  
 Tommi Nyronen<sup>14</sup>, Kristian Rother<sup>19</sup>, Gustavo A. Salazar<sup>18</sup>, Reinhard Schneider<sup>20</sup>,  
 Allegra Via<sup>21</sup>, Jose M. Villaveces<sup>22</sup>, Ping Yu<sup>23</sup>, Maria V. Schneider<sup>24</sup>, Teresa K. Attwood<sup>25</sup>  
 and Manuel Corpas<sup>24,\*</sup>

+ Author Affiliations

... an event should be curated only once



# iAnn

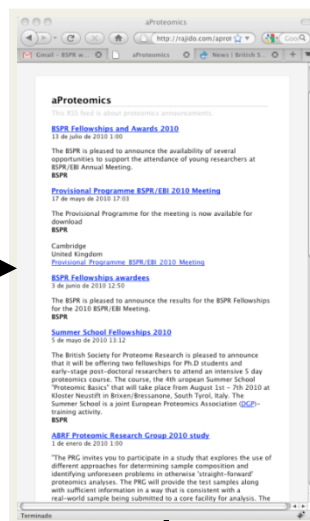
## iAnn Editor

The screenshot shows the iAnn Editor interface. It includes a 'Login' section with a password field. Below is an 'EVENT' section with fields for 'Title', 'Subtitle', 'Link', 'Date', 'Location' (Venue, City, County, Postcode, Country), 'Attachment' (Optional name, File), and 'Image/Runner' (File). At the bottom is an 'ANNOUNCEMENT' section with 'Classification' (Category, Subcategory) and 'Terminology'.

