Visualization for Scientists

We discuss how Deluge and Complexity call for new ideas in data exploration. Learn more, find tools at layerscape.org
What is Globus Online?

• **Transfer and synchronize files**
  – Easy “fire-and-forget” transfers
  – Automatic fault recovery
  – High performance
  – Across multiple security domains

• **Minimize IT costs**
  – Software as a Service (SaaS)
    • No client software installation
    • New features automatically available
  – Consolidated support & troubleshooting
  – Simple endpoint installation with Globus Connect and GridFTP

• **Recommended by XSEDE, Blue Water, NERSC, ALCF, ESnet, many Universities**
Eamonn Maguire. The ISA Software Suite and Format.

1. **Create templates** detailing the steps to be reported for different investigations, complying to community standards (listed at [isa-tips](https://isa-tips.org) and [isa-formats](https://isa-formats.org)), e.g., configuring fields to be (i) ontology terms, (ii) text (with/without regular expression testing), (iii) numbers etc.

2. **Report and edit** the description of the investigation using ISAcreator.
   - Comes complete with ontology search, browse and tagging support, publication search, file lookup (FTP and local file system), experimental design wizard, legacy format mapper and much more.

3. **Perform analysis** in GenomeSpace, R-based BioConductor or Galaxy using relevant ISA modules.

4. Share the investigation as Linked Data through conversion to RDF.

5. Convert to other formats for submission to public repositories.

6. Direct submission in ISA-Tab

7. Submit to public repositories

8. Store the investigations in the database, assign access rights and conduct maintenance tasks.
   - Command line version and API also available

9. ISA-Tab formatted investigations can also be created via a growing number of other tools (under the ISAcommons)

10. BI WEB APPLICATION

11. BI DATA MANAGER

12. Store the investigations in the database, assign access rights and conduct maintenance tasks.

13. Command line version and API also available

14. ISA CONVERTER

15. Direct submission in ISA-Tab

16. Submit to public repositories

17. Store the investigations in the database, assign access rights and conduct maintenance tasks.

18. Command line version and API also available

19. ISA EXPRESS

20. ISA PREDICT

21. ENA

22. BI WEB APPLICATION

23. BI DATA MANAGER

24. Store the investigations in the database, assign access rights and conduct maintenance tasks.

25. Command line version and API also available

26. ISA EXPRESS

27. ISA PREDICT

28. ENA

29. BI WEB APPLICATION

30. BI DATA MANAGER

31. Store the investigations in the database, assign access rights and conduct maintenance tasks.

32. Command line version and API also available

 ISAtools
 isatools.org

 ISAcommons
 isacommons.org
Increasing Frustration with Research Papers

Prior to the computational revolution in science, research papers provided the primary mechanism for sharing data. As the science becomes more computational, with experiments involving large amounts of data, scientific data is often shared through means other than the research paper. Readers increasingly require that data and other objects from a scientific research be closely associated with the paper for quick and reliable access while reading the paper.

How do readers and reviewers want to read research papers?

<table>
<thead>
<tr>
<th>Described Concept</th>
<th>Implementation in source code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset Description</td>
<td>Dataset metadata</td>
</tr>
<tr>
<td>Figures</td>
<td>Figure derivation and workflows</td>
</tr>
<tr>
<td>Data Values</td>
<td>Location in the digital source</td>
</tr>
</tbody>
</table>

Creating these linkable scientific objects (SO) imposes significant author burden.

Our approach: SOLE

Our Science Object Linking and Embedding (SOLE) system addresses these challenges by:

✧ Using tagging to convert science objects (source codes, datasets, workflows) to linked data objects (associating a URI)
✧ Maintaining their representation in a bibliography-like tool

Demonstration

✧ Tools for authors that make it far easier to associate science objects (source codes, datasets, workflows) to papers
✧ Associating online data repositories, such as PubChem, SDSS, PDB with research papers.
**Data Reuse**

**Data Sharing**

**Data Management**

---

**DATAUP**

Describe, Manage, & Share Your Data
A Service of the University of California Curation Center (UC3)

- Free & open-source
- Add-in or web app
- Connects to DataONE
- Features
  - Check for best practices
  - Create metadata
  - Create a citation
  - Archive data
- Learn more: dataup.cdlib.org

---

University of California
Funders: Microsoft Research Connections; Gordon and Betty Moore Foundation

---

From Flickr by c sessums
From Flickr by Flickmor
From Flickr by Rick by c sessums
Led by a group of advisors and editors from 13 different countries, Databib is identifying and cataloging the world’s repositories of research data.

- Find repositories to submit your data to
- Find data to reuse in your research or learning
- Open access, open source, curated, sustainable
- Easy to integrate (RSS, Linked Data, OpenSearch, RDF/XML, Twitter, etc.)

For more information, visit http://databib.org
Urban Myth at the dawn of eScience era: (with a flavour of Bioinformatics)

1. Bioinformatics should be done on Linux only (Mac allowed).
2. Because every tool is written on Linux for Linux.
3. It’ll be hard to port any of these C/C++ tools to Windows.
4. You just can’t do it, because Windows does not support Pthread.
5. Even you can, it will run much slower than on Linux.
6. Even if it’s faster, although I don’t believe it, who is going to use it?
7. Definitely nobody wants to use Windows for anything serious right?
8. No no, not again, we already got Amazon AWS and Google Engine, we don’t need yet another cloud, too much rain these days.
10. Big Data you mean, that’s Hadoop & HPC Cluster’s backyard.
12. You should really stop, there is better thing to do than challenging the status quo.

My 2 pence on changing the world:

1. How I ported BWA and SAMTOOLS in C99 to Windows
2. How I enhanced BWA on parallelization
3. How I built my Windows Azure pipeline for Whole Genome Sequencing data, see it live here!

Bonus: What is sequencing all about? Is there really a gene for Depression?
Environmental Informatics Framework (EIF)

Current State of Data Ecosystem

Applications
- Android
- iPhone
- Windows Phone
- WebOS
- Excel
- Java
- Silverlight
- .NET
- AJAX
- PHP
- MATLAB

Data Sources
- SQL
- XML
- Data Cube
- CSV

Open Data Protocol