Research Data: The First Mile

or...

“Where does the research data come from?”

Graham Klyne
October 2013

Background: Empty Data Archives

- I have for several years been working on projects related to research data sharing
- Repositories have been created for data storage and publication
- But there is not (yet) much data in them
  - (or not as much as there should be)
  - not counting large public databases

Populating Data Repositories

Data is increasingly seen as a first class product of research, underpinning trust in results

- “Research Objects”
- Reviewability
- Reproducibility
- Funder mandates

But who creates the data?

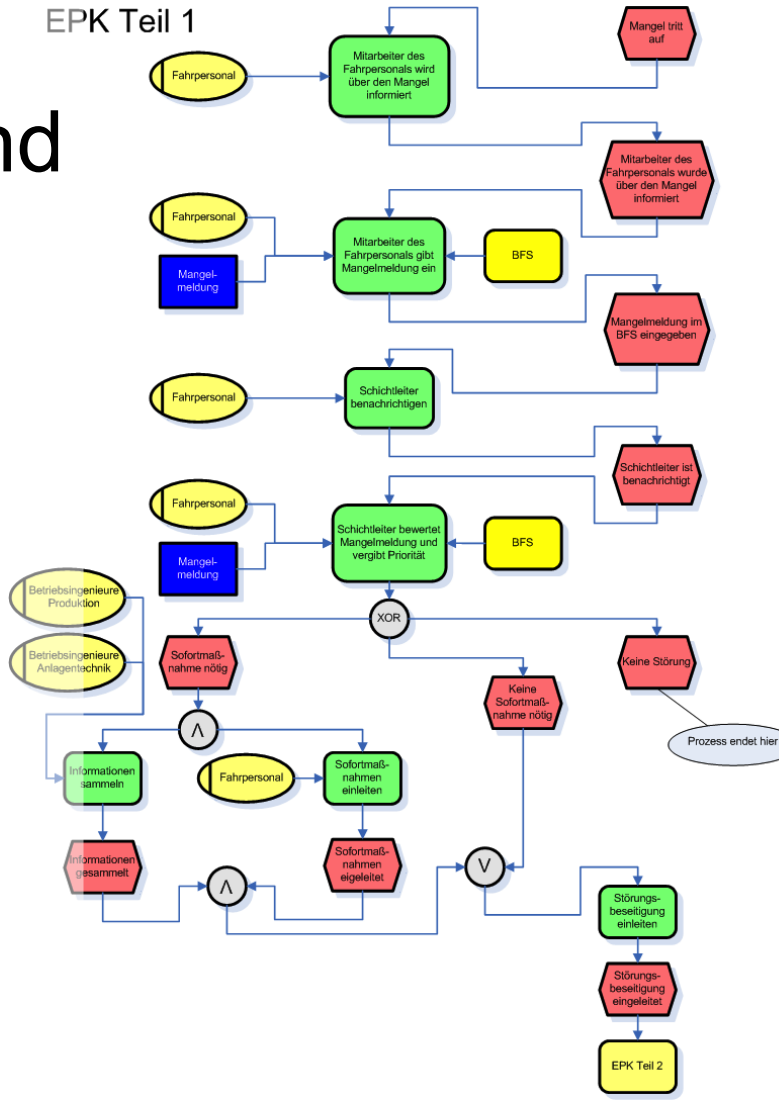
- Where does the data come from?

Large Research Projects

For large research projects, data management is planned and funded

The circumstances and methods of data generation are defined and managed

Dedicated IT support helps to make data acquisition, sharing and publication a reality



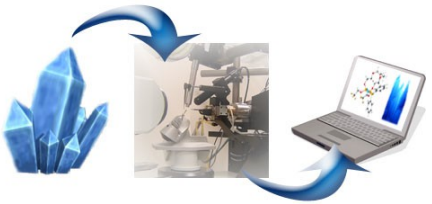
Subject and Other Databases

Examples:

- FlyBase, Beazeley Archive, eCrystals, UniProt, dbPedia
- many more – cf. http://databib.org

Separately funded

Often curated



Economies of scale?

Community portals, some with recognized academic value

Again: ***Dedicated IT support***

But What About the Little Guys?

