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

Judge
U. S. TREASURY
$600,000,000 TREASURY DEFICIT
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?

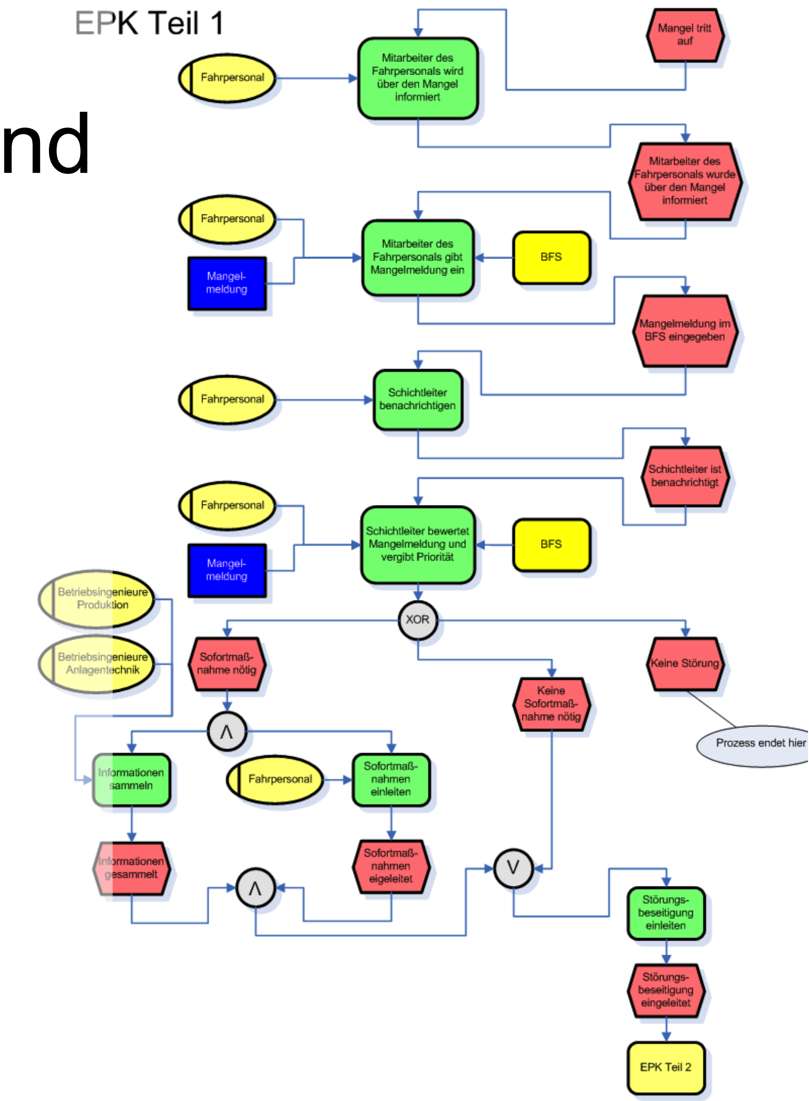
- Where does the data come from?

Large Research Projects

For large research projects, data management is planned and funded

The circumstances and methods of data generation are defined and managed

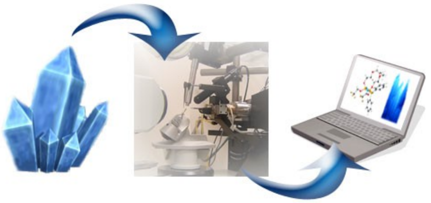
Dedicated IT support helps to make data acquisition, sharing and publication a reality


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 Störung
Keine Sofortmaß-nahme nötig
Prozess endet hier
Λ
Informationen sammeln
Fahrpersonal
Sofortmaß-nahmen einleiten
Informationen gesammelt
Λ
Sofortmaß-nahmen eigeleitet
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?

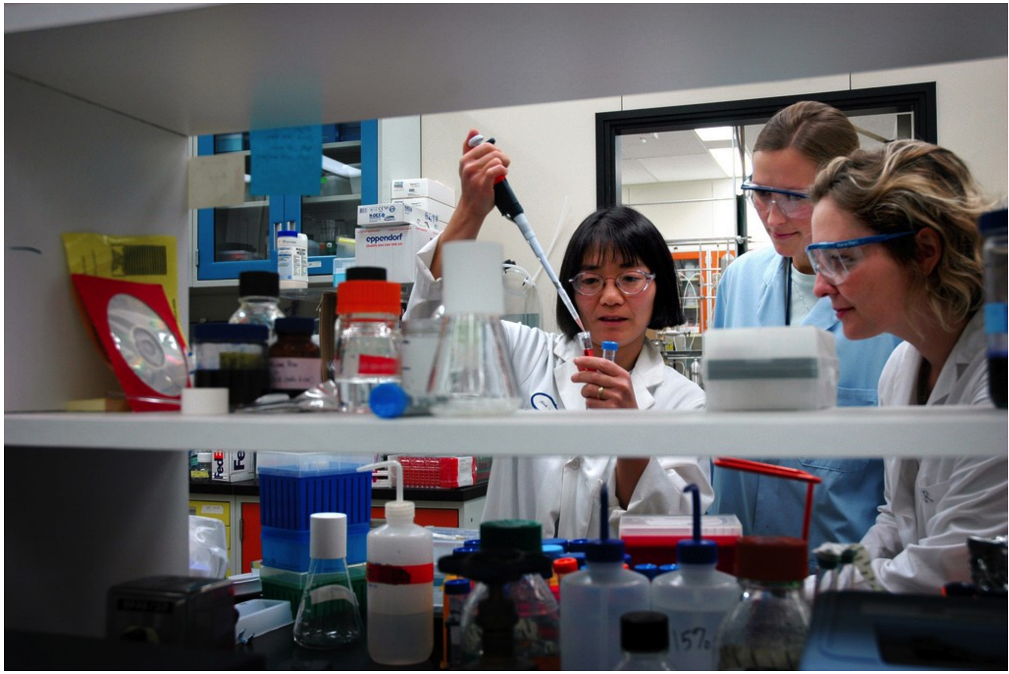
Small research groups

- 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

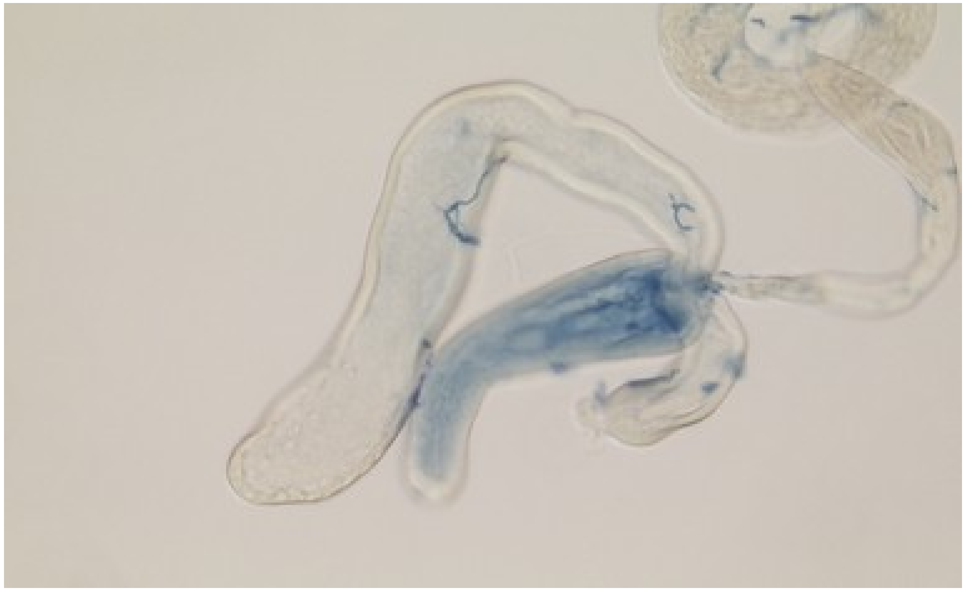


Image annotation

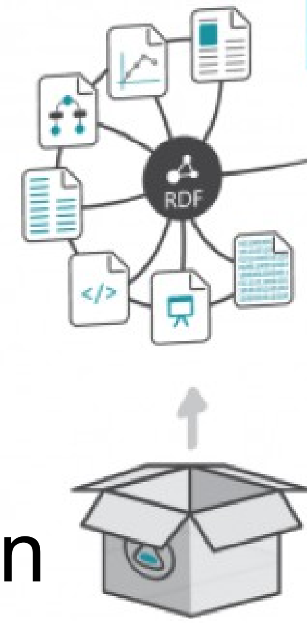
- cf. FlyWeb, Fly-TED

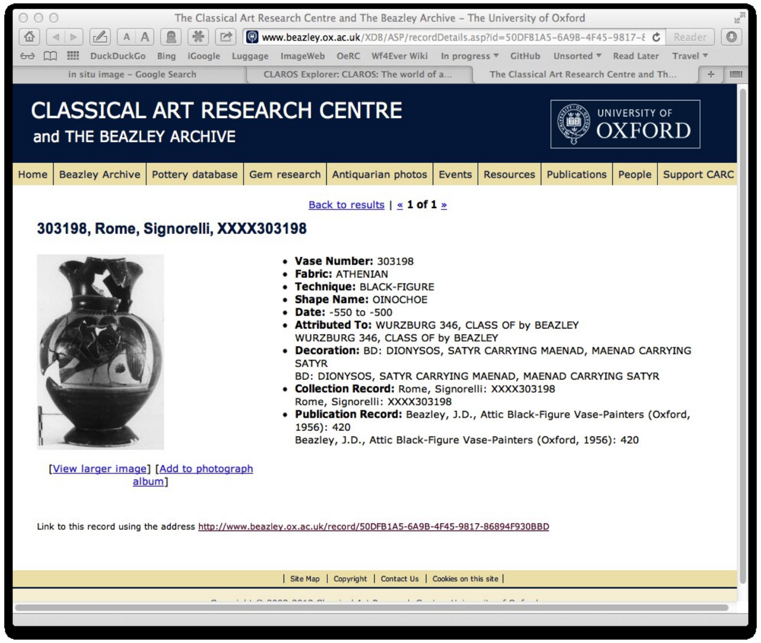
Personal web-research notebook

- 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...

