Data: The First Mile

or...

“Where does the data come from?”

This talk introduces an open source project
I plan to spool up at the end of this year

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
  - not counting large public databases

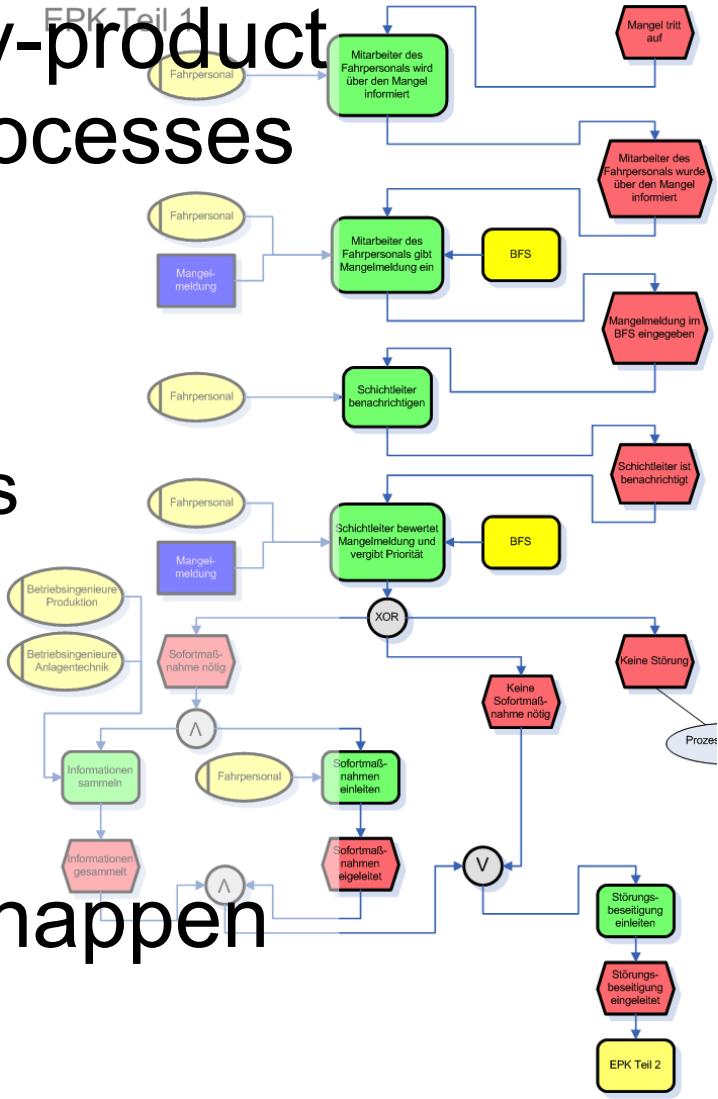
Feeding the “Big Data” Machines

- Much talk about “big data” and “open data”
  - Linked data
  - Government data
  - Data the “new oil”

- But who creates the data?
- Where does the data come from?

Large Organizations

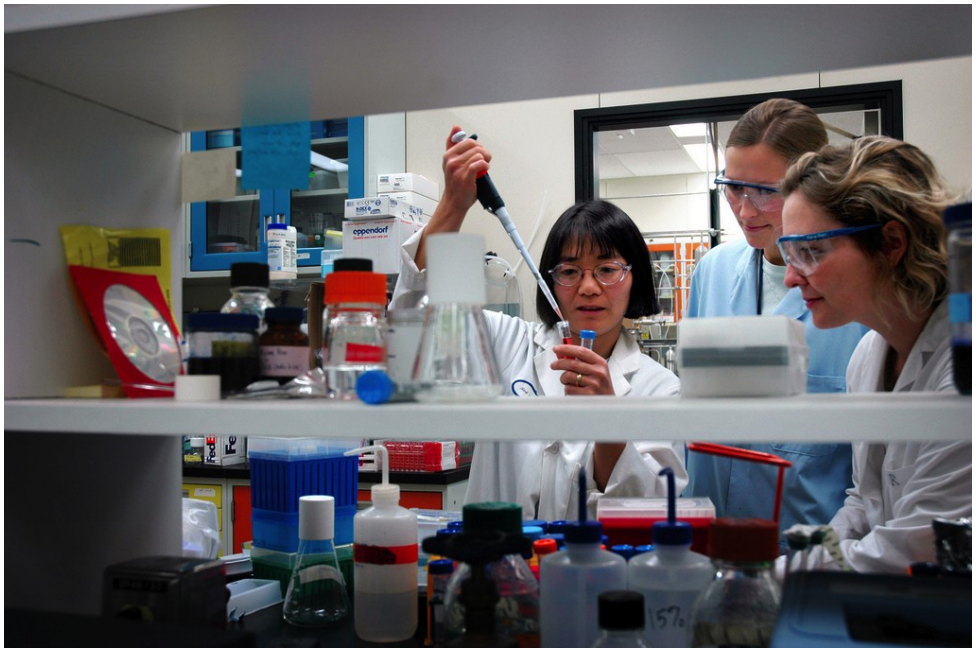
- Data is routinely created as a by-product of established, computerized processes
  - customer transactions
  - supplier transactions
  - (semi-)automated internal process workflows
  - automated monitoring systems
  - *etc*.
- IT departments to make all this happen
- Control over circumstances and methods of data generation



