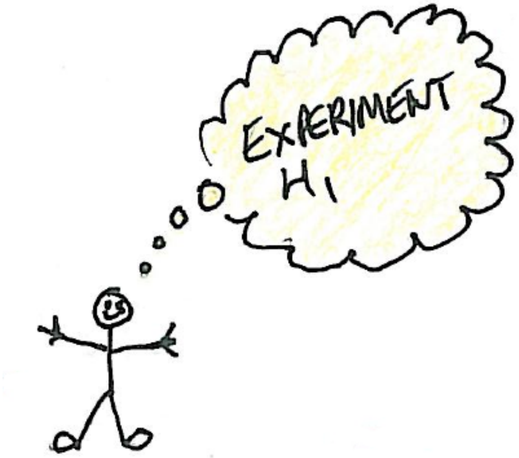
Research Objects: Towards Exchange and Reuse of Digital Knowledge

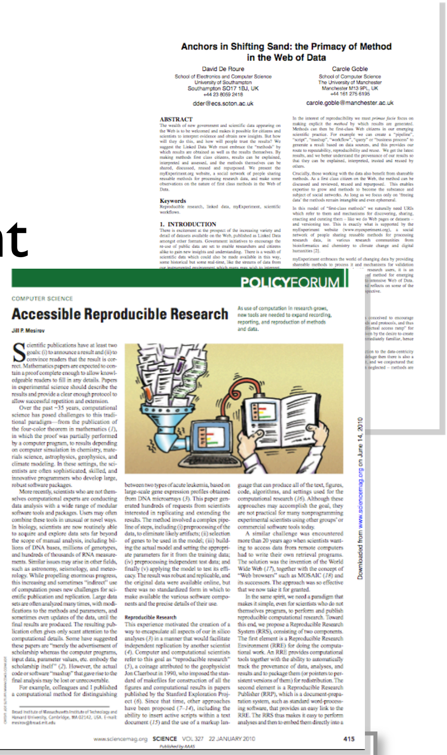


Sean Bechhofer
University of Manchester
sean.bechhofer@manchester.ac.uk
@seanbechhofer
http://humblyreport.wordpress.com

CERN Workshop on Innovations in Scholarly Communication (OAI7), Geneva, 22/6/11

Publication

- Argumentation: *Convince* the reader of the validity of a position [Mesirov]
  - Reproducible Results System: facilitates enactment and publication of reproducible research.



Anchors in Shifting Sand: the Primacy of Method in the Web of Data

David De Roure
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
+44 23 8059 2418
dder@ecs.soton.ac.uk

Carole Goble
School of Computer Science
The University of Manchester
Manchester M13 9PL, UK
+44 161 275 6195
carole.goble@manchester.ac.uk

ABSTRACT

The wealth of new government and scientific data appearing on the Web is to be welcomed and makes it possible for citizens and scientists to interpret evidence and obtain new insights. But how will they do this, and how will people trust the results? We suggest the Linked Data Web must embrace the "methods" by which results are obtained as well as the results themselves. By making methods first class citizens, results can be explained, interpreted and assessed, and the methods themselves can be shared, discussed, reused and repurposed. We present the myExperiment.org website, a social network of people sharing reusable methods for processing research data, and make some observations on the nature of first class methods in the Web of Data.

Keywords

Reproducible research, linked data, myExperiment, scientific workflows.

1. INTRODUCTION

POLICYFORUM

COMPUTER SCIENCE

Accessible Reproducible Research

Jill P. Mesirov

As use of computation in research grows, new tools are needed to expand recording, reporting, and reproduction of methods and data.

Scientific publications have at least two goals: (i) to announce a result and (ii) to convince readers that the result is correct. Mathematics papers are expected to contain a proof complete enough to allow knowledgeable readers to fill in any details. Papers in experimental science should describe the results and provide a clear enough protocol to allow successful repetition and extension.

J. Mesirov **Accessible Reproducible Research** *Science* 327(5964), p.415-416, 2010 http://dx.doi.org/10.1126/science.1179653

- Results are reinforced by *reproducability* [De Roure]
  - Explicit representation of *method.*

D. De Roure and C. Goble **Anchors in Shifting Sand: the Primacy of Method in the Web of Data** *Web Science Conference 2010, Raleigh NC*, 2010 http://eprints.ecs.soton.ac.uk/20817/

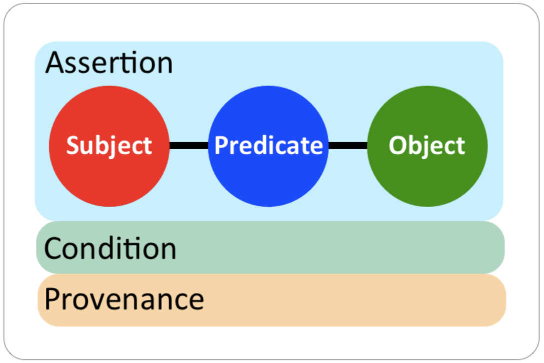
- *Verifiability* as a key factor in scientific discovery.

Stodden et. al. **Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science** *Computing in Science and Engineering* 12 (5), p.8-13, 2010 http://dx.doi.org/10.1109/MCSE.2010.113

Publication

- Nano-publications. Explicit representation at the *statement* level.

Groth et. al. **The Anatomy of a Nano-publication** *Information Services and Use* 30(1), p.51-56, 2010 http://iospress.metapress.com/index/FTKH21Q50T521WM2.pdf



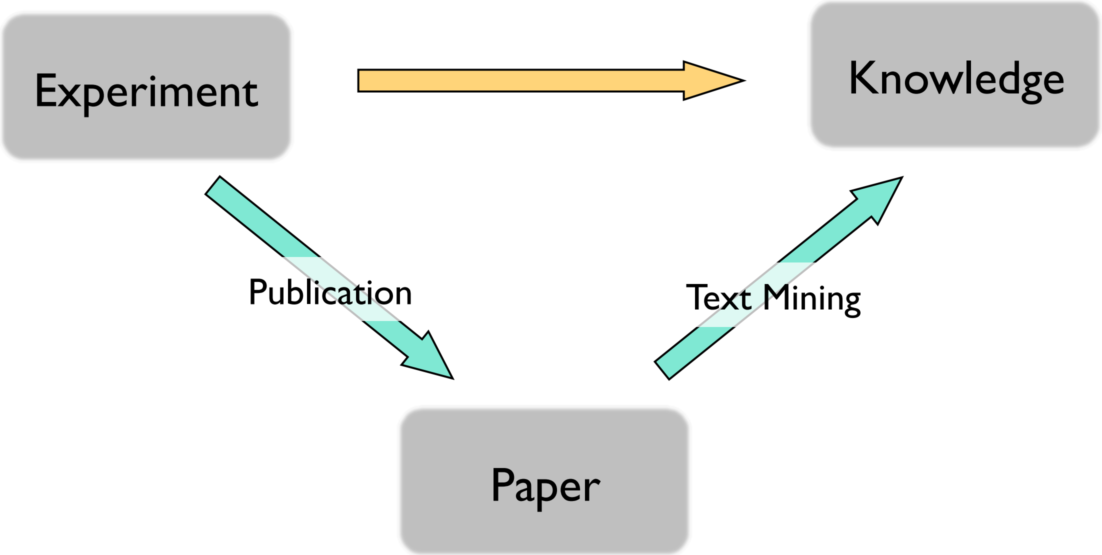
- Executable Papers
  - Collage
  - SHARE
  - Verifiable Computational Results