CLASSICAL ART RESEARCH CENTRE
and THE BEAZLEY ARCHIVE

UNIVERSITY OF OXFORD

Home | Beazley Archive | Pottery database | Gem research | Antiquarian photos | Events | Resources | Publications | People | Support CARC

Back to results | « 1 of 1 »

303198, Rome, Signorelli, XXXX303198

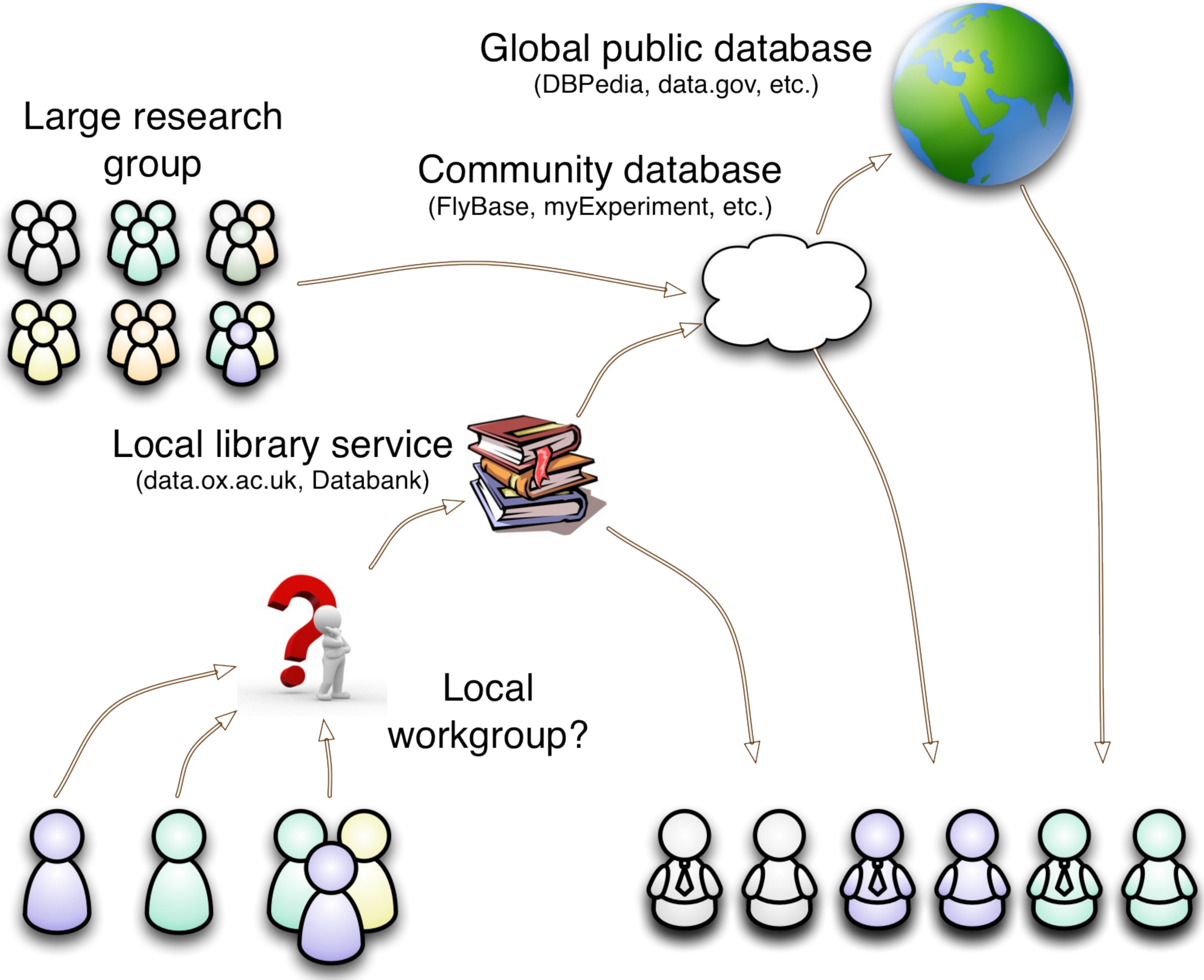
- **Vase Number:** 303198
- **Fabric:** ATHENIAN
- **Technique:** BLACK-FIGURE
- **Shape Name:** OINOCHOE
- **Date:** -550 to -500
- **Attributed To:** WURZBURG 346, CLASS OF by BEAZLEY
  WURZBURG 346, CLASS OF by BEAZLEY
- **Decoration:** BD: DIONYSOS, SATYR CARRYING MAENAD, MAENAD CARRYING SATYR
  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 |

A Missing Link



Global public database
(DBPedia, data.gov, etc.)

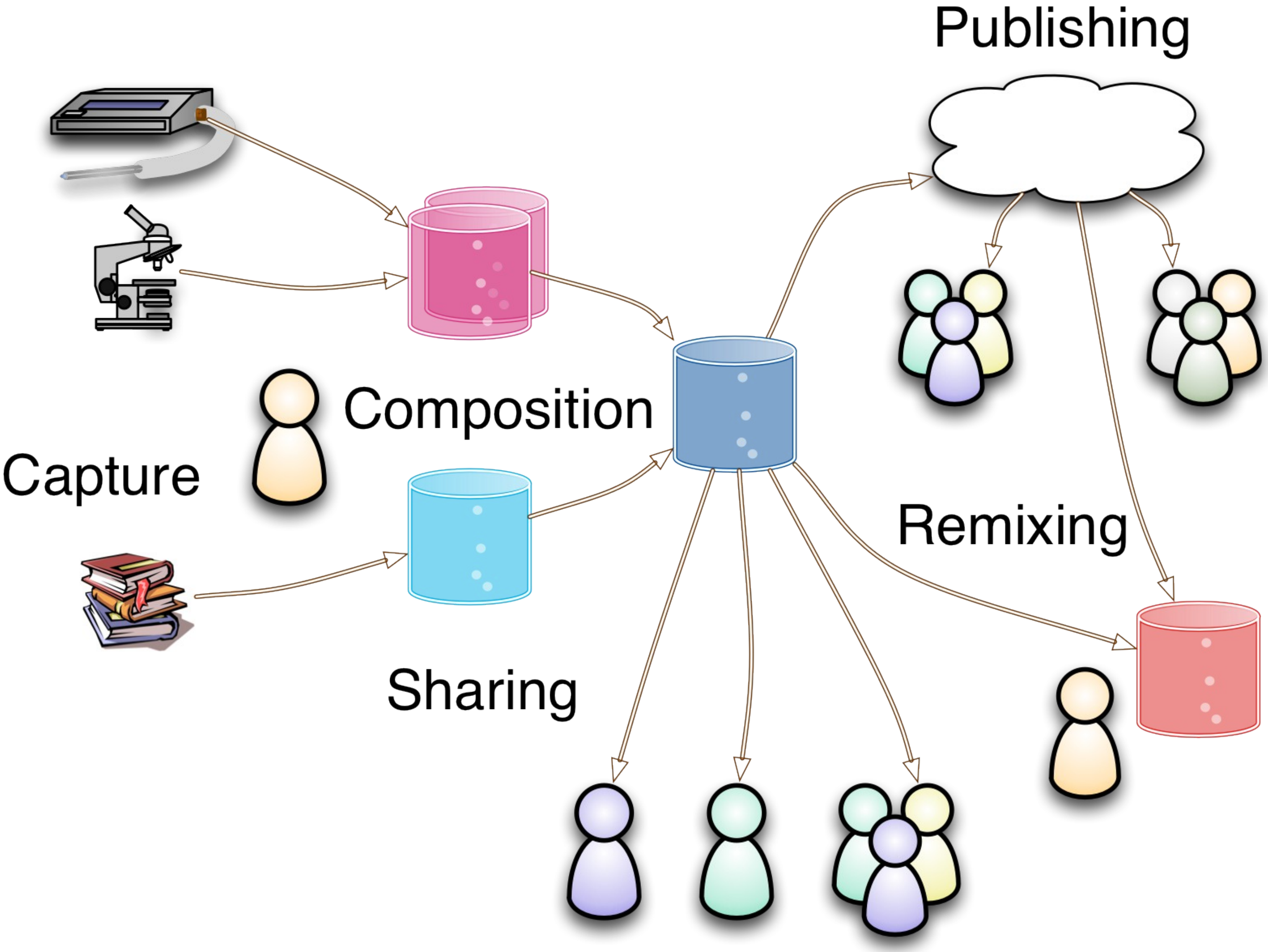
Large research group

Community database
(FlyBase, myExperiment, etc.)

Local library service
(data.ox.ac.uk, Databank)

Local workgroup?

Common Requirements



Publishing

Composition

Capture

Remixing

Sharing

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

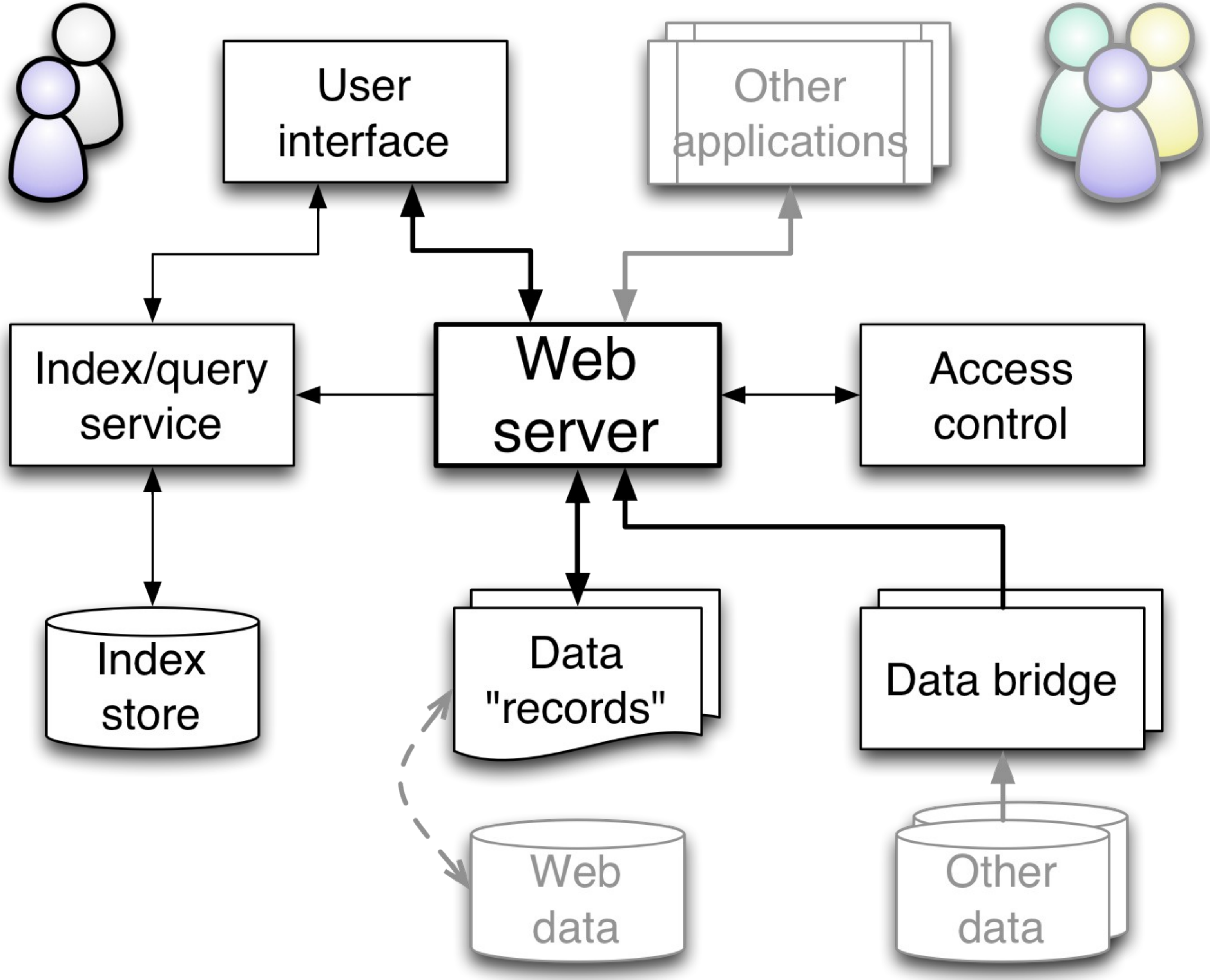
Configuration data is just data (easy replication of complete setup)

