the archival project
putting archival data on the semantic web, guerilla style...
the archival project

the data is out there...

big data ≠ web?

with linked open data, the web is data
the archival project
the evolution of the web

mashups linked data
WWW
the archival project
linked open data (what is it and why?)

use URIs as names for things.

use HTTP URIs so that people can look up those names.

when someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).

include links to other URIs so that they can discover more things
the archival project

how does linked data work?

Aaron Rubinstein

URI

HTML

representations

web page

RDF

Amherst, MA

FOAF profile

URI

info about Amherst
the archival project
big data, linked!
the archival project

giving linked data a shot

the questions...

how hard it is to publish linked data?

is it possible without unlimited resources and technical support?

is there a clear benefit?
the archival project
modeling names and collections

- URI
  - arch:correspondedWith
    - URI
      - arch:primaryProvenanceOf
        - URI
          - arch:correspondedWith
            - URI
              - arch:primaryProvenanceOf
                - URI
                  - arch:correspondedWith
                    - URI
                      - arch:primaryProvenanceOf
                        - URI
                          - arch:correspondedWith
                            - URI
                              - arch:primaryProvenanceOf
                                - URI
                                  - VIAF ID

Robert Francis
Marianne Moore
Robert Francis Papers
Predilections (book)
Marianne Moore Papers
the archival project
arch vocabulary

creating ontologies can be as simple or complex as you want it to be...

design principles

* re-use existing vocabularies

* only create terms that exist nowhere else

* keep it simple

http://purl.org/archival/vocab/arch
the archival project

tools

data extraction

python rdflib: [http://www.rdflib.net/](http://www.rdflib.net/)


googole refine: [http://code.google.com/p/google-refine/](http://code.google.com/p/google-refine/)

data publication

a webserver with URL rewriting

a web framework: [http://webpy.org](http://webpy.org)

an HTTP library: [http://pythonpaste.org/webob/](http://pythonpaste.org/webob/)
the archival project

the result...

http://gslis.simmons.edu/archival
the archival project

conclusions...

some hacking involved, however:

embedding RDFa is an easy way to integrate linked data into existing content.

URLs must be stable (which means institutional support).

beneficial side effects:

http://inkdroid.org/journal/2011/03/31/snac-hacks/
the archival project

next steps...

integrating structured data into applications built on Hydra.

automating linked data creation with a new CMS.

more large scale datasets

more applications for users

institutional commitment