Nowakowski et. al. **The Collage Authoring Environment** *ICCS 2011*, 2011 http://dx.doi.org/10.1016/j.procs.2011.04.064

Van Gorpet. al **SHARE: a web portal for creating and sharing executable research papers** *ICCS 2011*, 2011 http://dx.doi.org/10.1016/j.procs.2011.04.062

Gavish et. al. **A Universal Identifier for Computational Results** *ICCS 2011*, 2011 http://dx.doi.org/10.1016/j.procs.2011.04.067

Knowledge Burying in paper publication



- Publishing/mining cycle results in loss of knowledge
  - ≥ 40% of information lost
- RIP – *Rest in Paper*
- Need for mechanisms for publication of knowledge, preserving information about the *process*.

B.Mons **Which Gene Did You Mean?** *BMC Bioinformatics* 6 p.142 2005
http://dx.doi.org/10.1186/1471-2105-6-142

The Problem

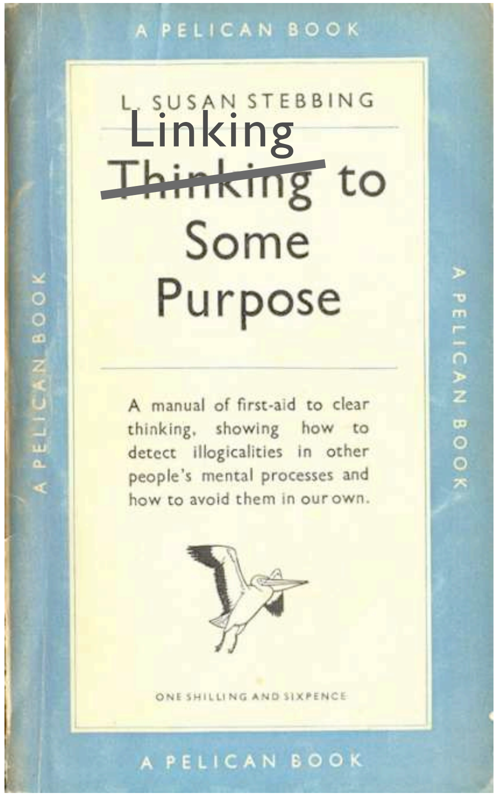
- Moving to digital environments
  - Workflows, protocols, algorithms
  - Consuming and producing data
  - Electronic publication methods
- From (linear) paper publications to….

???

- Need for frameworks for facilitating *reuse* and exchange of digital knowledge

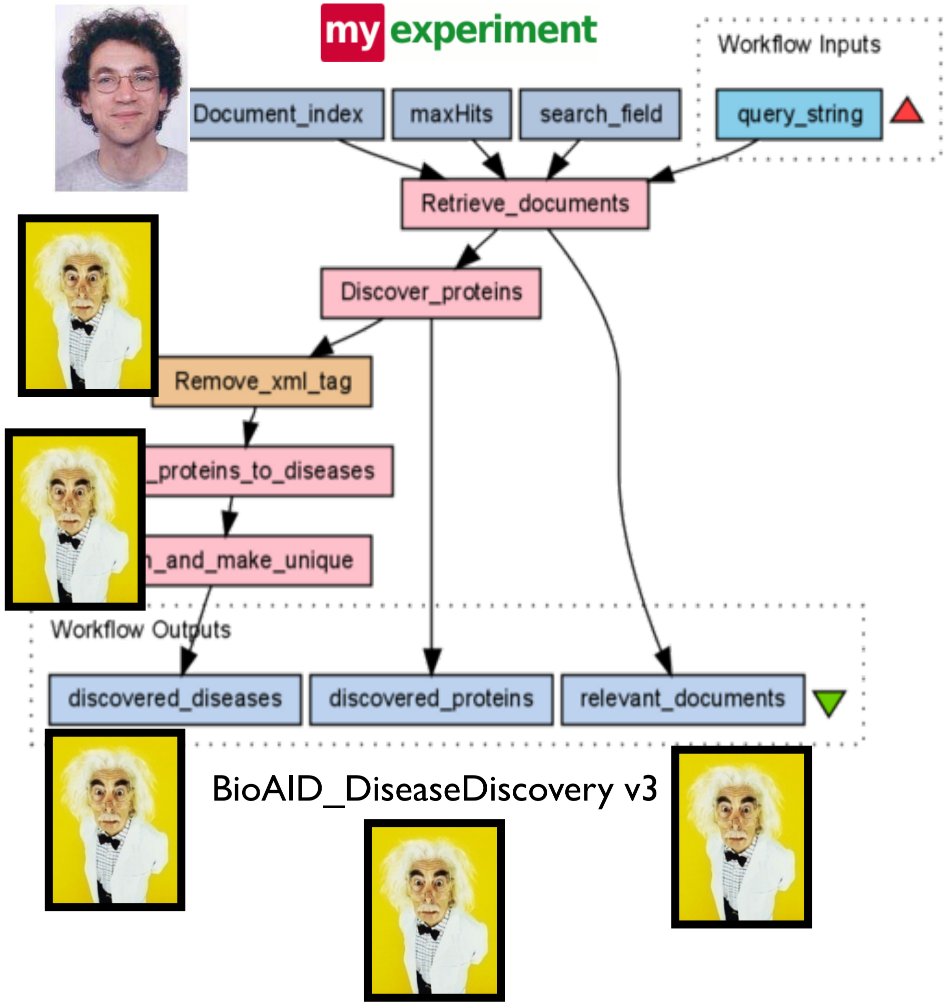
Research Objects

Semantically rich aggregations of resources, supporting a research *objective*


A PELICAN BOOK
L. SUSAN STEBBING
Linking
~~Thinking~~ to
Some
Purpose
A manual of first-aid to clear thinking, showing how to detect illogicalities in other people's mental processes and how to avoid them in our own.
ONE SHILLING AND SIXPENCE
A PELICAN BOOK
A PELICAN BOOK
A PELICAN BOOK