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

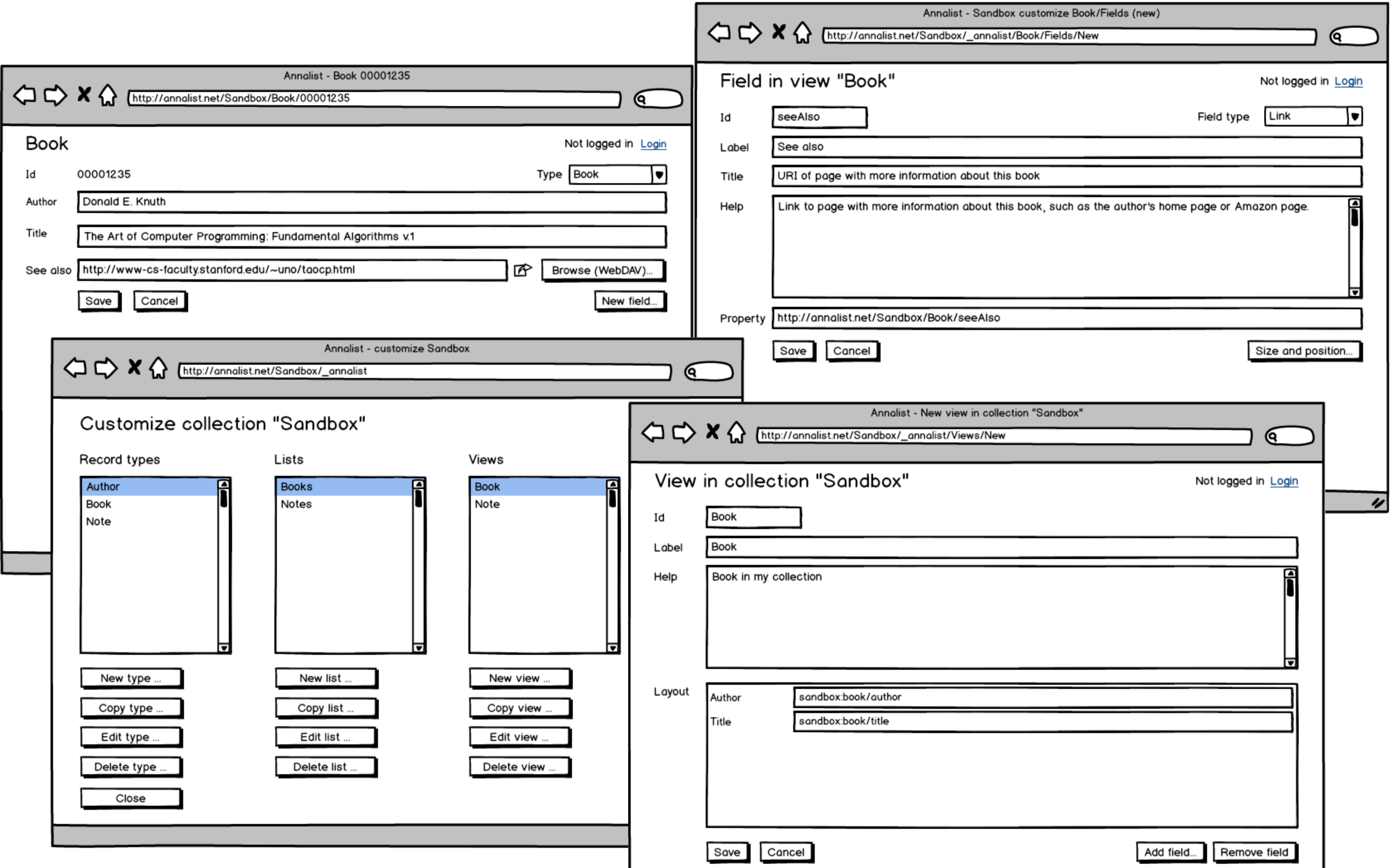
RDF-based format

- 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 and access control
  - Persona, OpenId Connect, UMA, ...
- 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

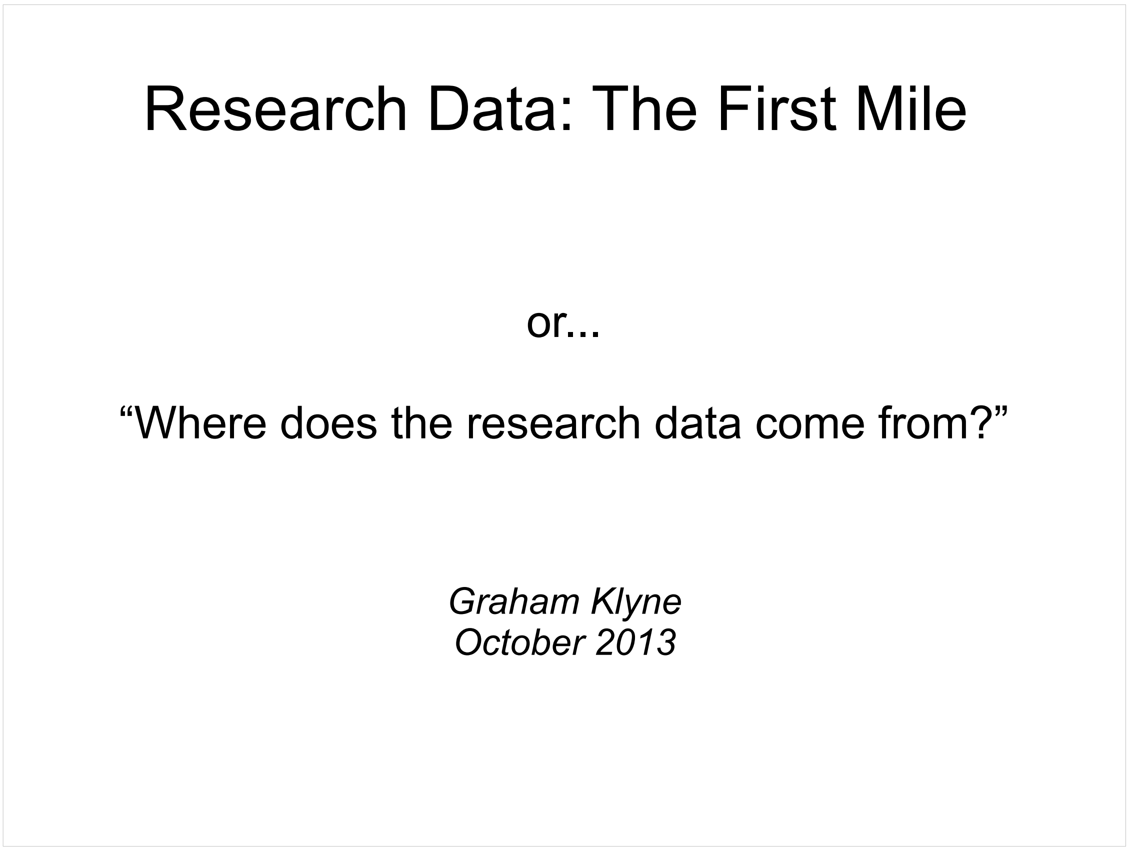
- Pin down data access API details
- Choose web server, indexing engine, etc
- Implement data acquisition/viewing UI
- Implement spreadsheet data bridge
- Work with user(s) to create demo application(s)

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.)



Research Data: The First Mile

or...

“Where does the research data come from?”

Graham Klyne
October 2013

Background: Empty Data Archives

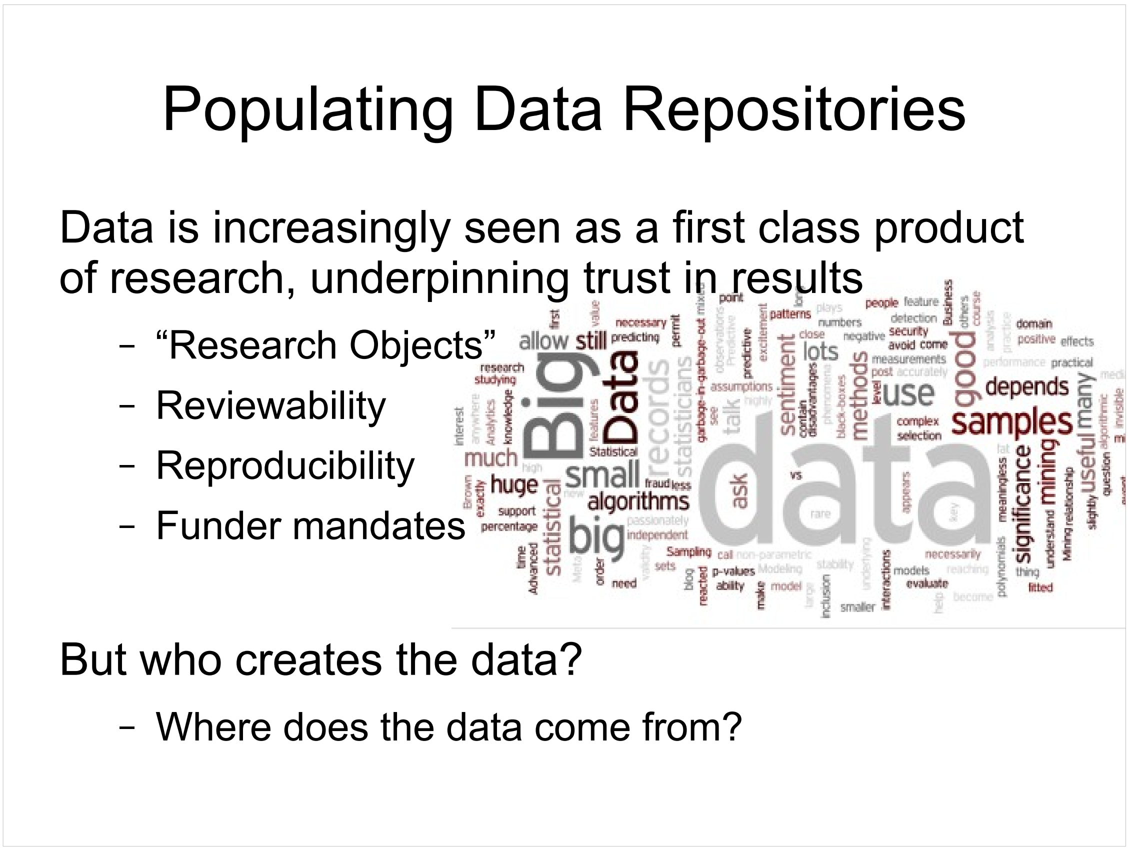
- I have for several years been working on projects related to research data sharing
- Repositories have been created for data storage and publication
- But there is not (yet) much data in them
  - (or not as much as there should be)
  - not counting large public databases

Judge
U. S. TREASURY
$400,000,000 TREASURY DEFICIT
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.