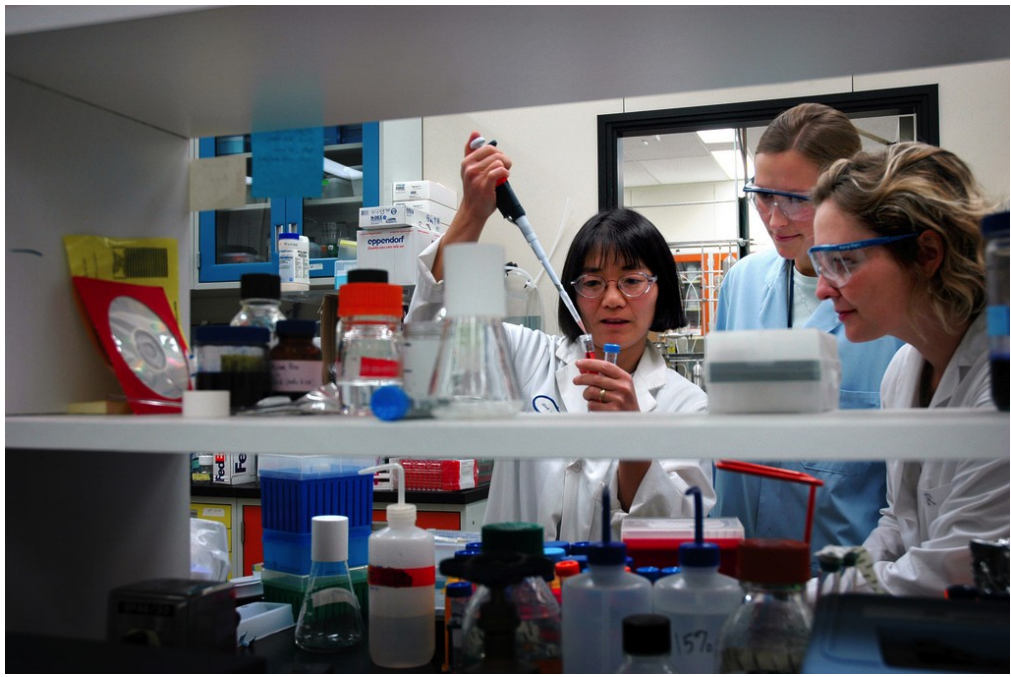
Small research groups

- e.g. 1-5 people
- Substantially manual processes
- Working with existing software tools
- No capability or capacity for custom software development

Large projects have small groups too

The “long tail” of data creation?