Workflows

A Scientific Workflow can be seen as the combination of data and processes into a configurable, structured set of steps that implement semi-automated computational solutions in scientific problem-solving



BioAID_DiseaseDiscovery v3

- Central in experimental science
  - Enable automation
  - Make science *repeatable* (and sometimes *reproducible*)
  - Encourage best practices
- Scientist-friendly
  - Aimed at (some types of) scientists, possibly even without strong computational skills
- Communities: Need for scientific data preservation
  - Enhance scientific development by building on, sharing, and extending previous results within scientific communities
- However, workflow preservation is especially complex
  - Workflows not only specified statically at design time but also interpreted through their execution
  - Complex models are required to describe workflows and related resources, including documents, data and services
  - Resources often beyond control of scientists

Motivating Projects

- myExperiment
  - Workflow sharing
- Sysmo-DB
  - Assets catalogue supporting exchange of data, models, SOPs
- Obesity e-Lab/MethodBox
  - Sharing survey data/analysis scripts

myexperiment

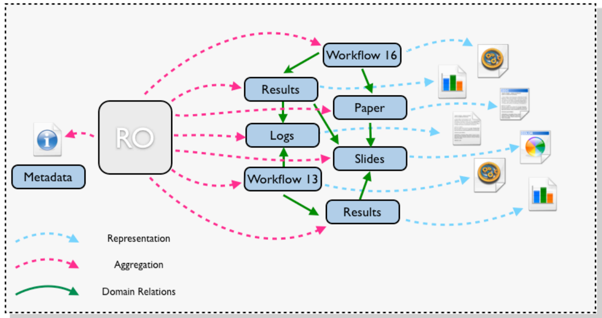
- "Facebook for Scientists" ...but different to Facebook!
- A repository of research methods
- A community social network of people and things
- A Social Virtual Research Environment
- A probe into researcher behaviour
- Open source (BSD) Ruby on Rails app
- REST and SPARQL interfaces, supports Linked Data
- Part of product family including BioCatalogue, MethodBox and SysmoDB

4000 members, 200 groups, ~1500 workflows, ~150 packs

myGrid MANCHESTER 1824 UNIVERSITY OF Southampton School of Electronics and Computer Science OXFORD e-Research CENTRE e-Research SOUTH JISC EPSRC Microsoft Research

Motivating Projects

- myExperiment
  - Workflow sharing
- Sysmo-DB
  - Assets catalogue supporting exchange of data, models, SOPs
- Obesity e-Lab/MethodBox
  - Sharing survey data/analysis scripts
- myExperiment packs
  - Packs supporting (simple) aggregations.
  - *Links* not just references
  - Packs as nascent ROs



Wf4Ever

…technological infrastructure for the preservation and efficient retrieval and reuse of scientific workflows in a range of disciplines.

- Architecture/implementation for workflow preservation, sharing and reuse
- Research Object models
- Workflow Decay, Integrity and Authenticity
- Workflow Evolution and Recommendation
- Provenance
- Driven by Use Cases

FP7 Digital Libraries and Digital Preservation
iSOCO, University of Manchester, Universidad Politécnica de Madrid, University of Oxford, Poznan Supercomputing and Networking Centre, Instituto de Astrofísica de Andalucía, Leiden University Medical Centre

Wf4Ever presented at the .Astronomy Conference

4/12/11 10:56 AM

Wf4Ever was presented at the third edition of the Astronomy Conference held in Oxford New College during 4th

Project ID card

Funded under: 7th FWP (Seventh Framework Programme)

Area: Digital Libraries and Digital Preservation. (ICT-2009.4.1)