The background that motivates this development is in my experience of working with systems for research data management, sharing and publication.

A fair amount of effort has gone into creating repositories and other supporting infrastructure for researchers to share and publish data, but they are not as well used as they should be, especially by individual and small group projects.

(I don't include community databases to which researchers can submit data, such as genomics and chemical crystallography community facilities for specified data.)



Populating Data Repositories

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

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

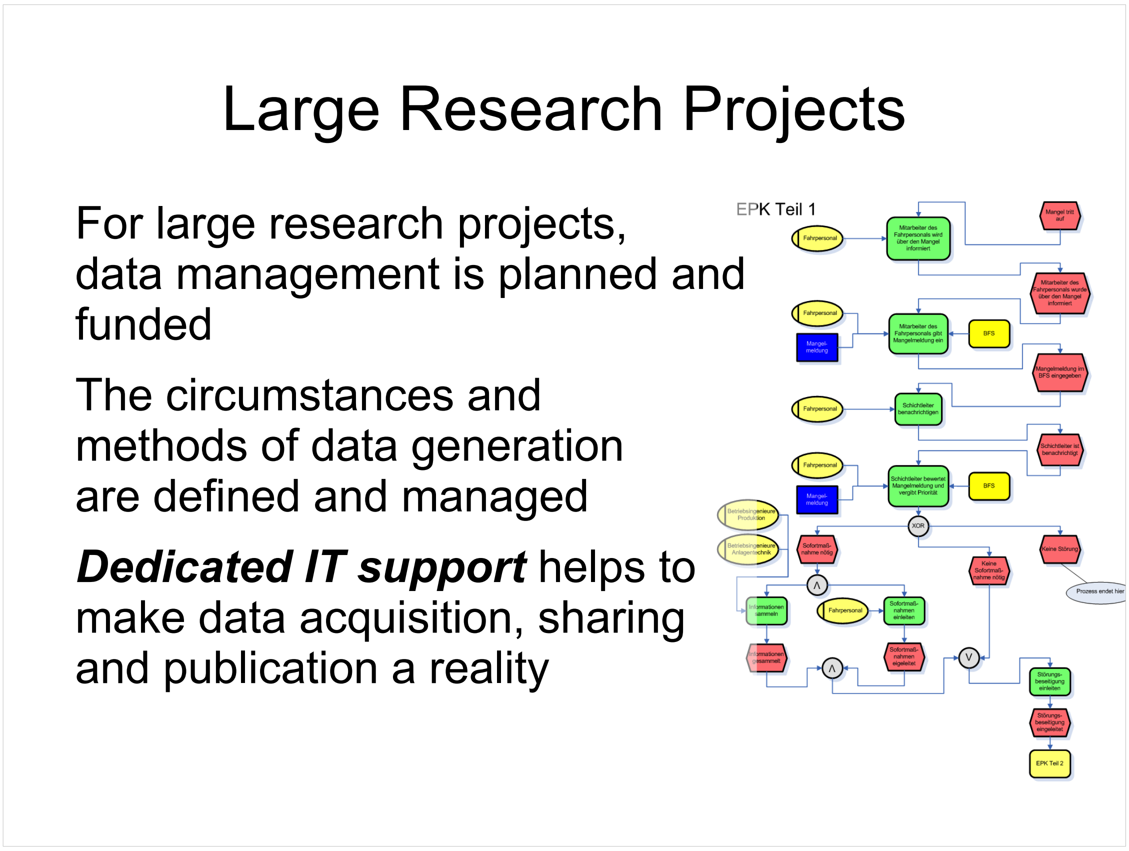
But who creates the data?

- Where does the data come from?

Data is increasingly see as a first class research output, and funders and others are requiring that data produced by funded research is made public.

Also, community expectations are increasingly (if not yet overwhelmingly) that data should be published to support replication and reproducibility of results.

But who is to create this data, and ensure that it is fit for re-examination by other researchers?



Large Research Projects

For large research projects, data management is planned and funded

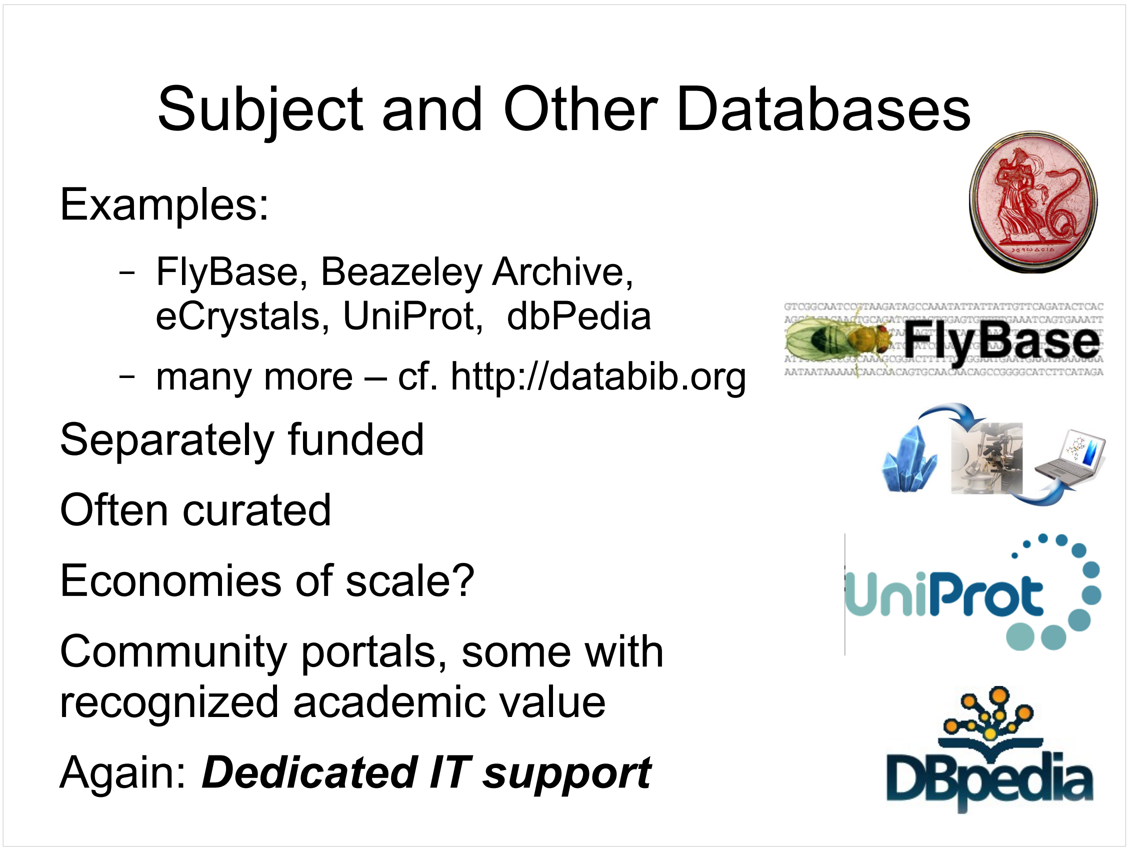
The circumstances and methods of data generation are defined and managed

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
Λ
Informationen sammeln
Fahrpersonal
Sofortmaß-nahmen einleiten
Informationen gesammelt
Λ
Sofortmaß-nahmen eigeleitet
V
Störungs-beseitigung einleiten
Störungs-beseitigung eingeleitet
EPK Teil 2

For large research projects, data management is generally planned and funded as part of the project.

To the extent needed, dedicated IT support resources may be brought to bear to ensure the data management plans can be followed.



Subject and Other Databases

Examples:

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

Separately funded

Often curated

Economies of scale?

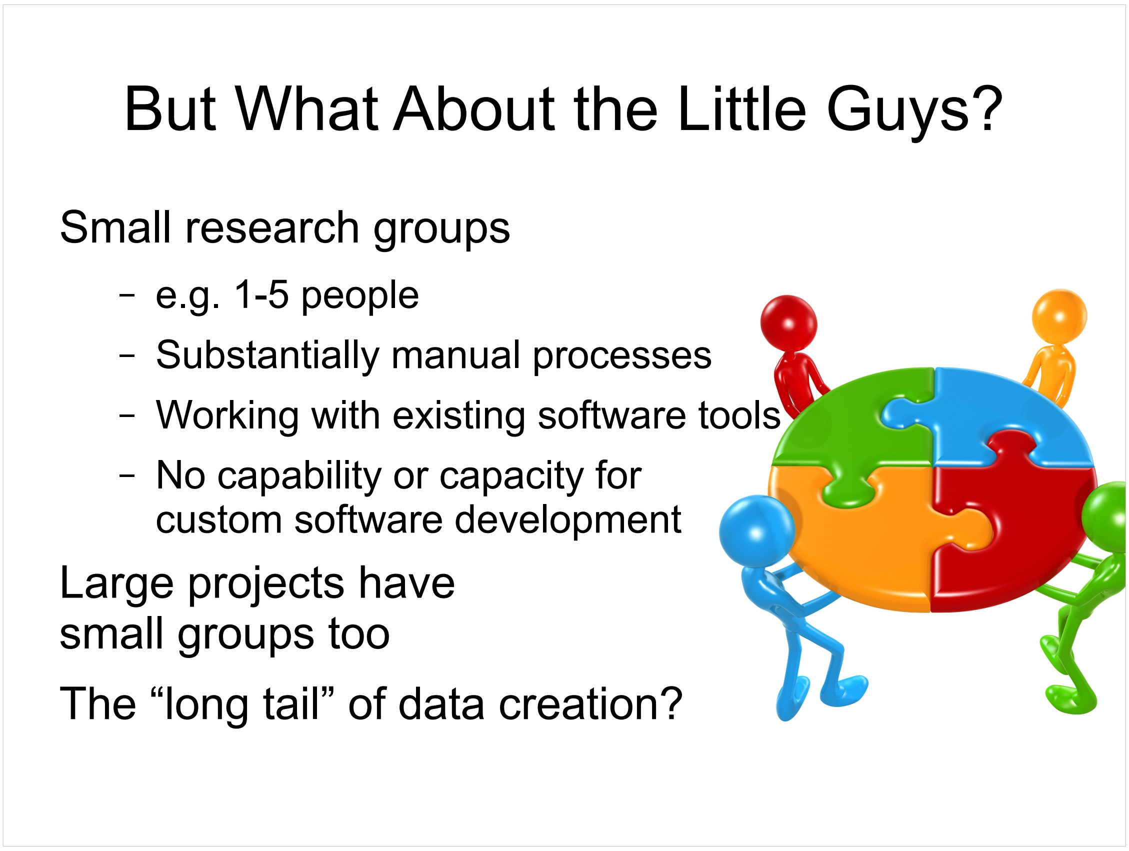
Community portals, some with recognized academic value

Again: ***Dedicated IT support***

FlyBase
UniProt
DBpedia

In some research communities, there are public databases, often separately funded and curated, to which researchers can contribute. But these are usually for specific kinds of data, and are not a complete data management solution for new research projects.