Small Research Group Data



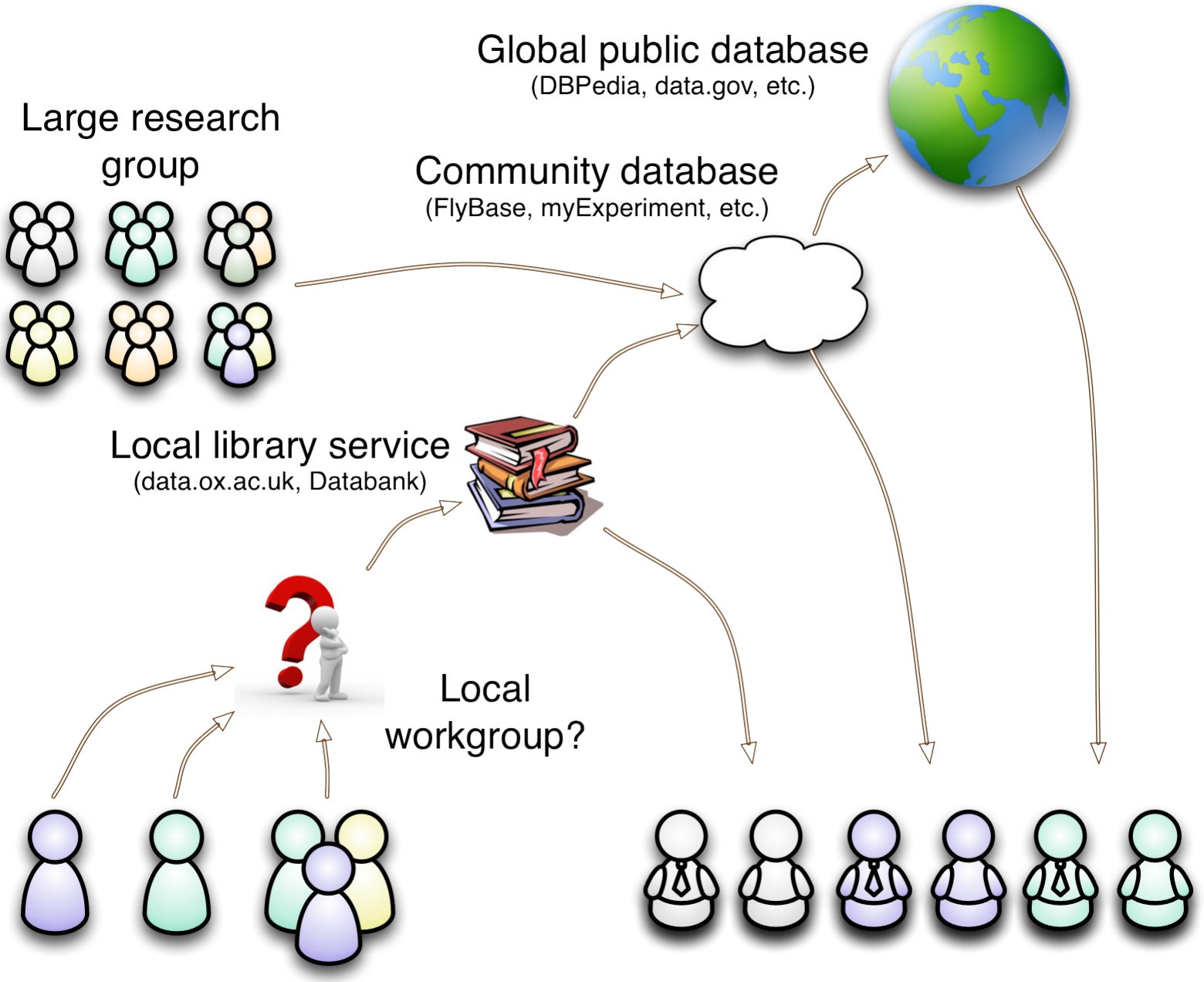
Data comes from:

- Hand-written notebooks
- Spreadsheets
- Documents (computer text)
- Instruments

  not necessarily networked
- Stand-alone software tools
- Web sites and online reference