Project reference: 270192

Total cost: 3.86 million euro

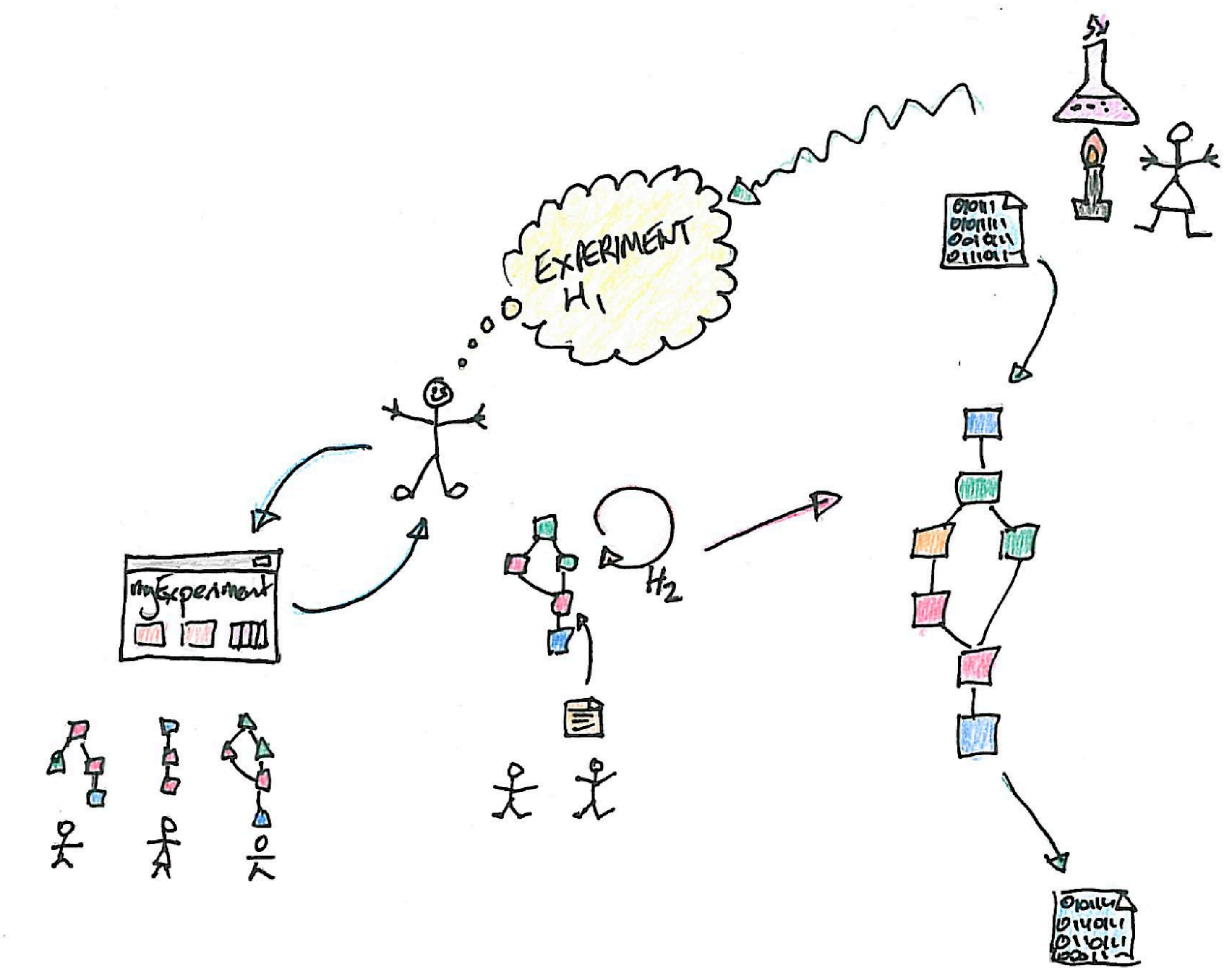
EU contribution: 2.94 million euro

Execution: From 2010-12-01 to 2013-11-30

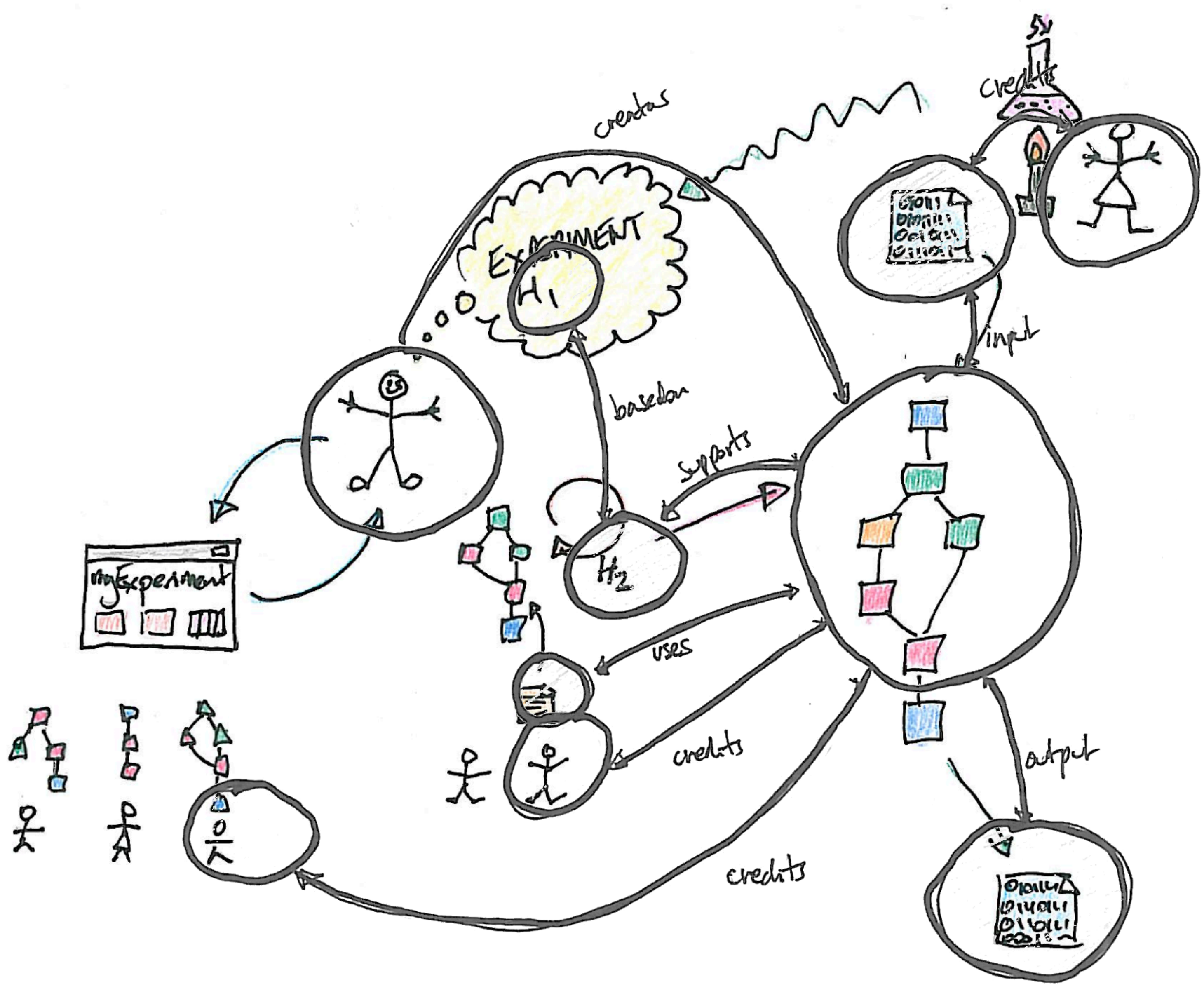
Duration: 36 months

Project status: Execution

Bio Scenario



Bio Scenario



Astronomers Questions

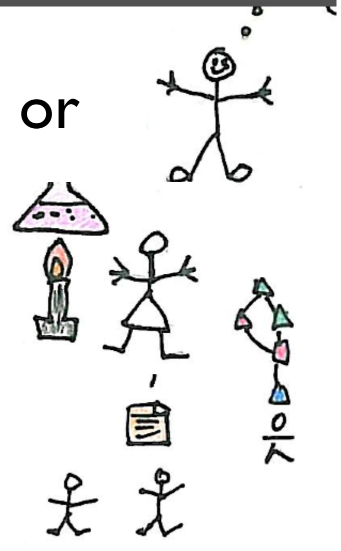
When *accessing* a workflow

- Can I use it for my purposes (in my words)?
- If I can expect it to run, when was it was last run, by whom?
- What it does quickly, by one of
  - example input / output (and trying it)
  - a description
  - 'reading' its key parts
  - what it was used for
  - related workflows its creator
  - contacting the creator or last user
- How I need to cite the author and workflow?

When *sharing* a workflow

- What rights others have?
- What a good workflow is to get a good score?
  - Make my workflow findable, reusable, and ready for review
  - Instructions to authors
  - Two types of contributions: serious science, preliminary/playing around
- If my workflow may have issues
  - What the system or other users think it does
- How it relates to other things
- Share freely or anonymously upon request?

User Roles



Creator. Collecting together resources as an RO for reuse or repurpose. May be for personal use.

Contributor. Providing materials to be used within an RO

Collaborator. Providing materials to be used without necessarily being aware of the RO

Reader. Looking for related works, state of the art.

Comparator. Looking for similar or previous work to a task in hand

Re-User. Understands the underlying methods encapsulated (e.g. workflow) and how to extract/replace components.

