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

Old Mother Hubbard, she went to the cupboard
To get her poor dog (Uncle Sam) a bone;
But when she got there the cupboard was bare,
And so the poor dog had none.

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?

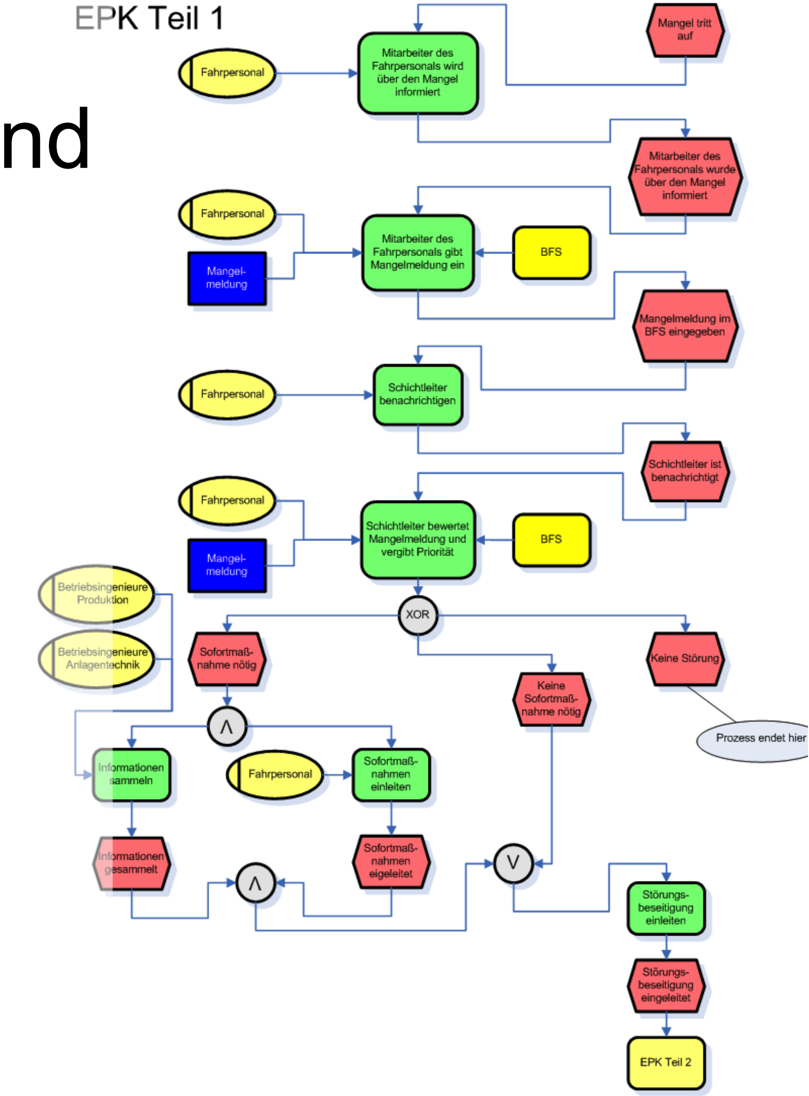
- Where does the data come from?

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


EPK Teil 1
Fahrpersonal
Mitarbeiter des Fahrpersonals wird über den Mangel informiert
Mangel tritt auf
Mitarbeiter des Fahrpersonals wurde über den Mangel informiert
Fahrpersonal
Mangel-meldung
Mitarbeiter des Fahrpersonals gibt Mangelmeldung ein
BFS
Mangelmeldung im BFS eingegeben
Fahrpersonal
Schichtleiter benachrichtigen
Schichtleiter ist benachrichtigt
Fahrpersonal
Mangel-meldung
Schichtleiter bewertet Mangelmeldung und vergibt Priorität
BFS
Betriebsingenieure Produktion
Betriebsingenieure Anlagentechnik
XOR
Sofortmaß-nahme nötig
Keine Sofortmaß-nahme nötig
Keine Störung
Prozess endet hier
∧
Informationen sammeln
Fahrpersonal
Sofortmaß-nahmen einleiten
Informationen gesammelt
Sofortmaß-nahmen eingeleitet
∧
V
Störungs-beseitigung einleiten
Störungs-beseitigung eingeleitet
EPK Teil 2

Subject and Other Databases

Examples:

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

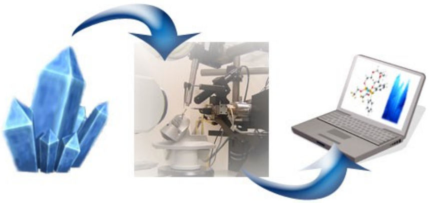
Separately funded

Often curated

Economies of scale?