Again, these databases are typically backed by some form of dedicated IT development and support capabilities.



But What About the Little Guys?

Small research groups

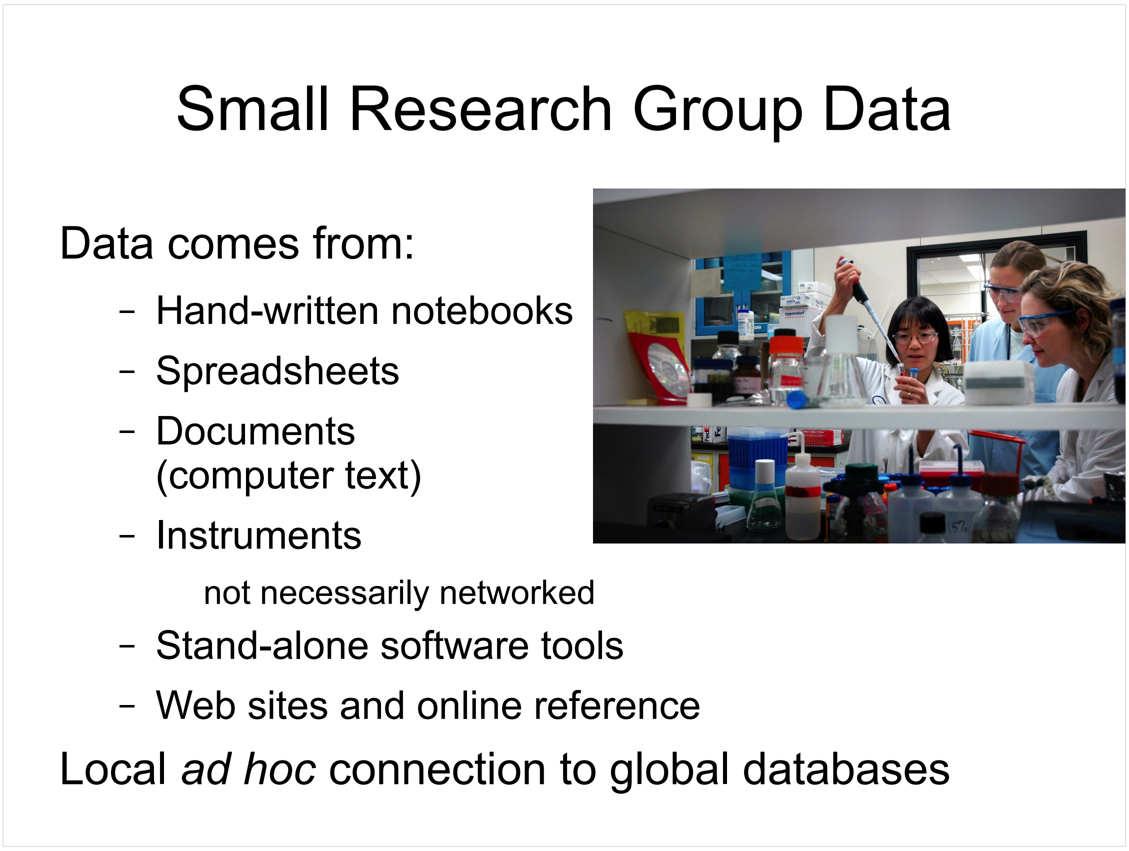
- e.g. 1-5 people
- Substantially manual processes
- Working with existing software tools
- No capability or capacity for custom software development

Large projects have small groups too

The “long tail” of data creation?

But for small research groups, who arguably may represent a “long tail” of data creation, there is no dedicated IT support or development capability. Thus, they have to work with existing data management tools.

(And even within large projects, there may be smaller groups creating data that don't fit the data management patterns of the overall project.)



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

Small-group data may be obtained from a number of sources, and stored in a variety of ways.

Data collection, integration and management may be somewhat less than fully automated.

Researchers ideas about what they want to do with the data may change as the data is collected, responding to what they are seeing in it.

Some Applications

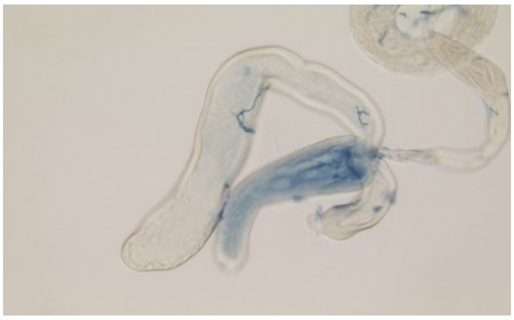


Image annotation

- cf. FlyWeb, Fly-TED

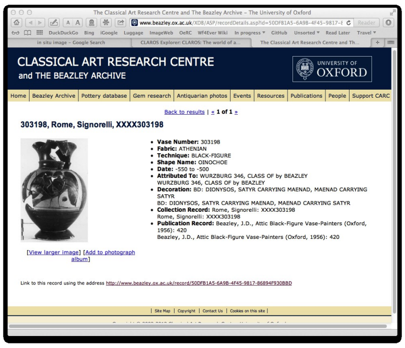
Personal web-research notebook

- investigations of CLAROS and related resources

Research Object creation

- aggregated context of an experiment

LINKED
OPEN DATA
RDF


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...
CLASSICAL ART RESEARCH CENTRE
and THE BEAZLEY ARCHIVE
UNIVERSITY OF OXFORD
Home
Beazley Archive
Pottery database
Gem research
Antiquarian photos
Events
Resources
Publications
People
Support CARC
Back to results | « 1 of 1 »
[View larger image] [Add to photograph album]
Site Map
Copyright
Contact Us
Cookies on this site

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

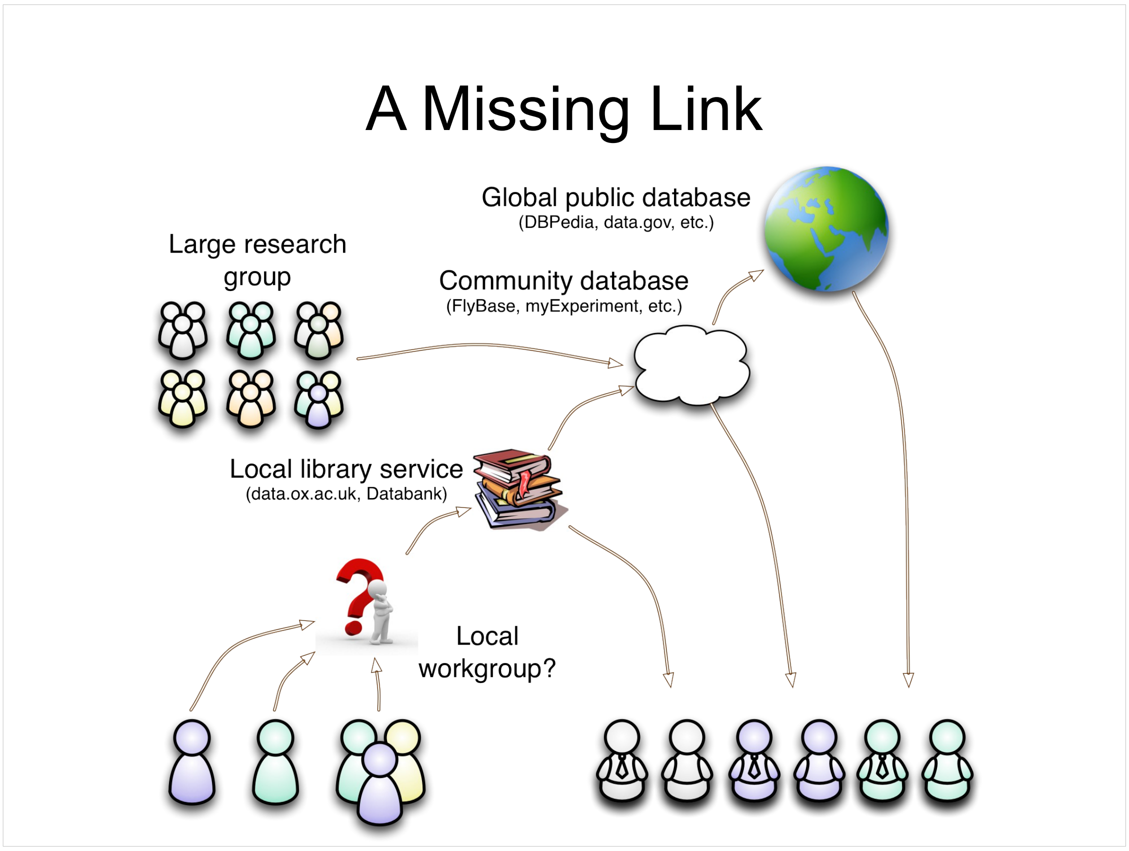
Link to this record using the address http://www.beazley.ox.ac.uk/record/50DFB1A5-6A9B-4F45-9817-86894F930BBD

Some examples of small-group activities that are not well served by available toolchains include:

Semantic annotation of in situ gene images, where lab observation of microscope images can yield important information about the relationship between gene expression activity and biological processes.

Collation and organization of information that is discovered by searching and analyzing the content of existing public databases. Ideally, this analysis might be able to feed new back to these data sources.

Creating Research Objects to describe the context of a lab experiment for other researchers to review and/or use.



A Missing Link

Global public database
(DBPedia, data.gov, etc.)