Publisher. Disseminating results or methods. Upload to repository, publish via myExp, embed in blog post.

Evaluator/Reviewer. Evaluating/validating or reviewing content. Confirmation of results or validation of process.

Brought to you by the letter..

The *n* Rs of Research Reuse (Historical)

- ***Reusable*** – used as part of new study;
- ***Repurposeable*** – reuse the pieces in a new (and different) study. Substitute alternative data sets, methods;
- ***Repeatable*** – repeat the study, possibly years later;
- ***Reproducible*** – a special case of repeatability with a complete set of information/results to work towards;
- ***Replayable*** – go back and see what happened;
- ***Referenceable*** – cite in publications;
- ***Revealable*** – provenance and audit;
- ***Re-interpretable*** – crossing boundaries;
- ***Respectful*** – appropriate credit and attribution;
- ***Retrievable*** – discover and acquire.

D. De Roure **Replacing the Paper: The Twelve Rs of the e-Research** http://blogs.nature.com/eresearch/2010/11/27/replacing-the-paper-the-twelve-rs-of-the-e-research-record

Dimensions

Repeatability. Sufficient information to allow others to rerun.

Reproducability. Sufficient information for an independent investigator to obtain the same results

Replayability. A comprehensive record of what happened (not necessarily execution)

Live/Refreshable. Dynamic links to content

Justification. Why/how were decisions made? Provenance

Resilience. Change/loss/errors

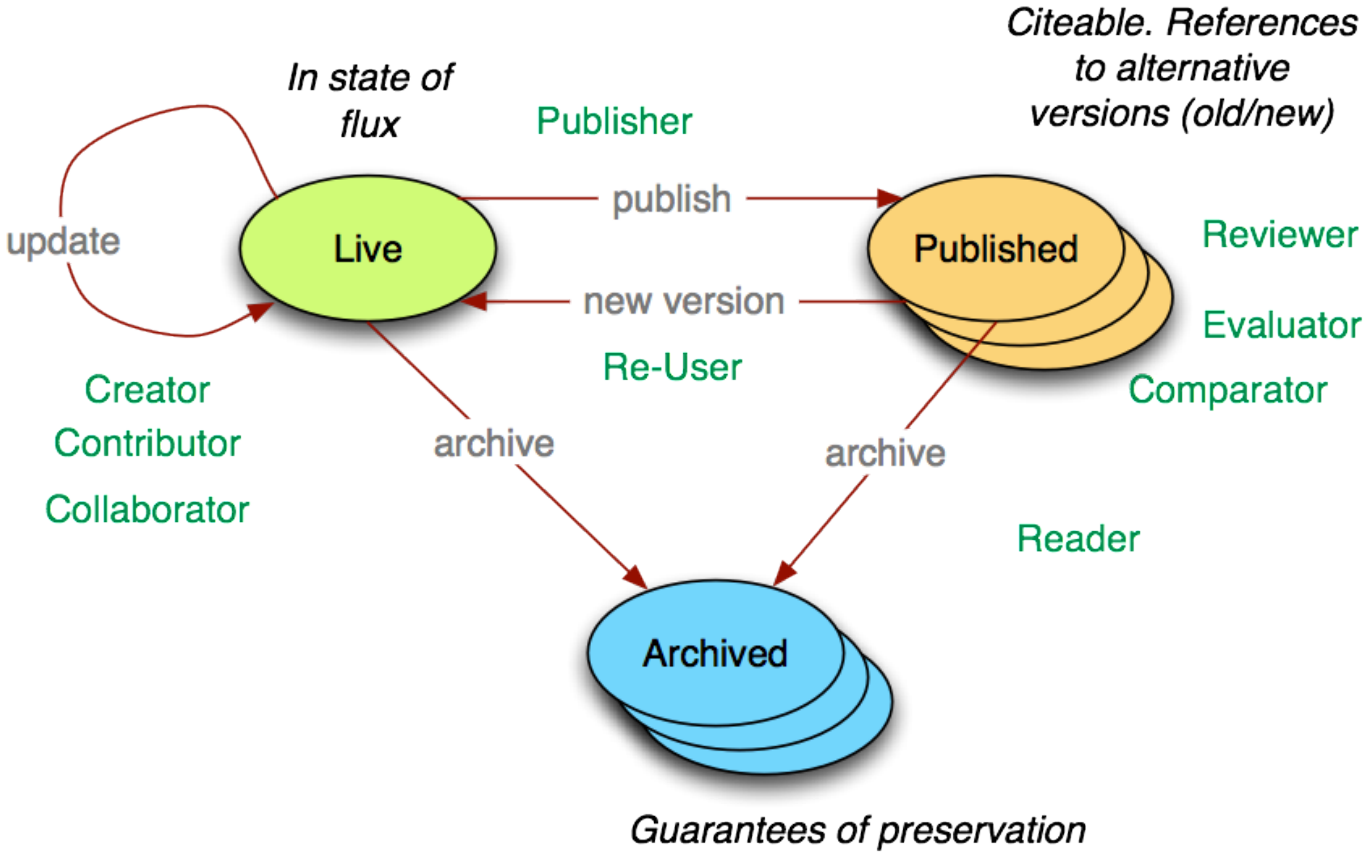
Discovery. Find/discover/index

Reference. Identification

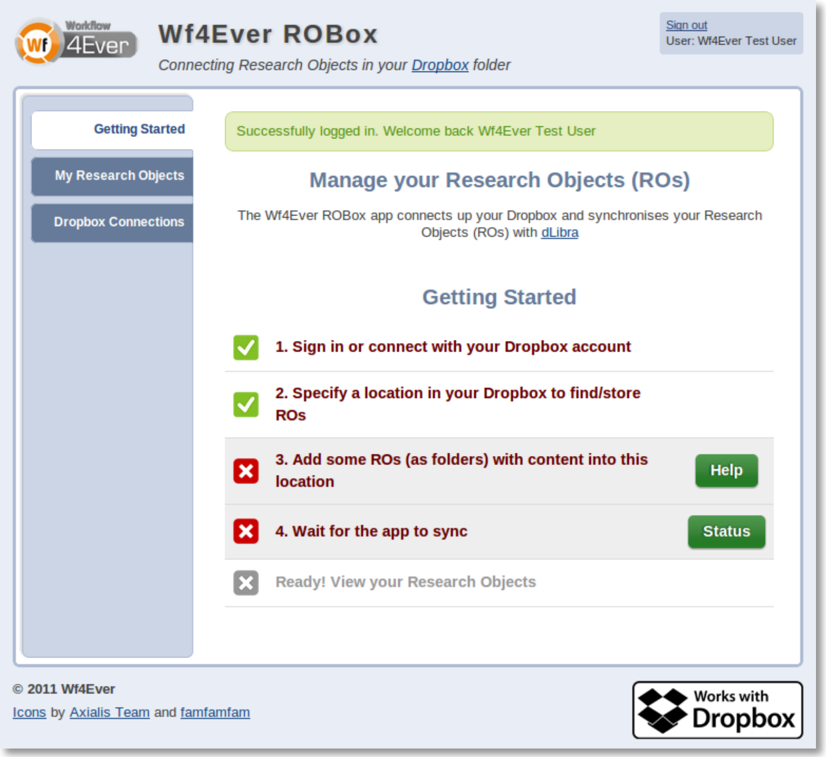
History. Rollback to retrace steps, fix errors.