Community portals, some with recognized academic value

Again: ***Dedicated IT support***



But What About the Little Guys?

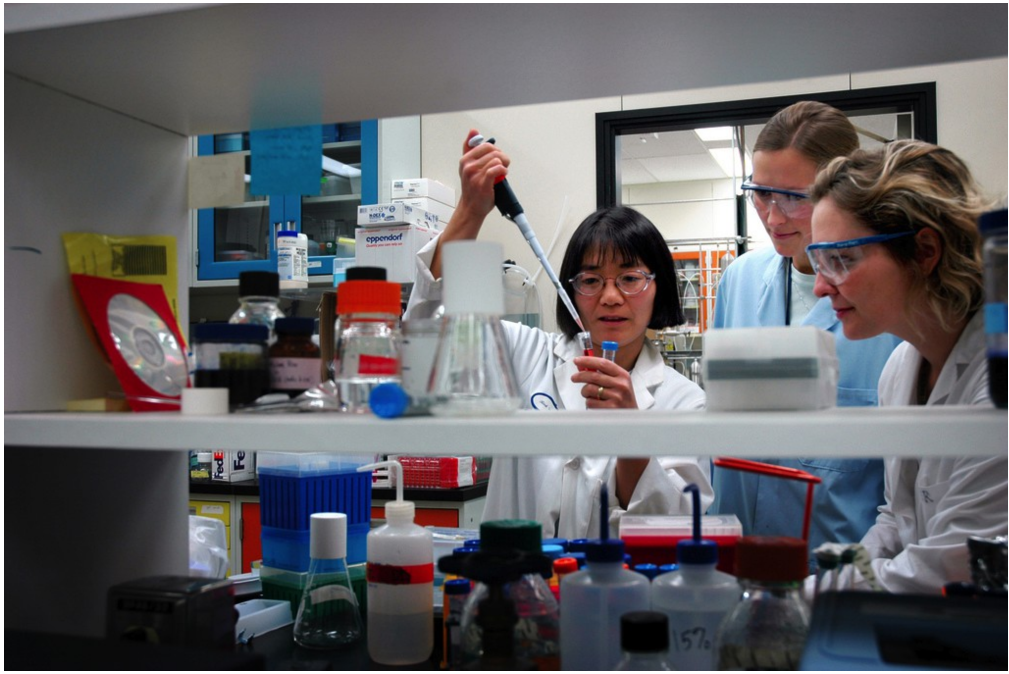
Small research groups

- 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



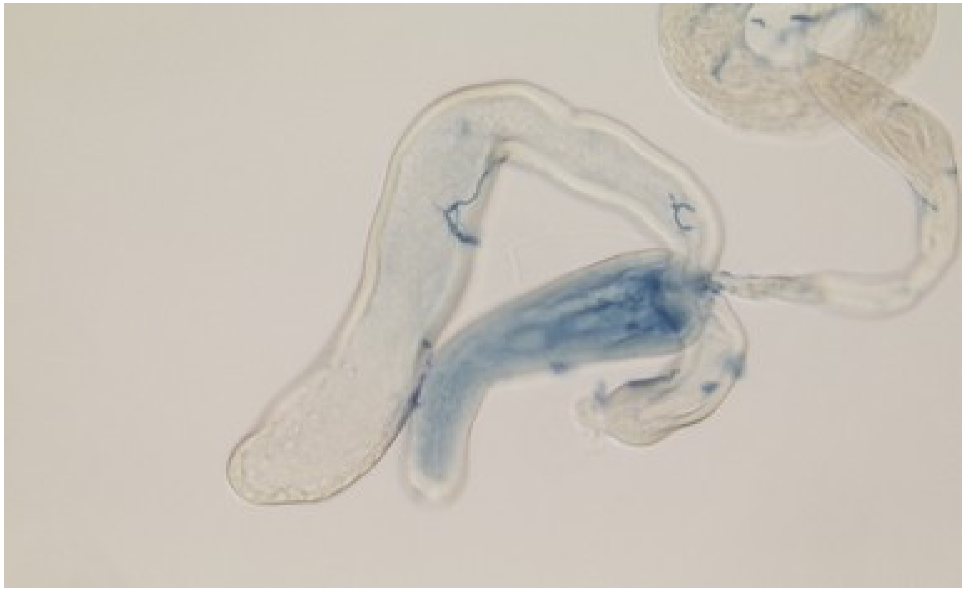
Data comes from:

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

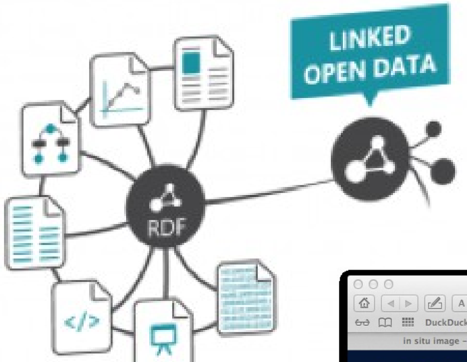
  not necessarily networked
- Stand-alone software tools
- Web sites and online reference

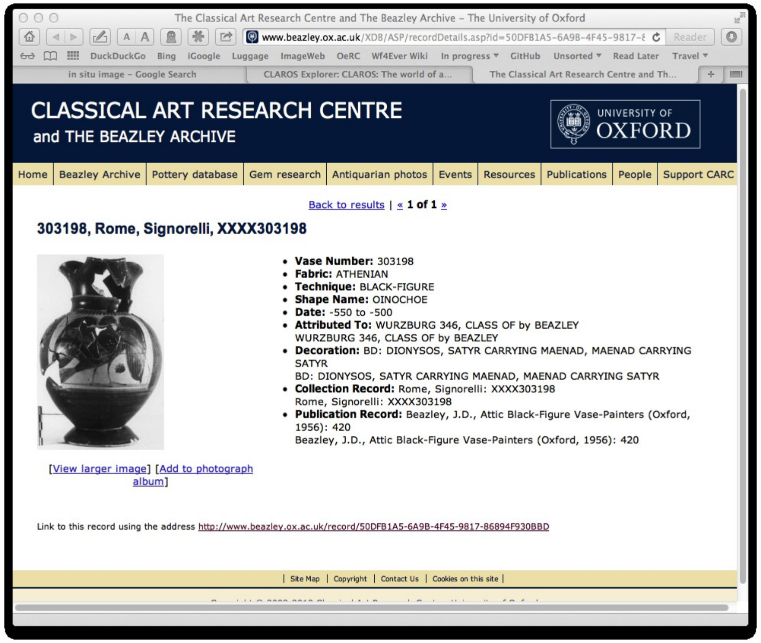
Local *ad hoc* connection to global databases

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




The Classical Art Research Centre and The Beazley Archive – The University of Oxford
www.beazley.ox.ac.uk/XDB/ASP/recordDetails.asp?id=50DFB1A5-6A9B-4F45-9817-8
Reader
DuckDuckGo
Bing
iGoogle
Luggage
ImageWeb
OeRC
Wf4Ever Wiki
In progress
GitHub
Unsorted
Read Later
Travel
in situ image – Google Search
CLAROS Explorer: CLAROS: The world of a...
The Classical Art Research Centre and Th...
UNIVERSITY OF OXFORD
Home
Beazley Archive
Pottery database
Gem research
Antiquarian photos
Events
Resources
Publications
People
Support CARC

CLASSICAL ART RESEARCH CENTRE
and THE BEAZLEY ARCHIVE

Back to results | « **1 of 1** »

303198, Rome, Signorelli, XXXX303198

- **Vase Number:** 303198
- **Fabric:** ATHENIAN
- **Technique:** BLACK-FIGURE
- **Shape Name:** OINOCHOE
- **Date:** -550 to -500
- **Attributed To:** WURZBURG 346, CLASS OF by BEAZLEY
  WURZBURG 346, CLASS OF by BEAZLEY
- **Decoration:** BD: DIONYSOS, SATYR CARRYING MAENAD, MAENAD CARRYING SATYR
  BD: DIONYSOS, SATYR CARRYING MAENAD, MAENAD CARRYING SATYR
- **Collection Record:** Rome, Signorelli: XXXX303198
  Rome, Signorelli: XXXX303198
- **Publication Record:** Beazley, J.D., Attic Black-Figure Vase-Painters (Oxford, 1956): 420
  Beazley, J.D., Attic Black-Figure Vase-Painters (Oxford, 1956): 420

[View larger image] [Add to photograph album]

Link to this record using the address http://www.beazley.ox.ac.uk/record/50DFB1A5-6A9B-4F45-9817-86894F930BBD

| Site Map | Copyright | Contact Us | Cookies on this site |

Common research requirements @@use image

Capturing data and metadata

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 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