But What About the Little Guys?

- Small research groups, SOHO businesses, freelancers, etc.
- Small groups
  - e.g. 1-5 people
  - Substantially manual processes
  - Working with existing software tools
  - No capability or capacity for custom software development
- Large organizations have small groups too
- The “long tail” of data creation?

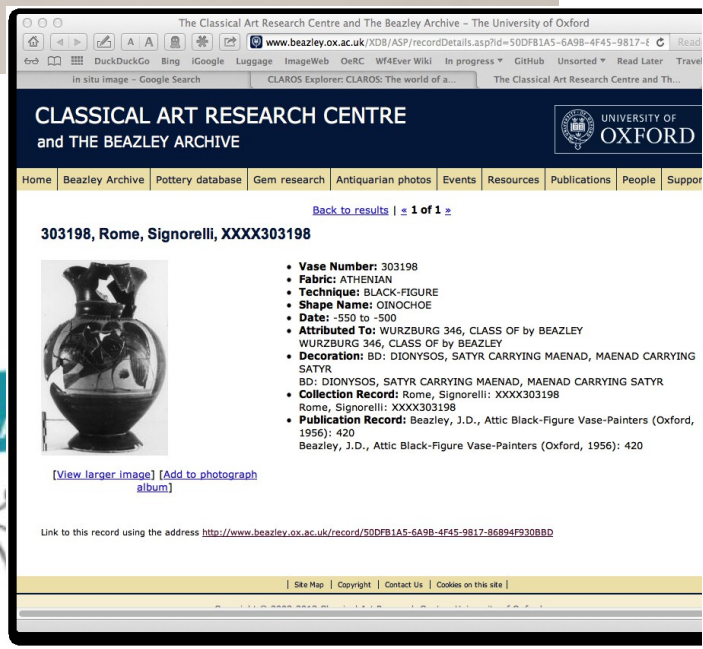
Small Research Group Data

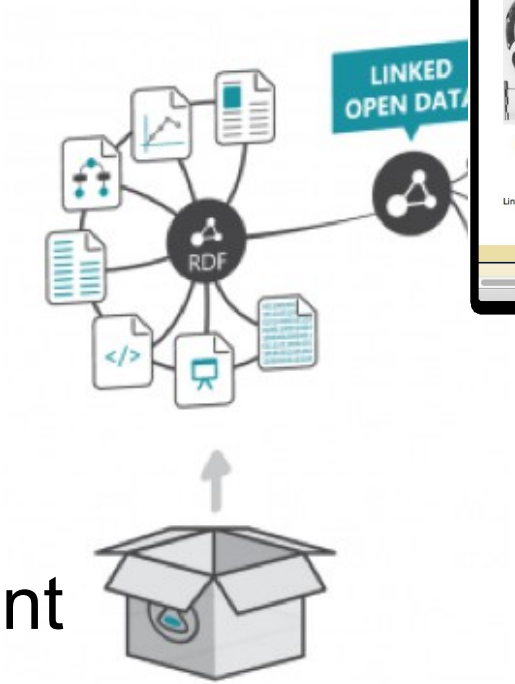


- Data comes from:
  - Hand-written notebooks
  - Spreadsheets
  - Documents (computer text)
  - Stand-alone software tools
  - Instruments
    - not necessarily networked
  - Web sites and online reference
- Local data is connected to global databases

Some Examples

- Image annotation
  - cf. FlyWeb, Fly-TED
- Personal web-research notebook
  - investigations of CLAROS and related resources
- Research Object creation
  - aggregating resources related to an experiment





Some Common Requirements

- Composition
  - comparing or combining data from diverse sources
- Sharing
  - selectively exposing data to collaborators
- Publishing
  - making selected data publicly available
- Remixing
  - connecting with third party data, often for new uses not originally envisaged

Small Research Group Challenges

- 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
- 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?