Credit and Attribution. Record where resources come from.

Lifecycle


In state of flux
Publisher
Citeable. References to alternative versions (old/new)
update
Live
publish
Published
Reviewer
new version
Evaluator
Re-User
Comparator
Creator
Contributor
Collaborator
archive
archive
Reader
Archived
Guarantees of preservation

ROBox prototype



Wf4Ever ROBox

Connecting Research Objects in your Dropbox folder

Sign out
User: Wf4Ever Test User

Successfully logged in. Welcome back Wf4Ever Test User

Manage your Research Objects (ROs)

The Wf4Ever ROBox app connects up your Dropbox and synchronises your Research Objects (ROs) with dLibra

Getting Started

✓ 1. Sign in or connect with your Dropbox account

✓ 2. Specify a location in your Dropbox to find/store ROs

☒ 3. Add some ROs (as folders) with content into this location

☒ 4. Wait for the app to sync

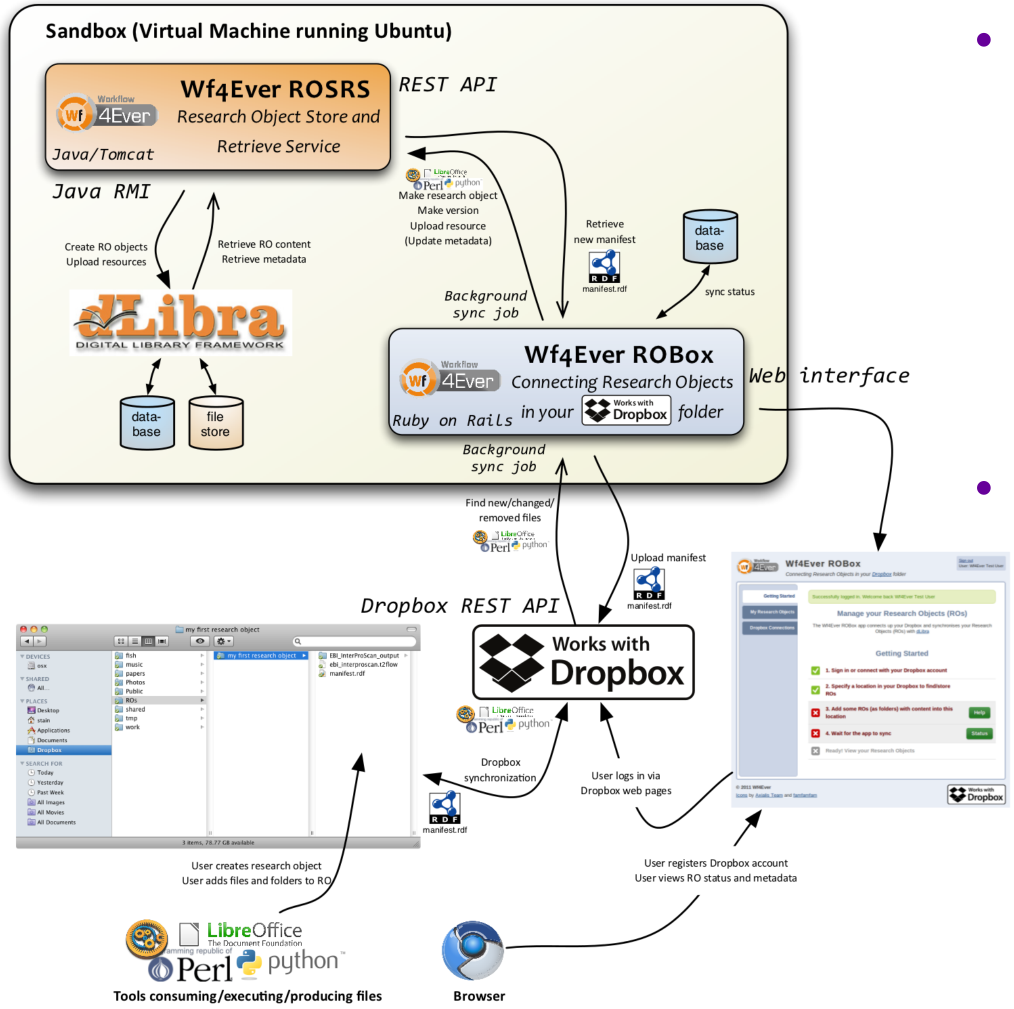
☒ Ready! View your Research Objects

© 2011 Wf4Ever
Icons by Axialis Team and famfamfam

- Collaboration support
- Shared Folder in Dropbox becomes working RO.
- Auto generation of metadata

ROBox prototype

Architecture of Prototype 1



Successfully logged in. Welcome back Wf4Ever Test User

Manage your Research Objects (ROs)

The Wf4Ever ROBox app connects up your Dropbox and synchronises your Research Objects (ROs) with dLibra