Local *ad hoc* connection to global databases

A Missing Link



Some Applications

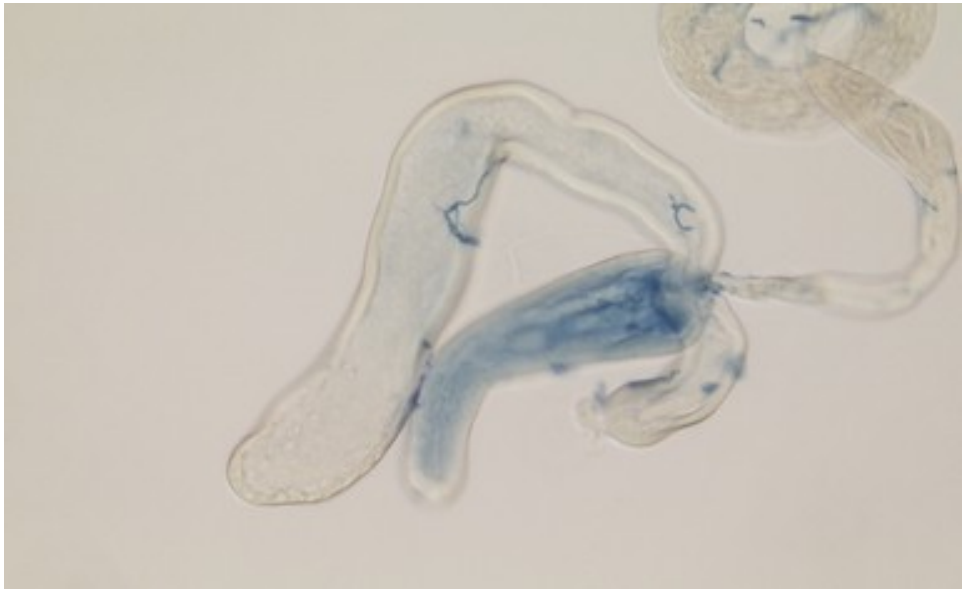


Image annotation

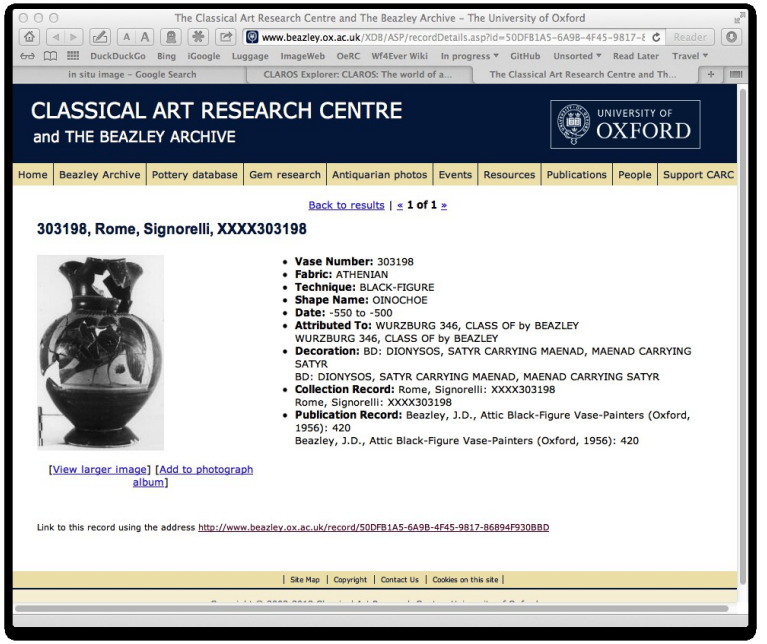
- cf. FlyWeb, Fly-TED

Personal web-research notebook

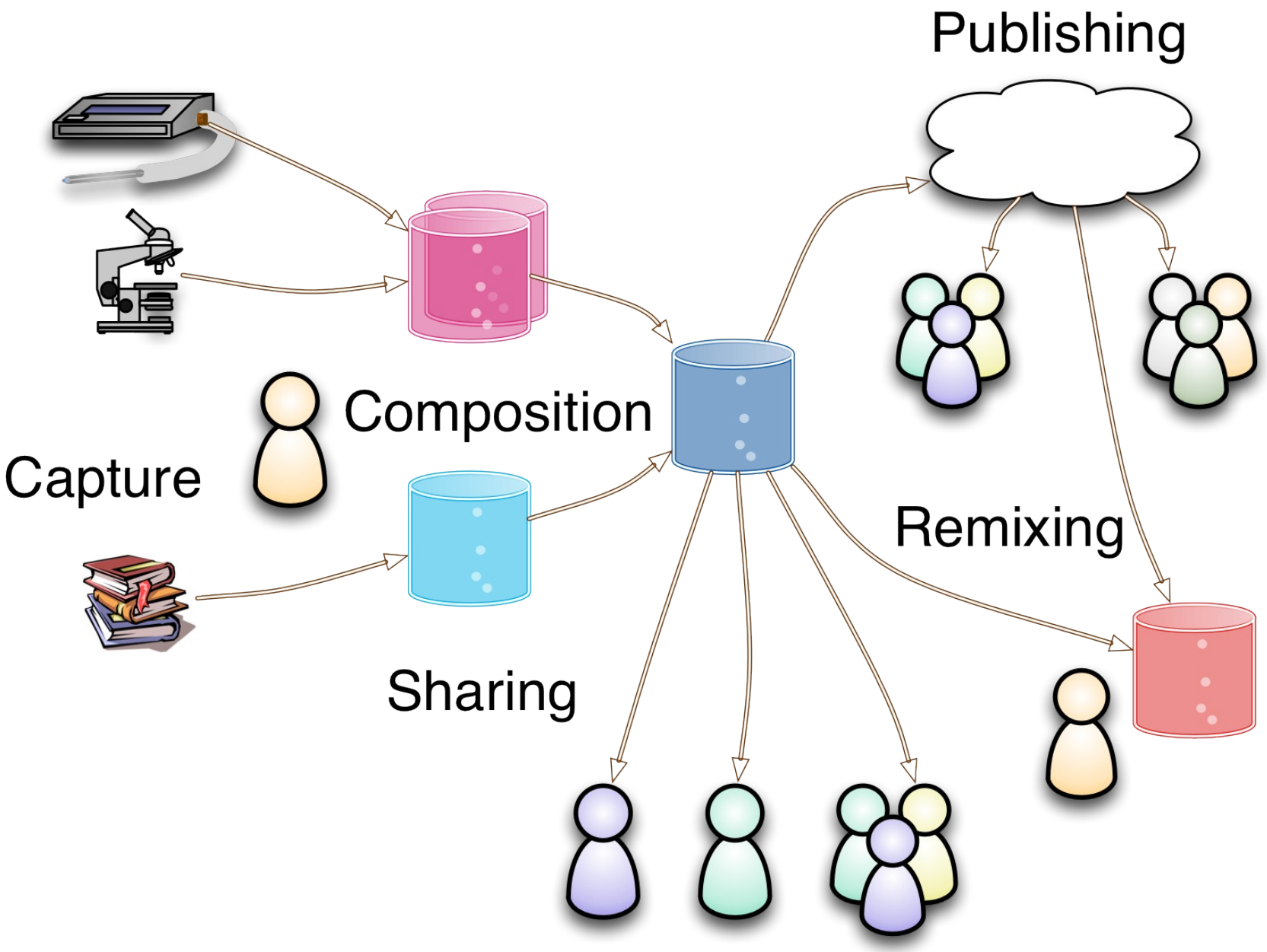
- investigations of CLAROS and related resources