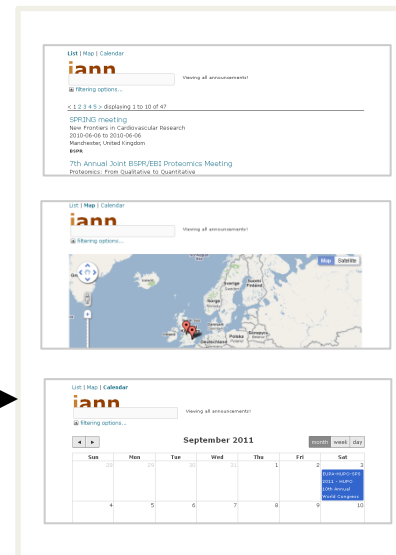
## iAnn Registry



## iAnn Web Service



## iAnn Visualization



AJAX-SOLR

## partners Integration



TeSS

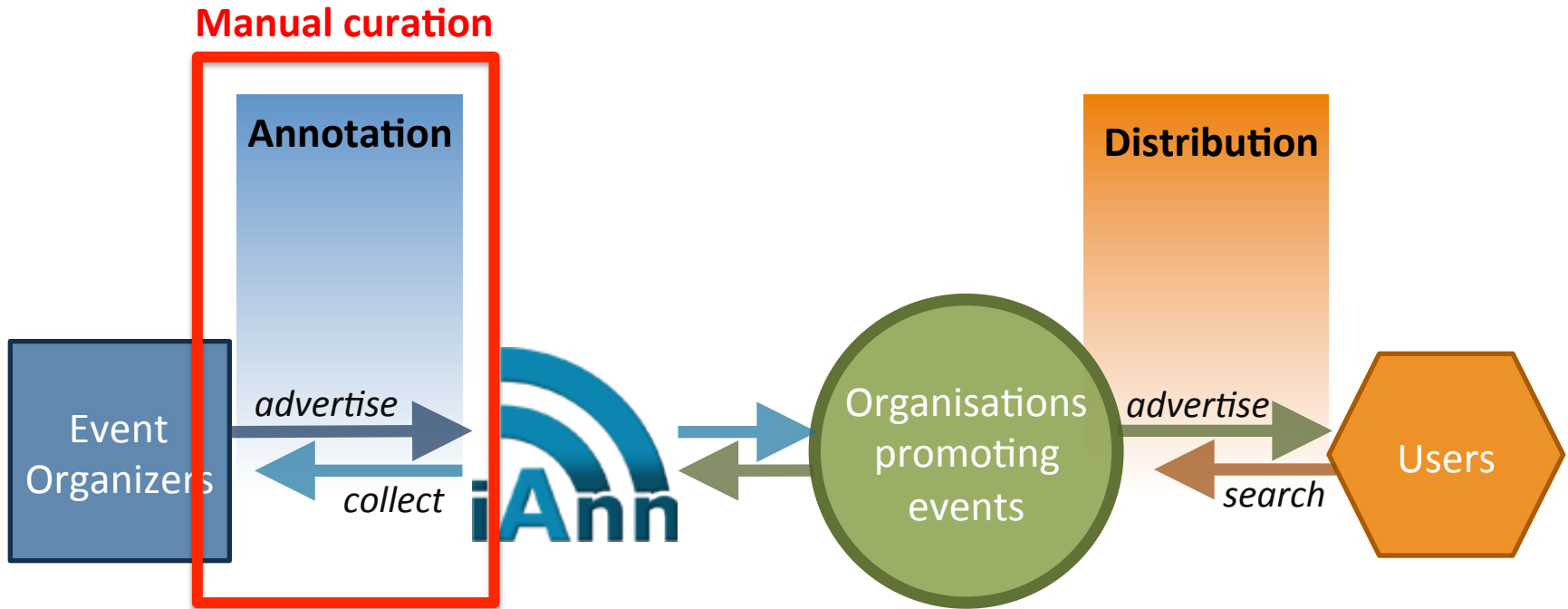
# Partial solution



**Automated dissemination**

focus on events

# Bottleneck



**NO community standard**

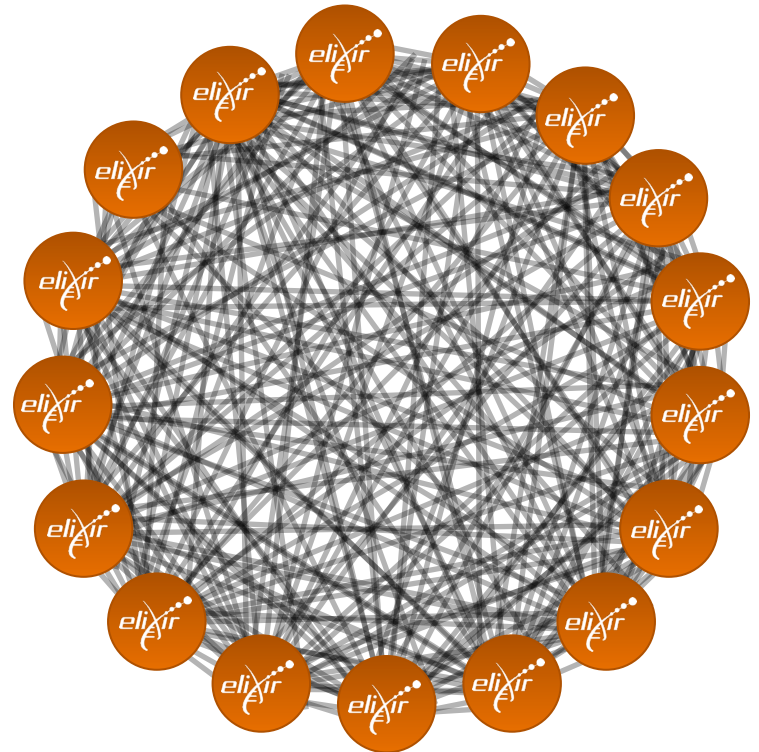