Large research group

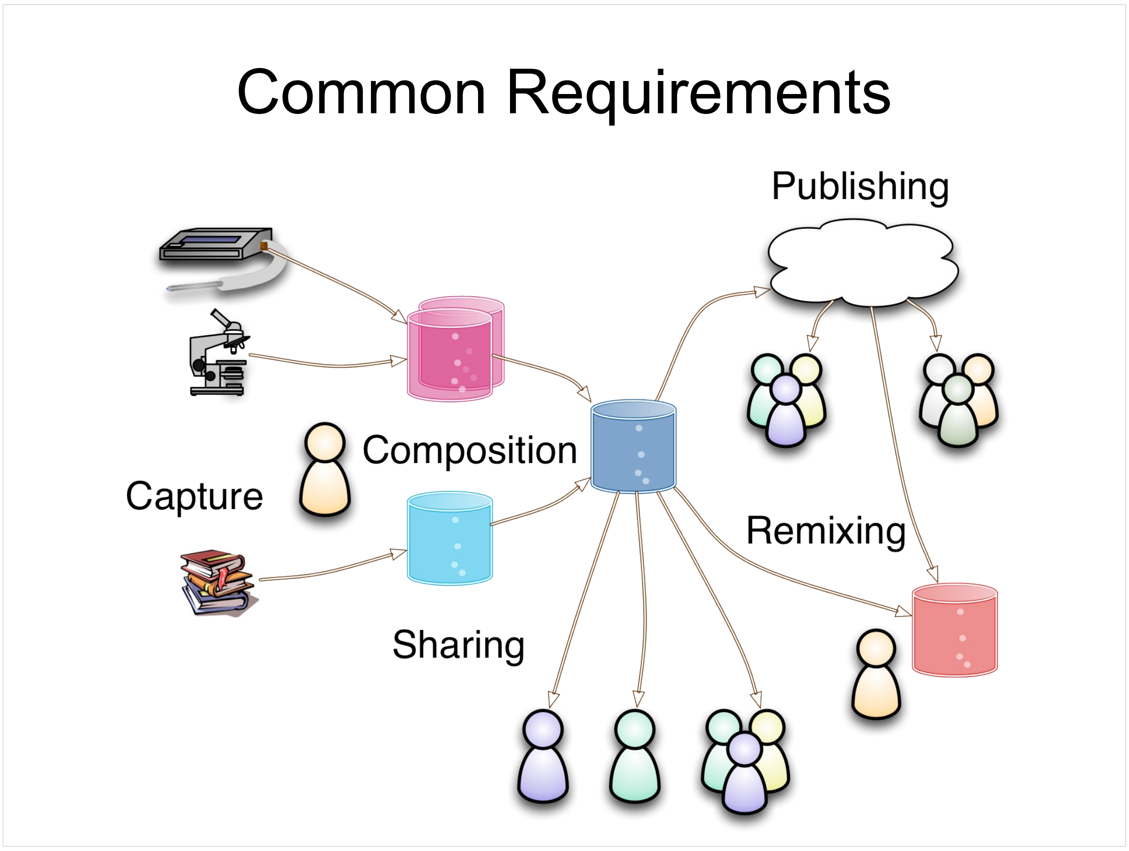
Community database
(FlyBase, myExperiment, etc.)

Local library service
(data.ox.ac.uk, Databank)

Local workgroup?

For these small research groups, it seems there is a missing link in the toolchain that conveys data from the lab bench to community and public repositories.

Lacking development capability and/or capacity, these groups are not well placed to supply the missing capabilities.



Common Requirements

These are some of the things that researchers would need to achieve to make their data meaningful and useful for sharing and publication:

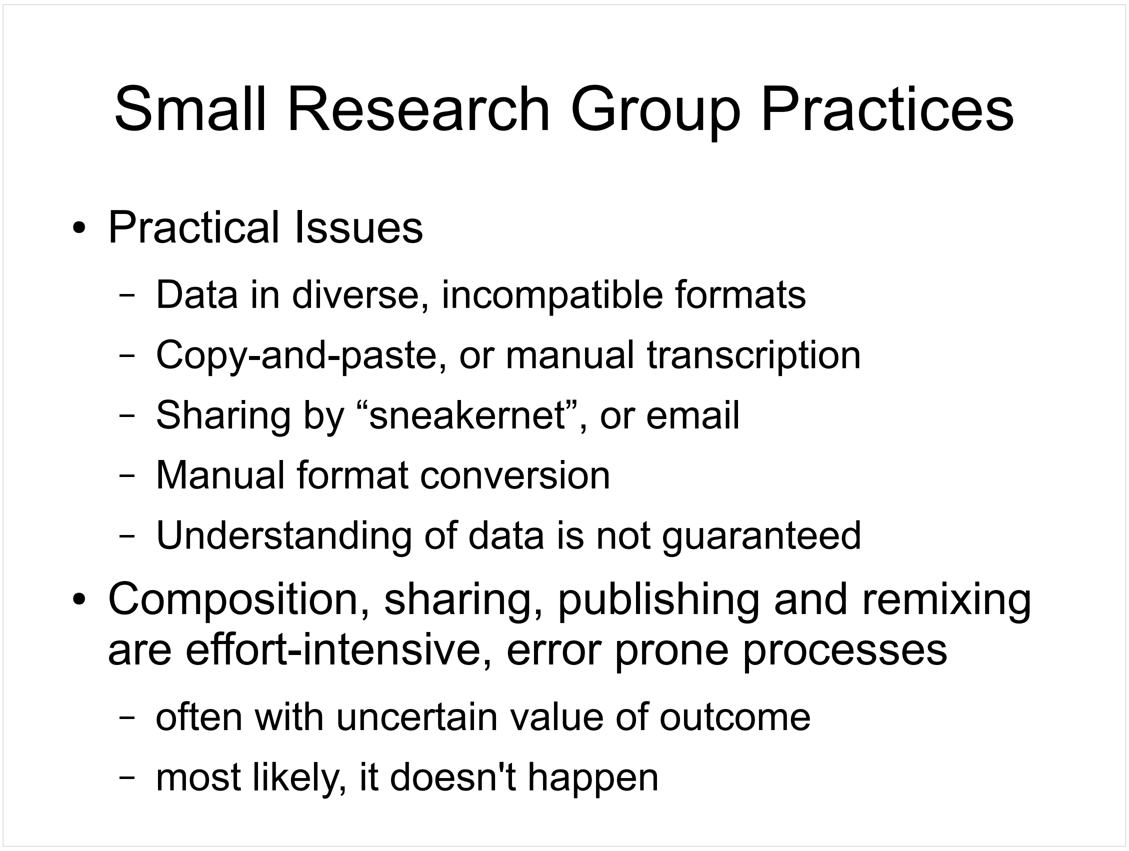
Initial capture from manual observation or lab instruments

Composition, or combining data from diverse local sources, capturing the relationships between them (e.g. microarray, in situ observations and literature based information about some set of genes).

Sharing with local colleagues for further analysis.

Publication for other researchers to use

Remixing with other published data to drive new research directions (e.g. FlyTED and FlyAtlas).



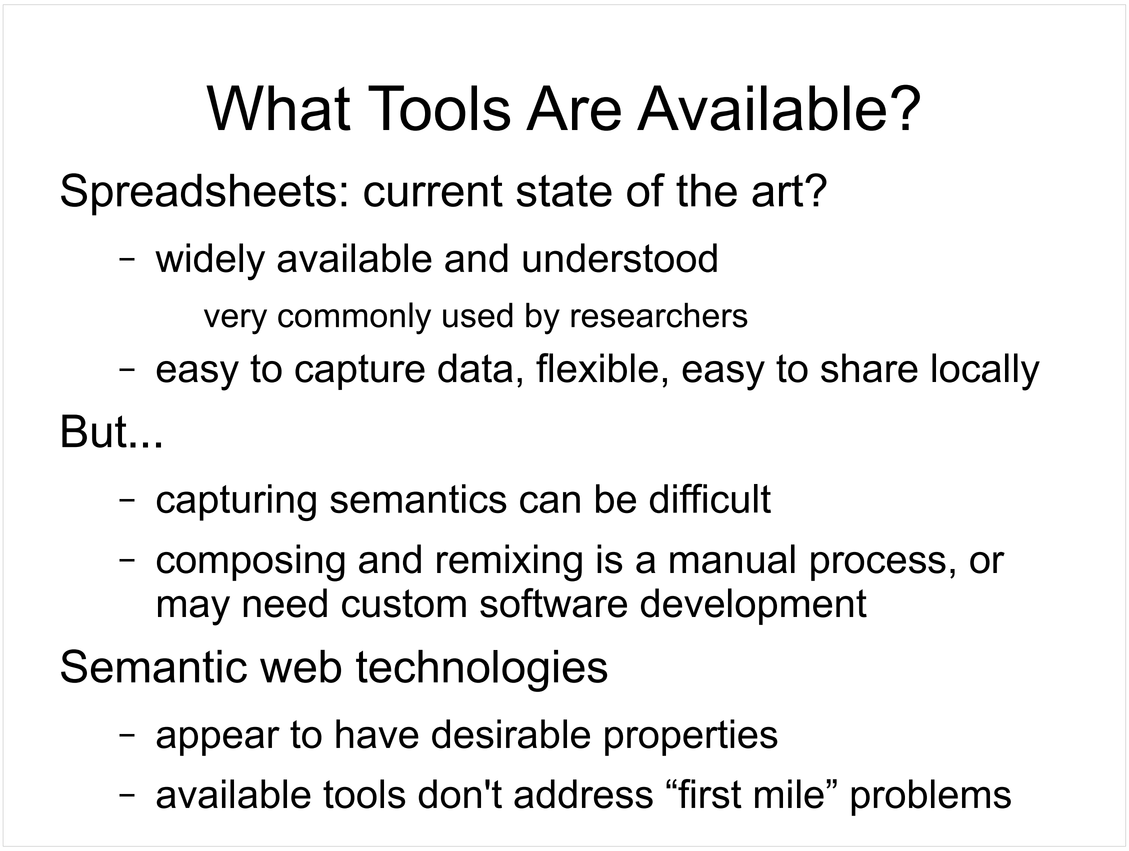
Small Research Group Practices

- 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 are the impediments for small groups in achieving these goals?

For this we need to examine the practical issues they face... incompatible formats, no automated transcription, transmission or format conversion, and formats that are not sufficiently self-describing to guarantee correct interpretation.

The steps required for composition, sharing, effective publishing and remixing are effort-intensive and error-prone, so unless there's an immediate need for the results of these, they may just not 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

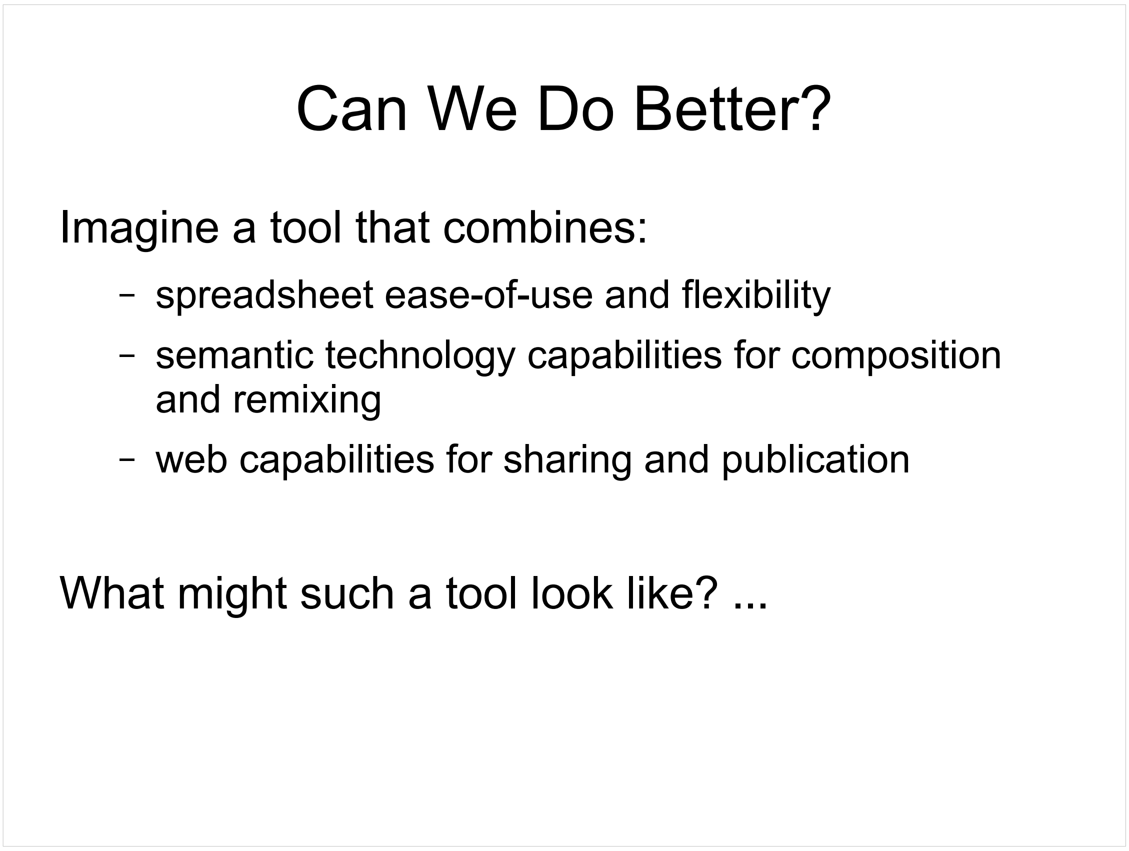
Semantic web technologies

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

Of all the off-the-shelf data management tools available to and used by researchers, the current state-of-the-art is probably spreadsheets. They are widely used and understood, flexible and forgiving in use for data acquisition, and work well for local data sharing.