Getting Started

☑ 1. Sign in or connect with your Dropbox account

☑ 2. Specify a location in your Dropbox to find/store ROs

☒ 3. Add some ROs (as folders) with content into this location

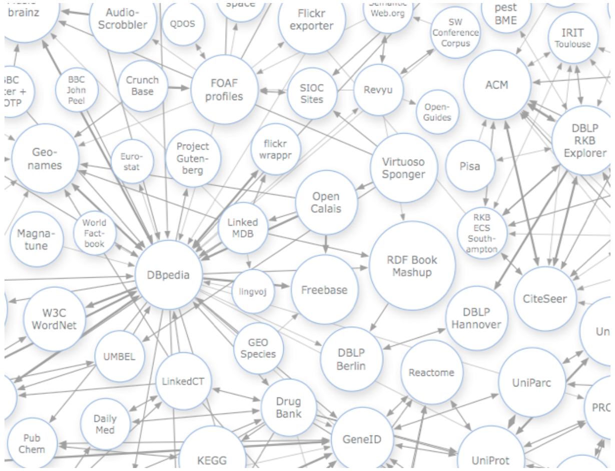
☒ 4. Wait for the app to sync

☒ Ready! View your Research Objects

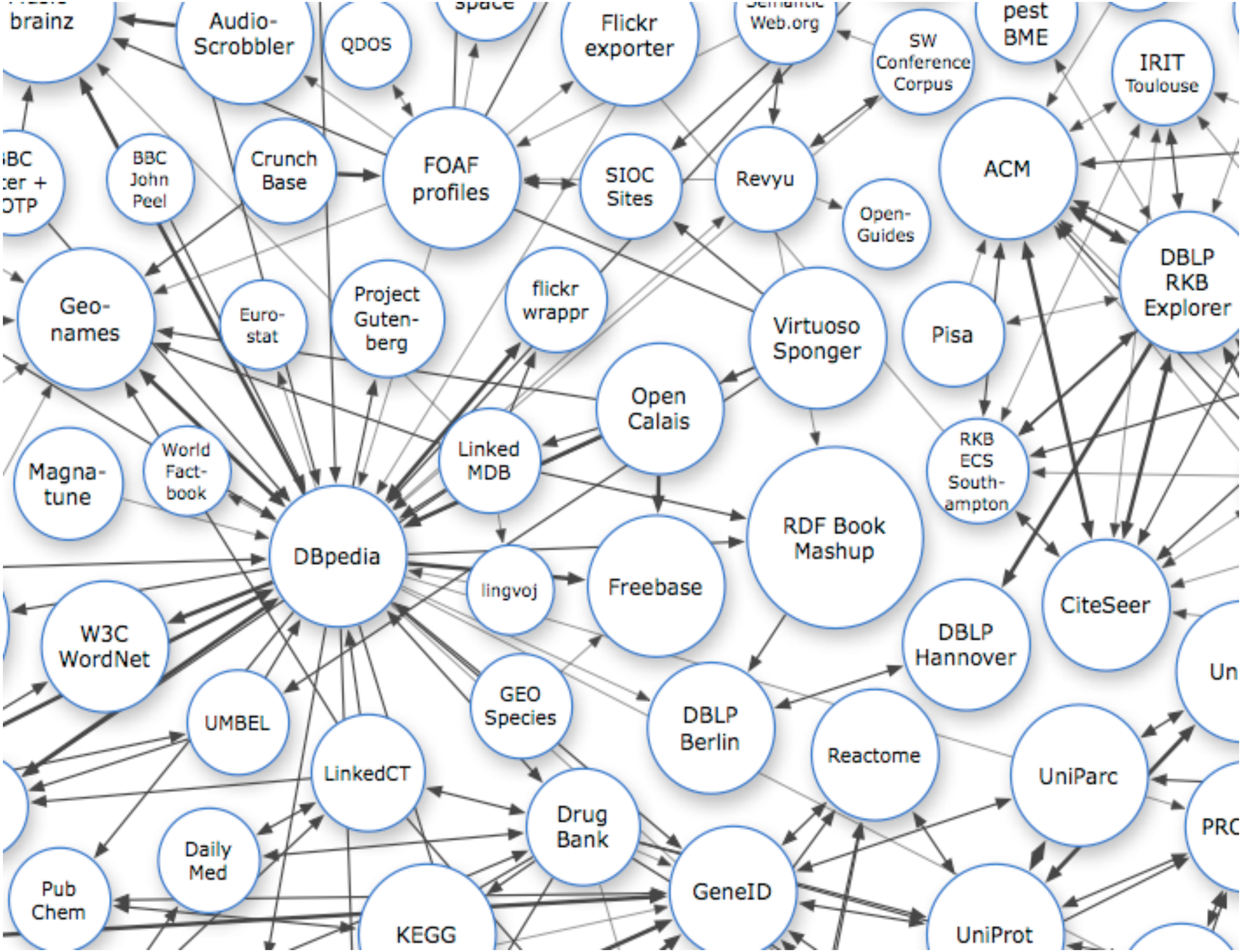
© 2011 Wf4Ever
Icons by Axialis Team and famfamfam

- dLibra backend for resources
- Manifest describes data package contents
  - Drawing on Admiral data package information (OXF)
  - DC terms
  - OAI-ORE aggregation vocabularies
- Editing/adding content triggers synchronisation and update to manifest

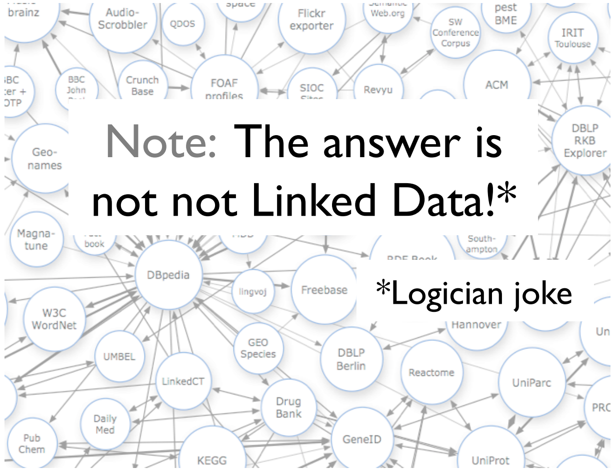
Linked Data



- A set of best practices for publishing and connecting data on the Web
  1. Use URIs to name things
  2. Use dereferencable HTTP URIs
  3. Provide useful content on lookup using standards
  4. Include links to other stuff


brainz
Audio-Scrobbler
QDOS
space
Flickr exporter
Web.org
SW Conference Corpus
pest BME
IRIT Toulouse
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
ACM
Open-Guides
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
DBLP RKB Explorer
Open Calais
Linked MDB
RKB ECS South-ampton
Magna-tune
World Fact-book
DBpedia
lingvoj
Freebase
RDF Book Mashup
CiteSeer
DBLP Hannover
W3C WordNet
UMBEL
GEO Species
DBLP Berlin
Reactome
UniParc
Un
LinkedCT
Drug Bank
PRC
Daily Med
Pub Chem
KEGG
GeneID
UniProt

Linked Data is not Enough!