Research Object creation

- aggregated context of an experiment



Common Requirements



Small Research Group Practices

- Practical Issues
  - Data in diverse, incompatible formats
  - Copy-and-paste, or manual transcription
  - Sharing by “sneakernet”, or email
  - Manual format conversion
  - Understanding of data is not guaranteed
- Composition, sharing, publishing and remixing are effort-intensive, error prone processes
  - often with uncertain value of outcome
  - most likely, it doesn't happen

What Tools Are Available?

Spreadsheets: current state of the art?

- widely available and understood

  very commonly used by researchers

- easy to capture data, flexible, easy to share locally

But...

- capturing semantics can be difficult
- composing and remixing is a manual process, or may need custom software development

Semantic web technologies

- appear to have desirable properties
- available tools don't address “first mile” problems

Can We Do Better?

Imagine a tool that combines:

- spreadsheet ease-of-use and flexibility
- semantic technology capabilities for composition and remixing
- web capabilities for sharing and publication

What might such a tool look like? ...

Out-of-box key features

Easy data entry and acquisition

- Fire up and start collecting data

Flexible evolution of data structures

- Add new fields, record types on-the-fly, as required

Controlled sharing of data with collaborators

- Access using standard mechanisms and formats
- Flexible access control

Remixing data with third party sources

- Support for linking in and out (hypermedia)

Additional features

Portable data (just copy)