But, with these tools, capturing the semantics in a standard fashion can be difficult or unreliable without a fair amount of planning, and combining different data sets may be either a manual process or require custom software development.

Semantic Web technologies have many desirable properties, but available tools don't handle the data acquisition problem very well.



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? ...

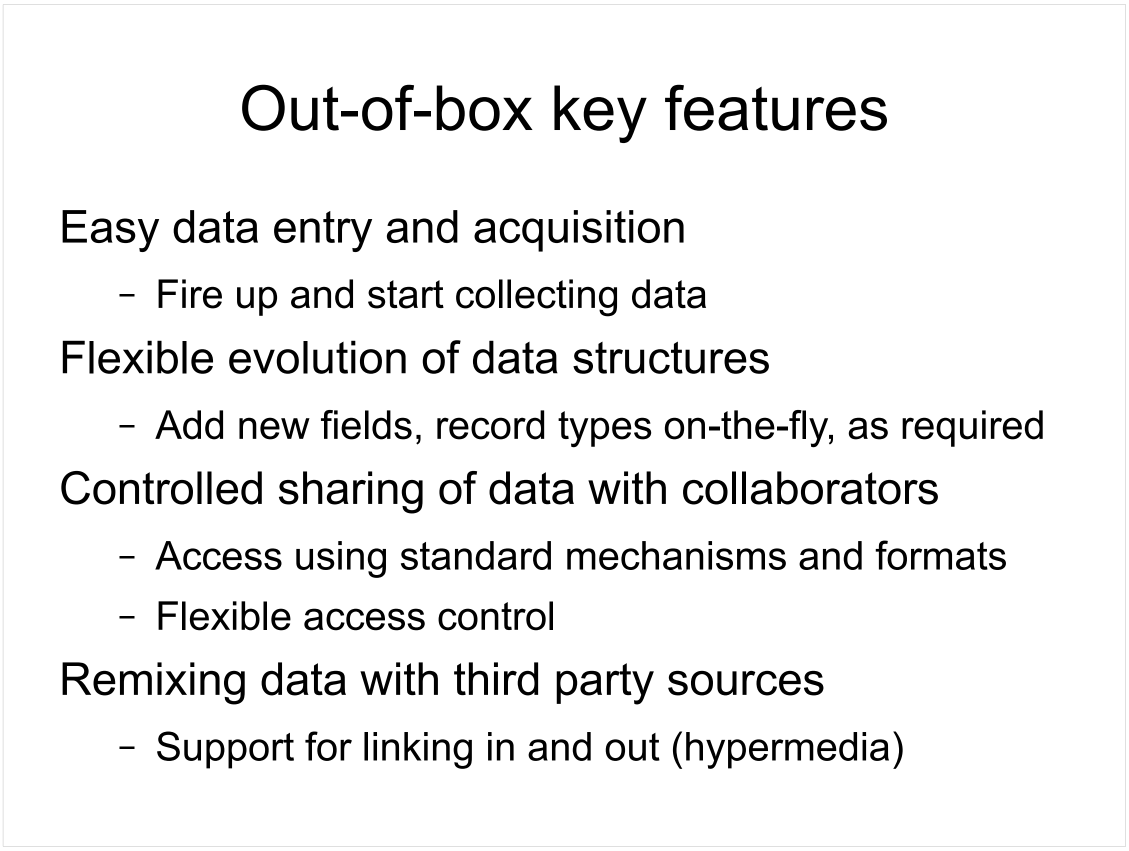
Can we make a tool that combines the best of these?

1. spreadsheet ease-of-use and flexibility

2. semantic technology capabilities for composition and remixing

3. web capabilities for sharing and publication

What might it 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)

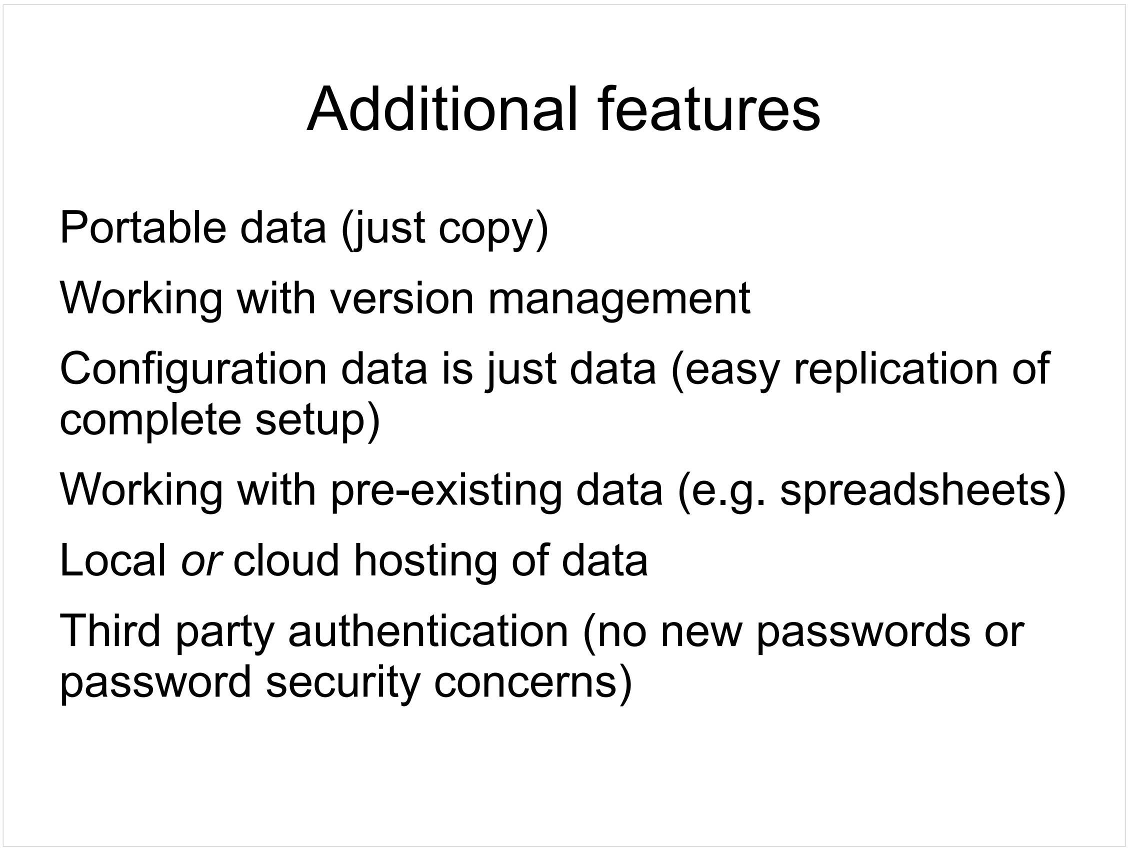
I think such a tool would have the following key features, out of the box, without prior configuration:

Easy data entry and acquisition

Flexible development of structure around the data

Controlled sharing with collaborators using common web tools

Remixing with third party resources – links in and out, and some form of co-reference management



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)

There are some further features, not immediately central to the key goals, that it seems would facilitate data sharing and good data management:

1. Copyable – no need for special software to use the data

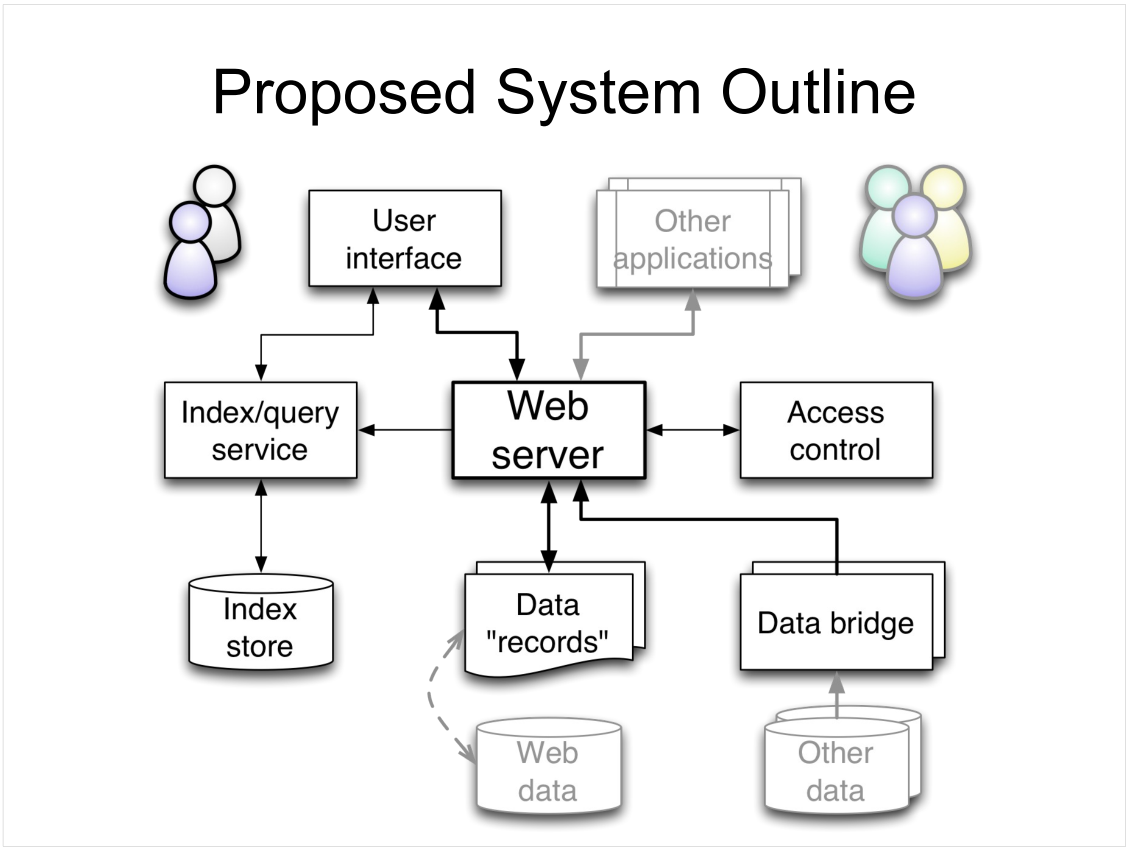
2. Compatible with use of version control