# Current dissemination

1 event per Node

Total of 306 events displayed

**289 manual copies**

17 Nodes



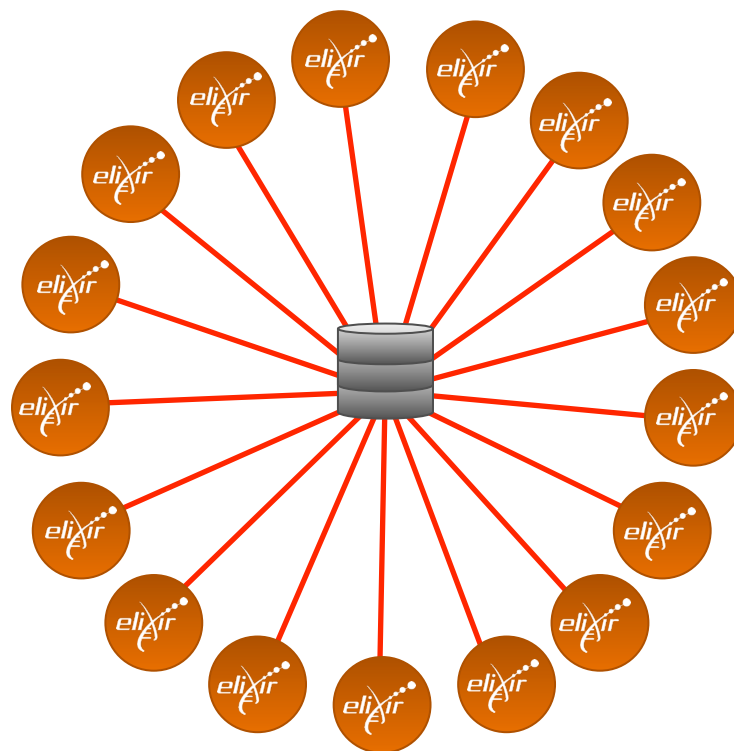
# iAnn dissemination

17 Nodes

1 event per Node

Total of 306 events displayed

**1 manual** copies to do



An event is curated ~~once~~ (twice)