Note: The answer is not not Linked Data!*

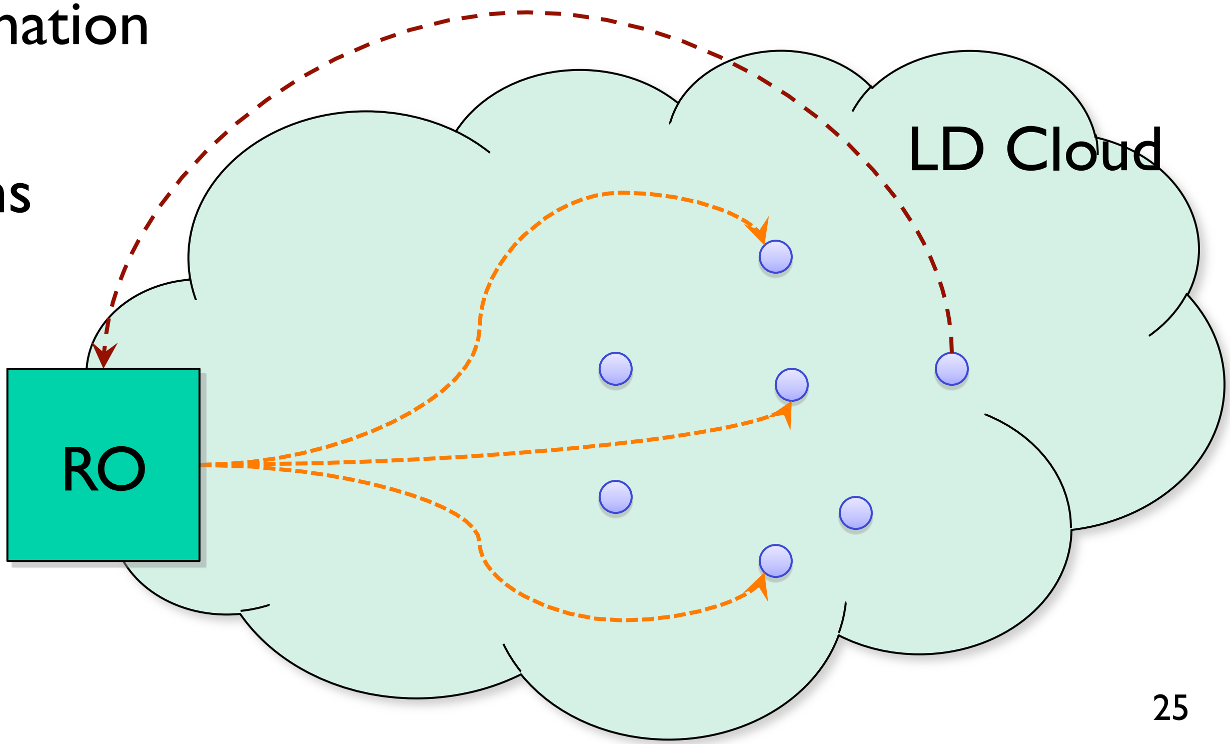
*Logician joke

- A set of best practices for publishing and connecting data on the Web
  1. Use URIs to name things
  2. Use dereferencable HTTP URIs
  3. Provide useful content on lookup using standards
  4. Include links to other stuff
- All very nice, lots of publishing going on, but no common models for *lifecycle*, *aggregation*, *ownership*, etc
- A platform for sharing and publishing, but *more* is needed

Bechhofer et al **Linked Data is not Enough for Scientists** *Sixth IEEE e-Science Conference*, 2010 http://dx.doi.org/10.1109/eScience.2010.21

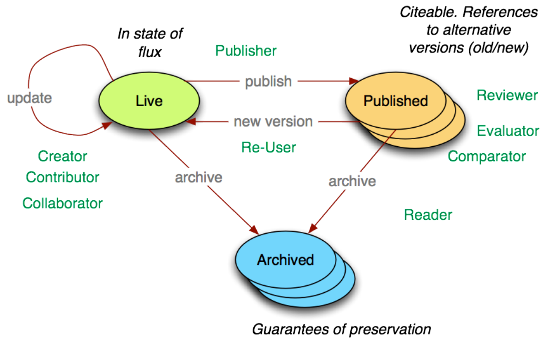
ROs and Linked Data

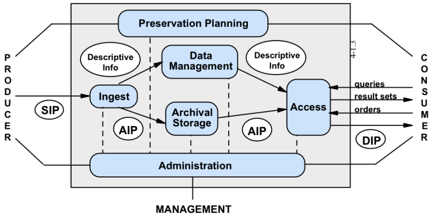
- Linked Data: Collection of best practices for publishing and connecting structured data on the web.
- ROs should be independent of mechanisms for representation and delivery
- ROs as non-information resources
  - "Named Graphs for LD"

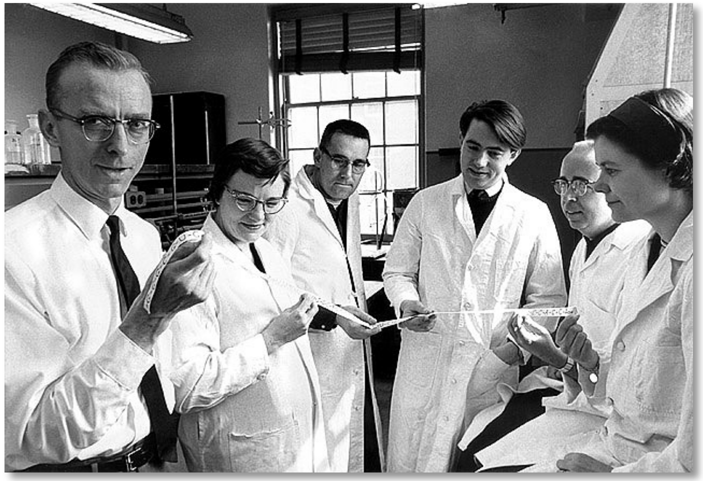


Where Next/Challenges

- Further Prototypes
- Models for Research Objects
  - Vocabularies
  - Refinement of lifecycle states
  - Provenance
- How much “sharing” can/should one support?
  - Intra group/Intra community/Anybody…
  - *Designated Communities*
- Identifiers
- *Publishing?*
- Versioning
- Trust







The Vision

- ROs: Aggregations to support sharing/publication.
- Incorporating methods, data, people
- Research Objects will allow us to conduct research in ways that are
  - *Efficient*: cheaper to borrow than recreate;
  - *Effective*: larger scale through reuse;
  - *Ethical*: Benefiting wider communities, not just individuals.
- *Could I have a copy of your Research Object please?*

Thanks!

- Manchester Information Management Group
  - http://img.cs.manchester.ac.uk
- myGrid Team
  - http://www.mygrid.org.uk/
- Wf4Ever Team
  - http://www.wf4ever-project.org/

Image Sources

- Cookie Monster: http://www.flickr.com/photos/nickstone333/3135318558
- Present: http://www.flickr.com/photos/powerhouse_museum_photography/3128638021
- Question Mark: http://www.flickr.com/photos/-bast-/349497988/
- R: http://www.flickr.com/photos/deks/185651630/
- Round Table: http://www.flickr.com/photos/svwscoop/3962423902/
- Scientists: http://www.flickr.com/photos/marsdd/2986989396/