3. Descriptive metadata and configuration is just part of the data

4. Ability to work with pre-existing local data (e.g, spreadsheets)

5. Local or cloud hostable

6. Usable with third party authentication and access control systems (SSO, etc)



Proposed System Outline

User interface

Other applications

Index/query service

Web server

Access control

Index store

Data "records"

Data bridge

Web data

Other data

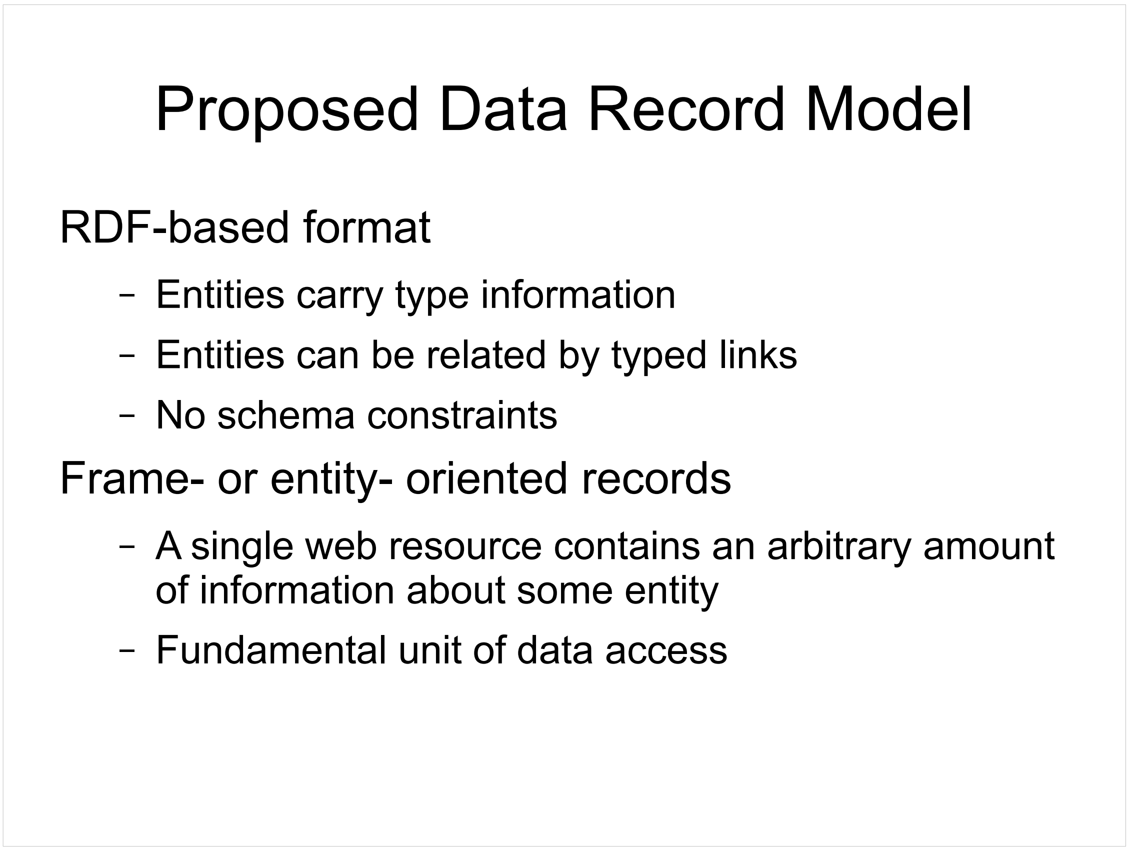
This is how I propose we might build such a system.

The key thing to note is that at its heart is a standard web server. All applications, including the primary data acquisition and viewing interface just use standard HTTP requests to access and update the data.

The next thing to note is the use of data bridges modules behind the server to access existing data, or data in different formats, locally or from the web. This would be the way of accessing and working with existing spreadsheet data.

Third, and one of the technical challenges, is that access control has to be aplied through the web server, not the applications.

Finally, the primary storage of data is as flat source files. Applications can work with indexing and query services as needed to optimize their own operations, but the data can still stand alone.



Proposed Data Record Model

RDF-based format

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

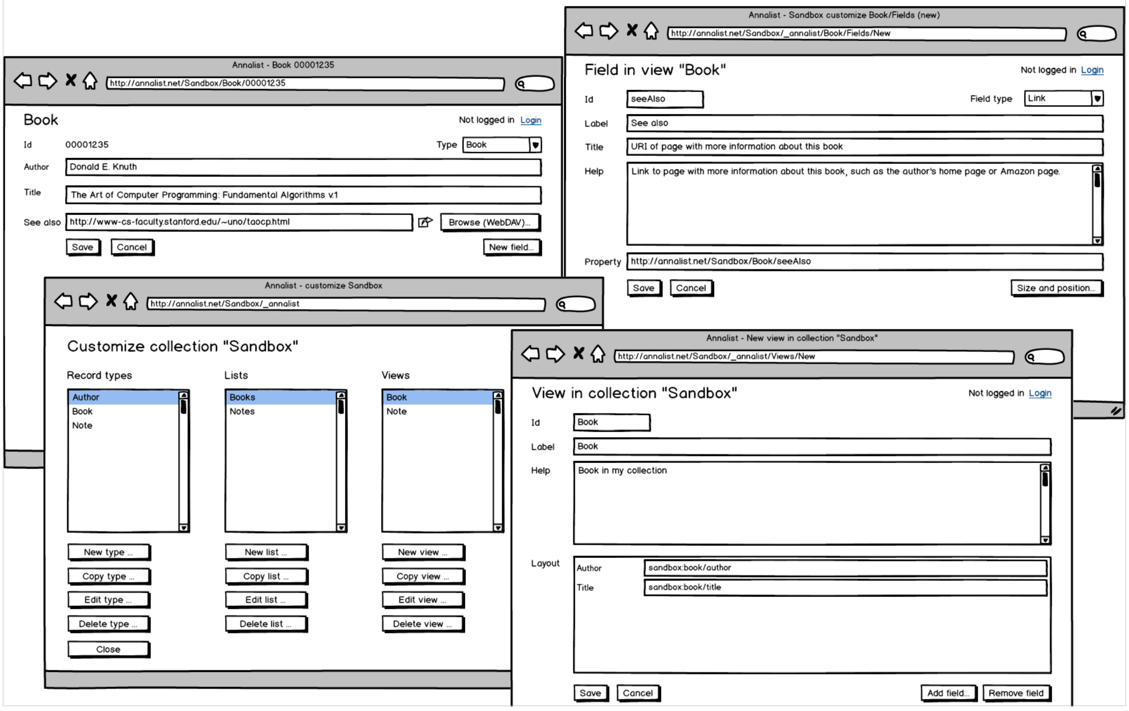
Frame- or entity- oriented records

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

I propose RDF as the primary data model, but stored as entity- or frame- oriented records. Thus a frame, or a collection of information about some resource, would be the primary unit of data access. E.g. information about a gene, or a species – the kind of level of information that might be supplied in a spreadsheet or database row (but without the structural constraint).

I believe this would form a basis for an easily understood, coherent exchange of information.

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

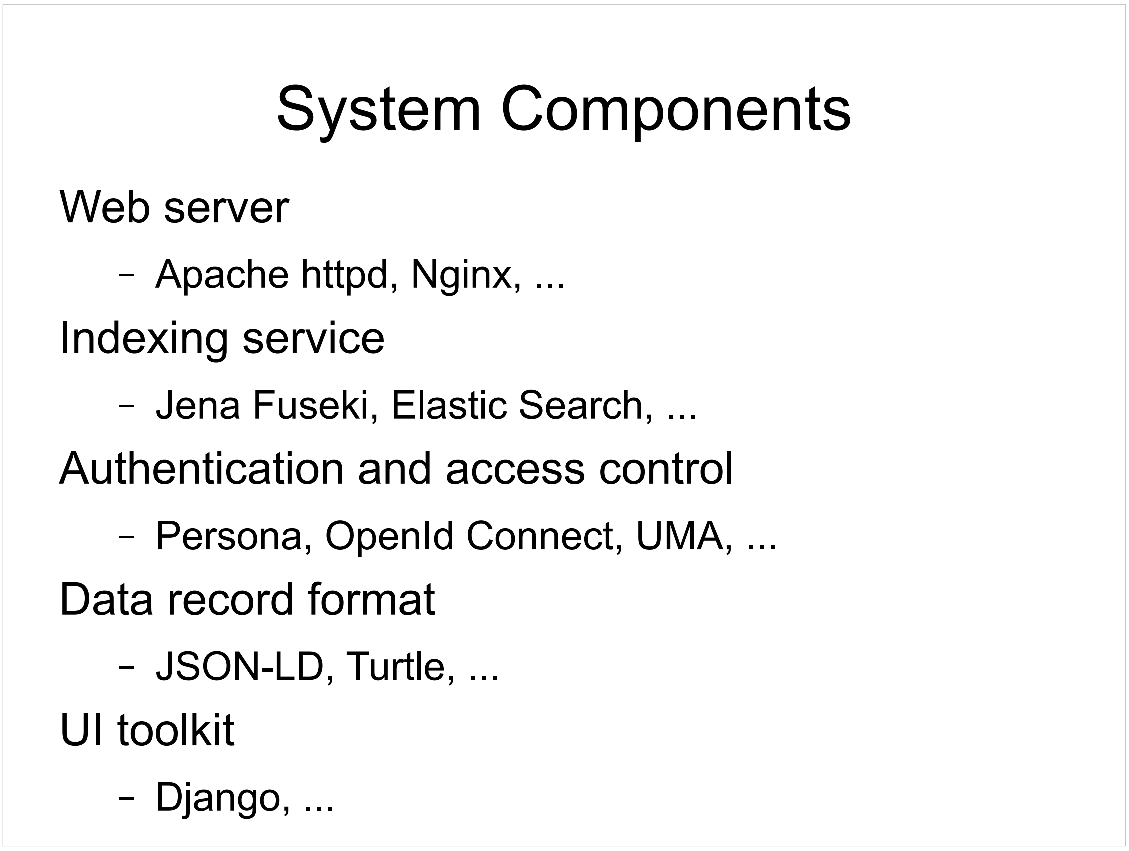
These are some wireframe mock-ups for the primary data entry, editing and viewing interface.

Key ideas here include:

- essentially the same interface for data entry and data structure definitions – no major mode changes

- aesthetics are more spreadsheet than product brochure.

Although web technology is envisaged, the emphasis here is for local, spreadhseet-like use by a researcher, not to provide a public web site.



System Components

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

Implementation is still at the planning stage, but these are some components under consideration.

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.)