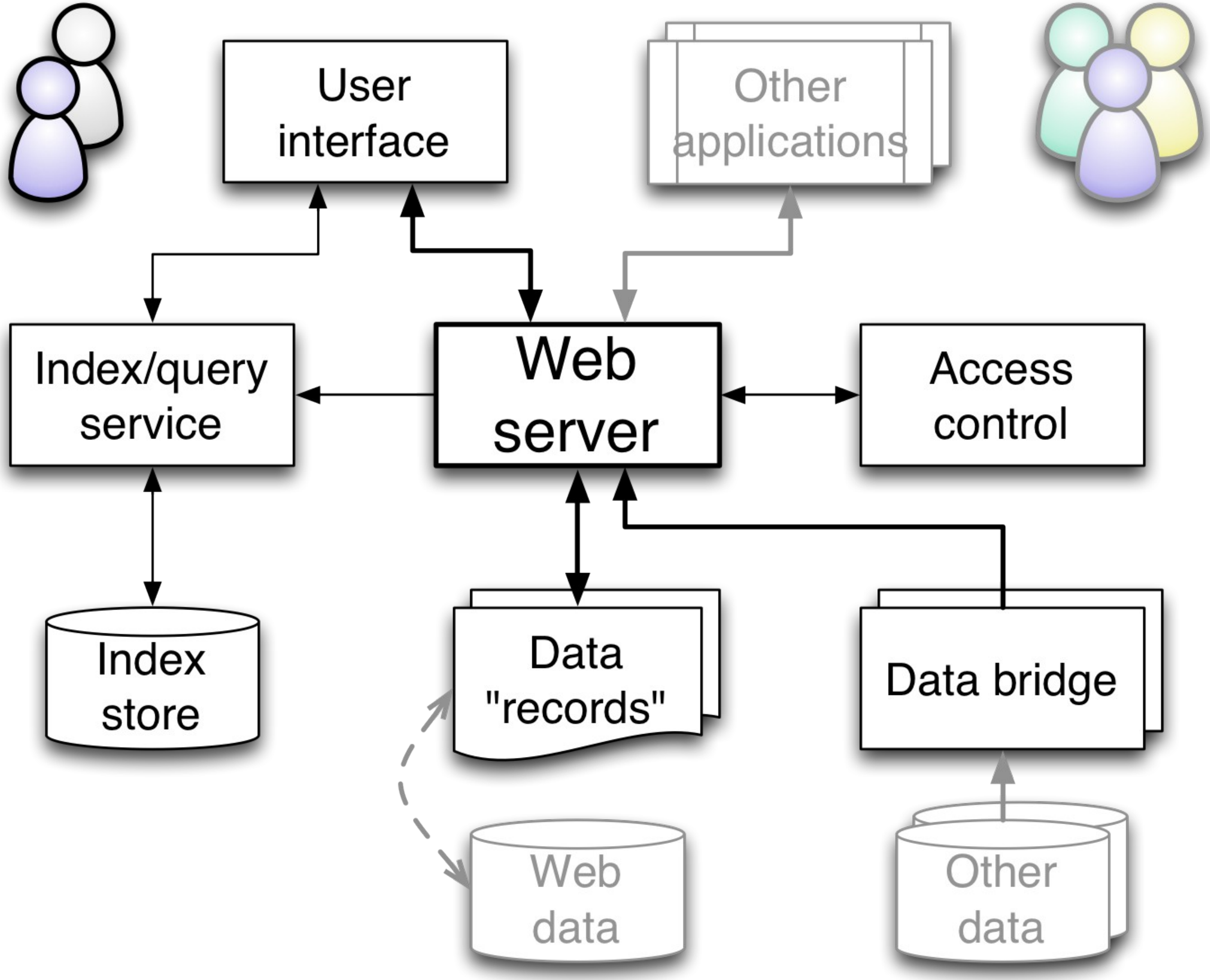
Configuration data is just data (easy replication of complete setup)