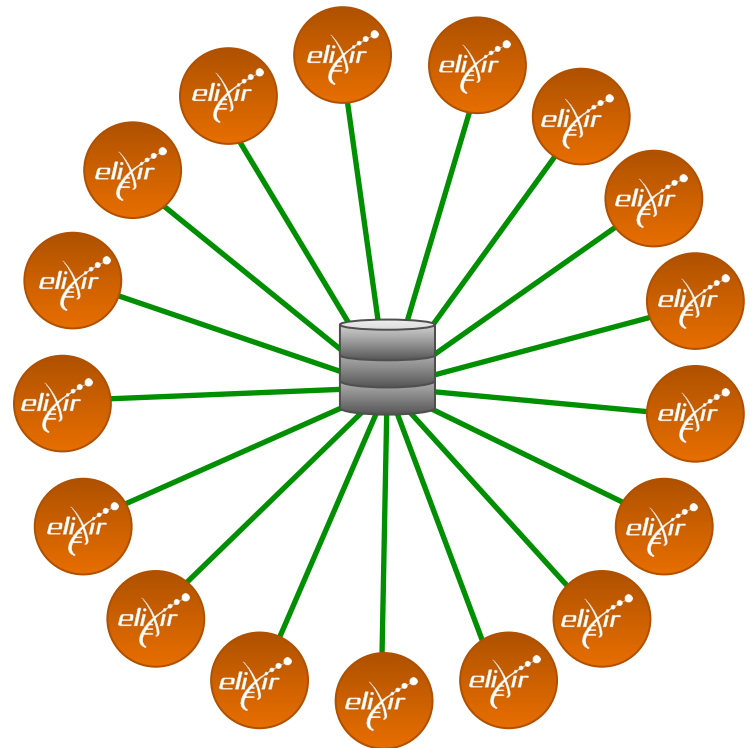
# Automated dissemination

1 event per Node

Total of 306 events displayed

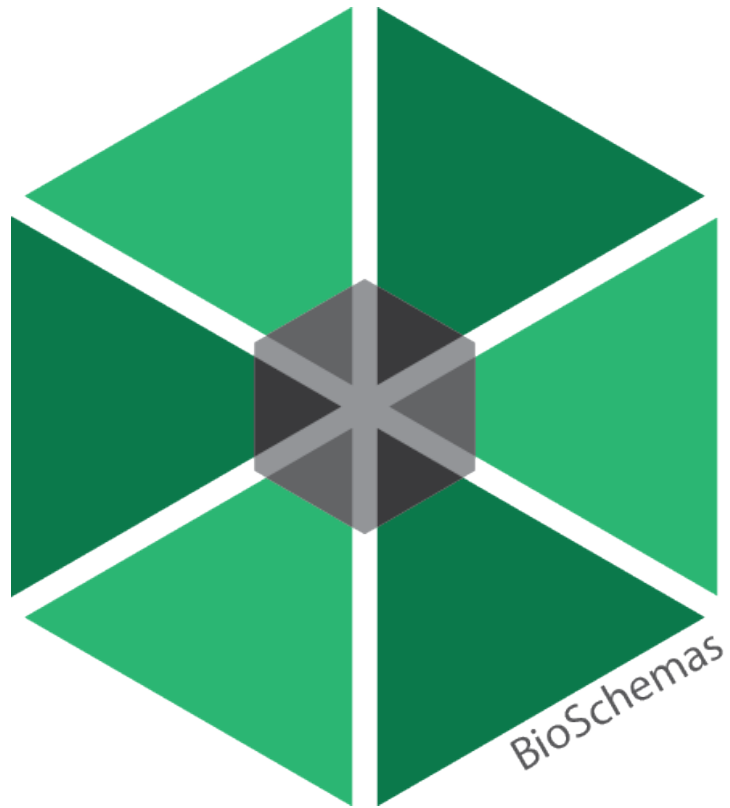
**0 manual copies**

17 Nodes



Create it **once**, share it everywhere

# BioSchemas

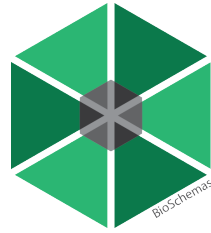


# Problem

Information about life science  
**events, training materials, people,**  
**organizations, standards, ...**  
is vast, scattered and lacks common description