Working with version management

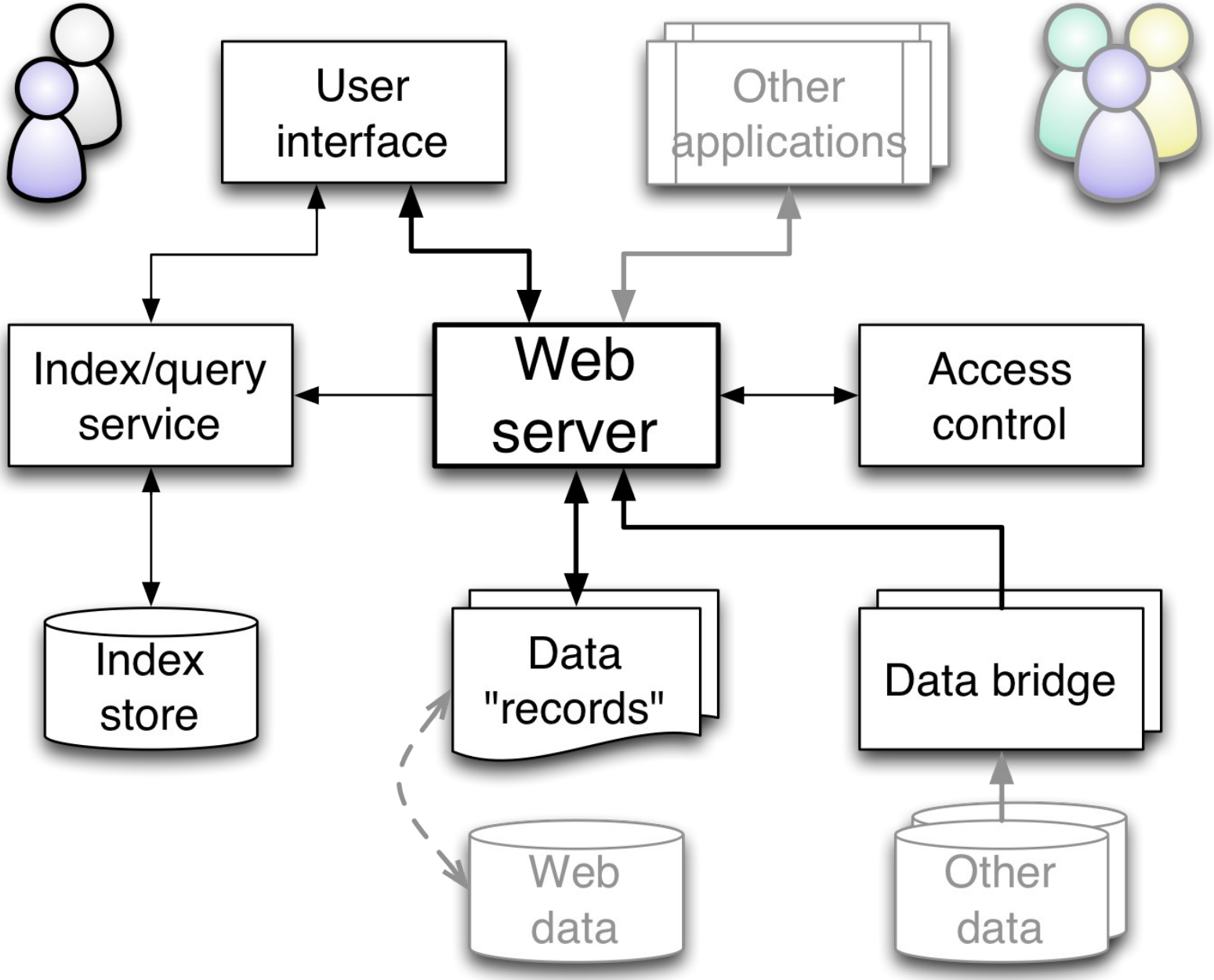
Configuration data is just data (easy replication of complete setup)

Working with pre-existing data (e.g. spreadsheets)

Local *or* cloud hosting of data

Third party authentication (no new passwords or password security concerns)