- I am contemplating 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? ...

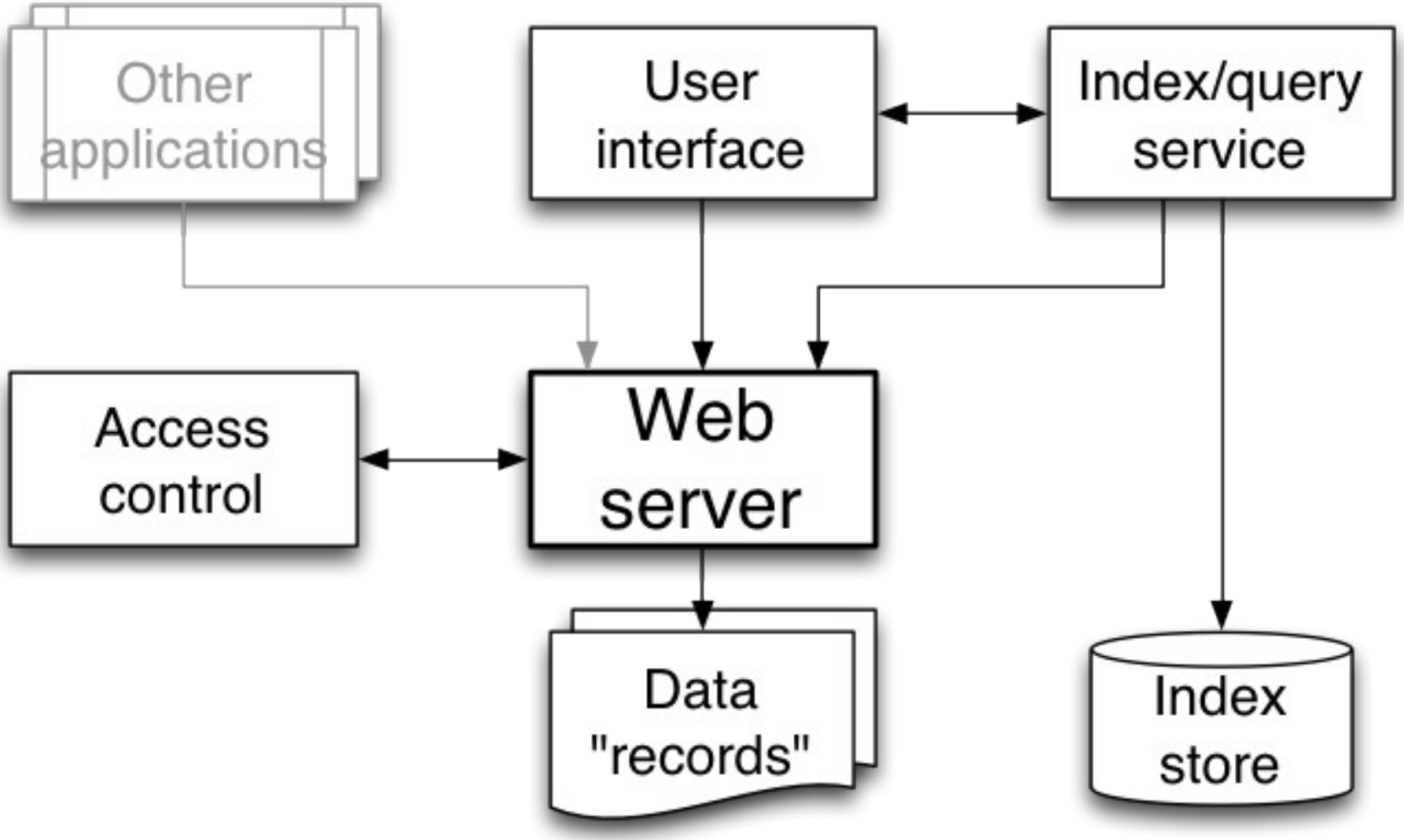
Out-of-box features

- 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
  - Use standard web file format
  - Expose using standard web mechanisms
  - Access control
- Remixing data with third party sources
  - Support for linking in and out (hypermedia)