Difficult to exchange and integrate



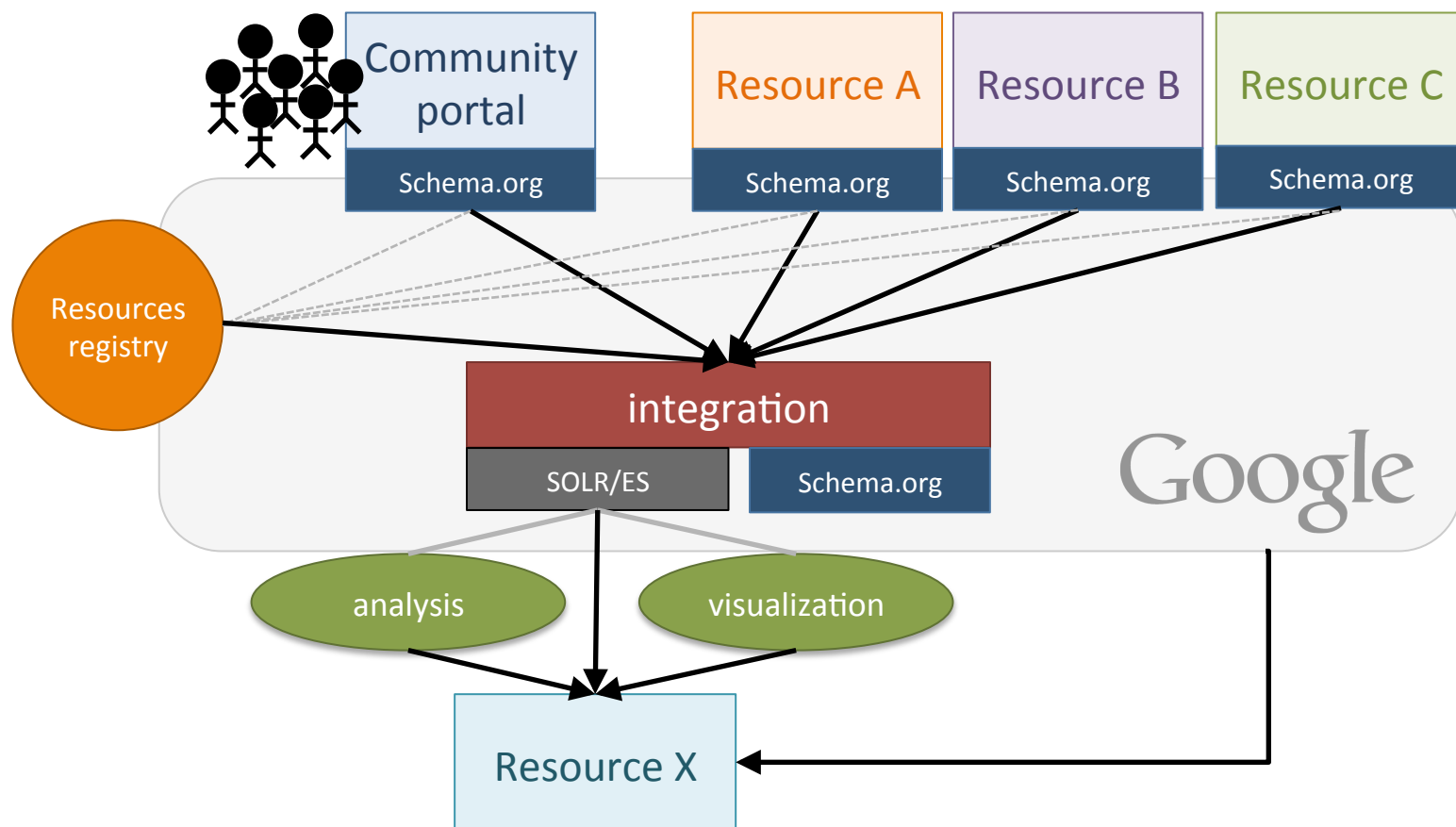


# BioSchemas

- **Metadata description spec**  
... for events, training materials, people, organizations, standards, ...
- **Relying on schema.org standards**  
... facilitating work of generic search engines
- **Tailored to life science**  
... including agreements on CV, Minimum guidelines and cardinality
- **Minimum effort to adopt and maintain**  
... including documentation and examples
- **Easy to contextualize across content types**  
... with common metadata properties across different types

# Application

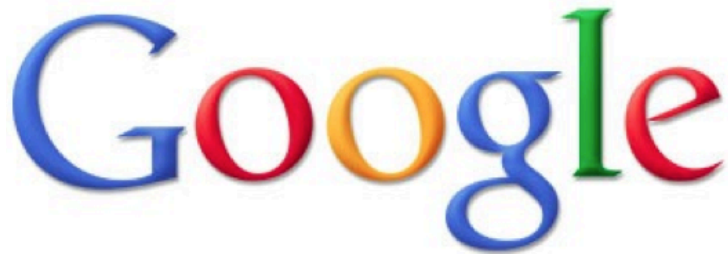
## ELIXIR event portal



# schema.org

metadata properties

agreed by major search engines  
to improve search & discoverability

The Google logo, featuring the word "Google" in its characteristic multi-colored font (blue, red, yellow, green, red).The Yahoo! logo, featuring the word "YAHOO!" in a purple, serif font.The Yandex logo, featuring the word "Yandex" in a black, sans-serif font, with a red "Y" at the beginning.The Bing logo, featuring the word "bing" in a blue, sans-serif font, with a small orange dot above the "i" and a "TM" trademark symbol.

# Components

- Metadata definition
  - **Schema.org**
- Formats
  - **Microdata**
    - Easy HTML tagging
  - **RDFa**
    - HTML tagging using a lite variant of RDF
  - **JSON-LD**
    - JSON used for link data.
    - No HTML tagging

# Components

- Metadata definition
  - **Schema.org**
- Formats

**GUI**

- **Microdata**

- Easy HTML tagging

- **RDFa**

- HTML tagging using a lite variant of RDF

- **JSON-LD**

- JSON used for link data.
- No HTML tagging

**GUI & WS**



# Lessons learnt

- Not just use but reuse

Data store, Interfaces, Query language, Libraries, Visualization, ...

- Model based on agreed (bio)standards
- Contribute and leverage search engines  
metadata

More synergies between data repositories, custom search engines and generic search engines.

# Metadata searching meeting

@ELIXIR All-Hands, 10 March, Barcelona, Spain

Join us!

- Sessions

- Microtagging for biological entries and datasets

Brainstorming and exchange of ideas

- Federated metadata searching

Presentation and review the state of the art

- Hands-on searchathon

testing, adoption or pushing forward ideas

- Expected outcomes

- Project proposal for NIH/BD2K and ELIXIR
- A review of existing efforts in ELIXIR and NIH/BD2K
- Tests, prototypes, examples to engage community

<http://tinyurl.com/MetadataSearching>



# Acknowledgements

- Many people from ...
  - ELIXIR
  - GOBLET
  - BBMRI
  - Pistoia Alliance
  - TeSS
  - iAnn
  - BioSchemas