Proposed System Outline



Proposed Data Record Model

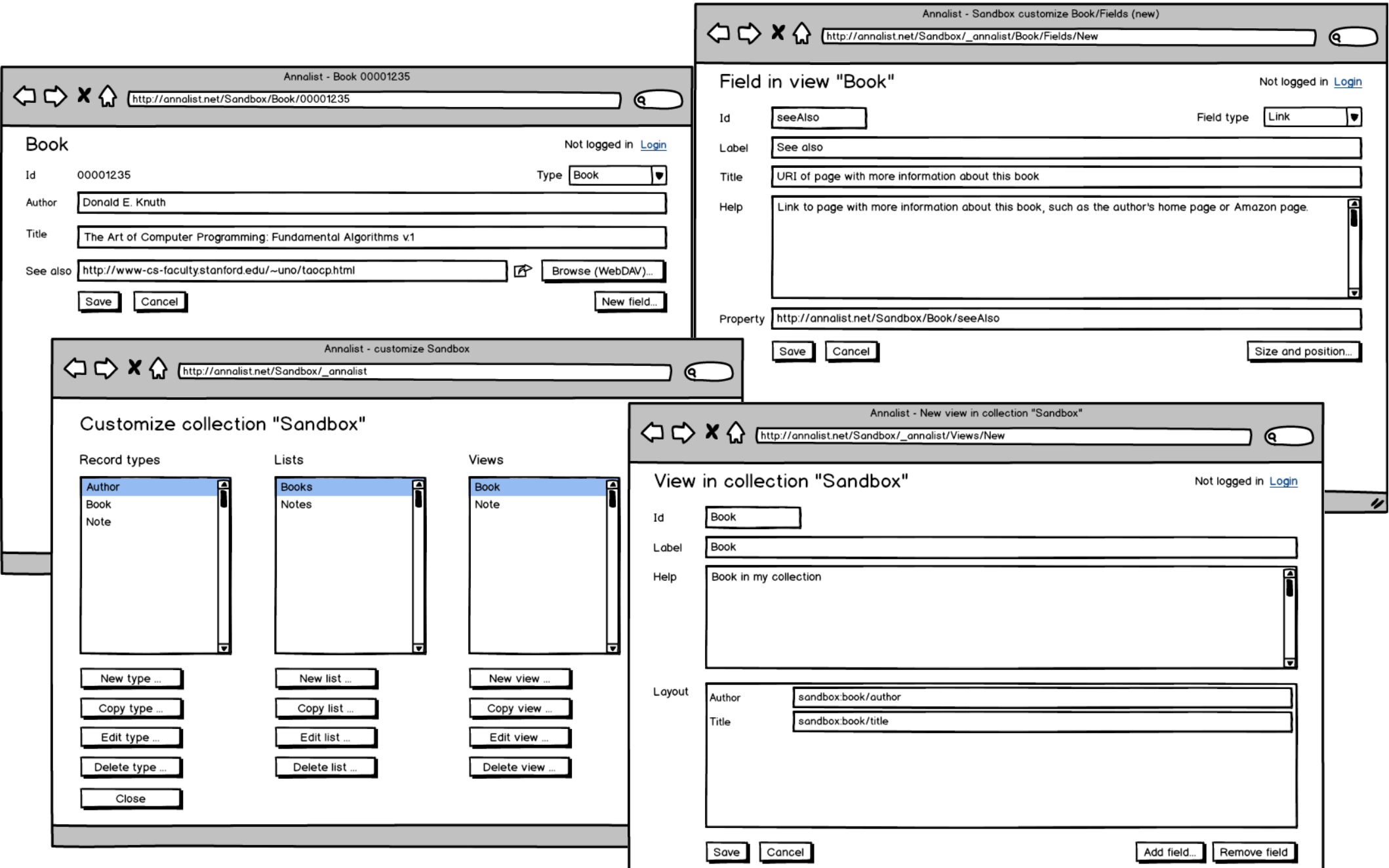
RDF-based format

- Entities carry type information
- Entities can be related by typed links
- No schema constraints

Frame- or entity- oriented records

- A single web resource contains an arbitrary amount of information about some entity
- Fundamental unit of data access

Data Editing User Interface



System Components

- Web server
  - Apache httpd, Nginx, ...
- Indexing service
  - Jena Fuseki, Elastic Search, ...
- Authentication
  - Persona, OpenId Connect, ...
- Data record format
  - JSON-LD, Turtle, ...
- UI toolkit
  - Django, ...

The Story So Far...

Working title: “annalist”

(as in creator of “annals”, or records)

Open source, open development

Github project

- https://github.com/gklyne/annalist
- (no code yet, just vapourware)

… Next Steps

2013-Q4

- Investigate authentication/IDP technologies
- Investigate web server access controls
- Identify potential user collaborations

2014-Q1 onwards

- Pin down data access API details
- Choose web server, indexing engine, etc
- Implement data acquisition/viewing UI
- Implement spreadsheet data bridge
- Work with user(s) to create demo application(s)

Opportunities for collaboration?

When initial demo capability is implemented, I would like to work with one or two active research activities to refine requirements

Develop support services and community to enable wider adoption

Create domain-tailored configurations to support community activities (e.g. MIBBI support, etc.)