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

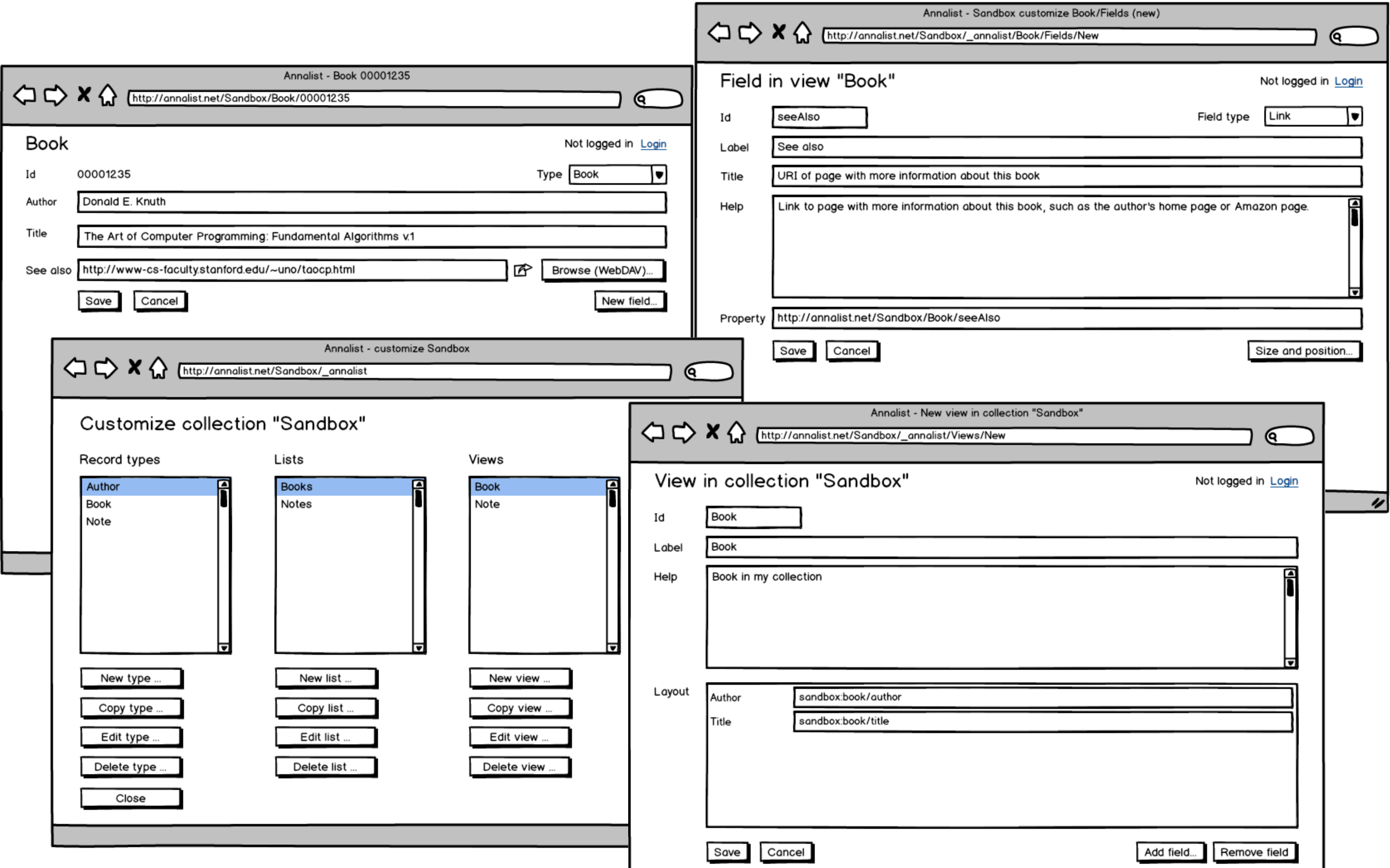
RDF-based format

- 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



Annalist - Book 00001235

http://annalist.net/Sandbox/Book/00001235

Book

Not logged in Login

Id	00001235
Type	Book
Author	Donald E. Knuth
Title	The Art of Computer Programming: Fundamental Algorithms v.1
See also	http://www-cs-faculty.stanford.edu/~uno/taocp.html

Annalist - Sandbox customize Book/Fields (new)

http://annalist.net/Sandbox/_annalist/Book/Fields/New

Field in view "Book"

Not logged in Login

Id	seeAlso
Field type	Link
Label	See also
Title	URI of page with more information about this book
Help	Link to page with more information about this book, such as the author's home page or Amazon page.
Property	http://annalist.net/Sandbox/Book/seeAlso

Annalist - customize Sandbox

http://annalist.net/Sandbox/_annalist

Customize collection "Sandbox"

Record types	Lists	Views
Author	Books	Book
Book	Notes	Note
Note		

Annalist - New view in collection "Sandbox"

http://annalist.net/Sandbox/_annalist/Views/New

View in collection "Sandbox"

Not logged in Login

Id	Book
Label	Book
Help	Book in my collection

Layout

Author	sandbox:book/author
Title	sandbox:book/title

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.)