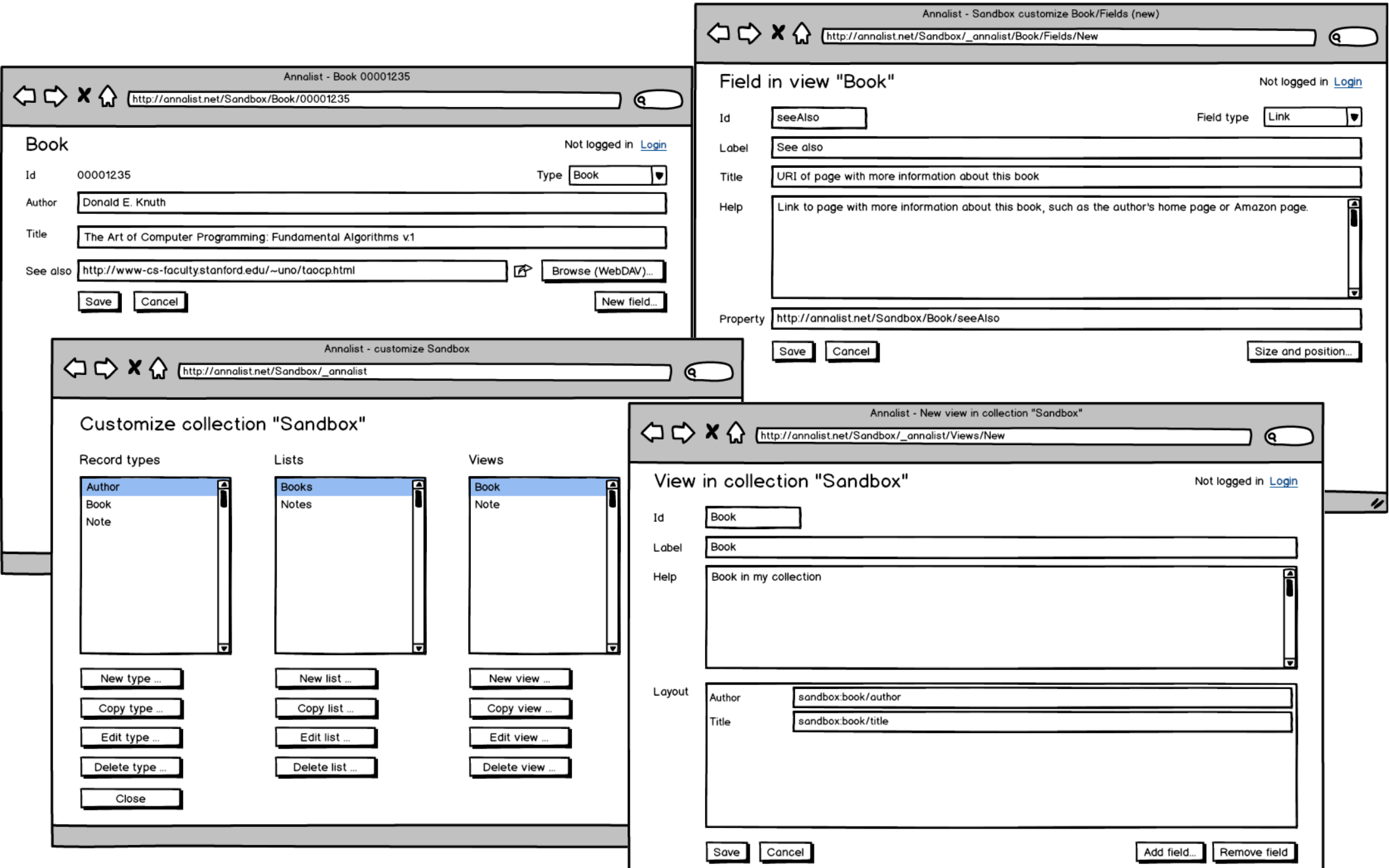
Out-of-box features

- 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
  - Use standard web file format
  - Expose using standard web mechanisms
  - Access control
- Remixing data with third party sources
  - Support for linking in and out (hypermedia)

Proposed System Outline



Data Editing User Interface



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

System Components

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

The Story So Far...

- Working title: “annalist”

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

(pause for breath)

The end