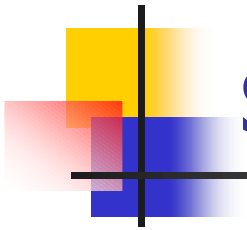
Setting up large-scale archive networks

Thomas Krichel
Long Island University
krichel@openlib.org

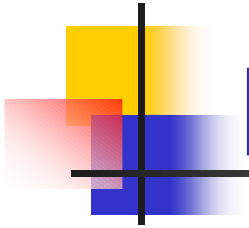
Open Archives Workshop
Oct. 17-19, 2002
CERN, Geneva



structure

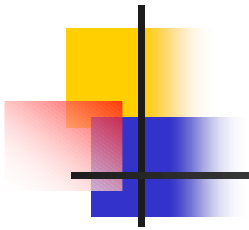
- general remarks
  - essentially backward looking
- specific business framework theory
  - essentially forward looking
- rallying cry (if I get there)
- 22 slides, no questions before the end, please

general remarks first



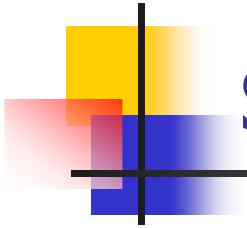
personal

- formerly lecturer in economics at the University of Surrey, Guildford
- now assistant professor at the Palmer School of Library and Information Science, at Long Island University, New York
- my attendance today is sponsored by the Open Society Institute.



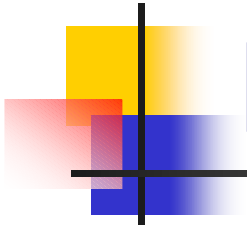
first experience

- started self-archiving in Economics in 1993
- work on archiving papers but
- main work on collecting data about papers archived elsewhere
- respect pre-Internet publication tradition of issuing working papers



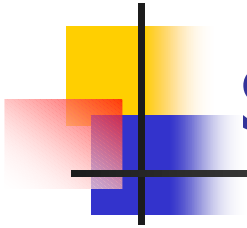
second experience

- founded RePEc, an digital library for the economics discipline, in 1997
- implements a data provider vs service provider architecture with crude, but robust technical means
  - static file harvesting
  - *attribute*: *value* templates
- over 200k records contributed by 260 archives, 10 user services
- a bit of a dinosaur



but still an economist

- "reputation for getting a lot done with limited resources" (Bill Arms)
- key: get others to do it!
- thus it all boils down to getting incentives right, an economic problem, fundamentally. Solutions
  - distributed volunteer power
  - alliences



searching for allies

- I started with no data
- I started with no computer on which to put data
- alliances are crucial
- seek out important partners first, the smaller ones will follow.
- example: US Fed gave me bibliographic data, other central bank follow

the world is yours to snarf

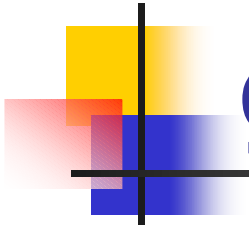
- it is critical to get a collection up to critical scale
- offer others to provide OAI services for them
  - already existing material in the institution
  - stuff that is held by others, but perceived as valuable

adequate as a strategy for the short run

in the long run we are all dead

- but we need to have a picture of what we want to achieve before we decide how to achieve it
- social transformation is more important than all technological features
- we need to grasp scholarly communication as an organic process within a community

theory of evaluative aggregation

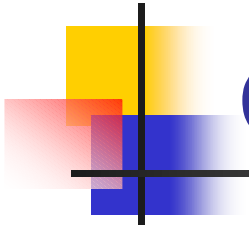


getting scholars involved

- it can argued that scholarly publishing is more a service to the authors than to the readers.
- getting scholars and their institutions involved seems crucial.
- this is best done on a per-discipline basis.

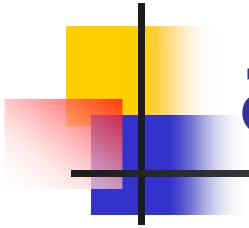
take a second look

- in addition to running open archives, we need open secondary datasets.
- "discipline-based" approach as a pile of secondary data is in no contradiction with institution-based archives, as long as there is a basic grouping in archives.
- we need aggregates of data…



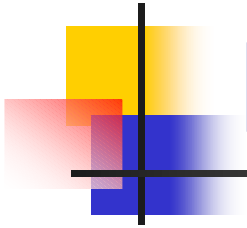
data? what data?

- basic aggregation
  - Persons
  - Institutions
  - Documents
  - Collections of documents
- documents provided by institutional archives.
- but we need relational dataset



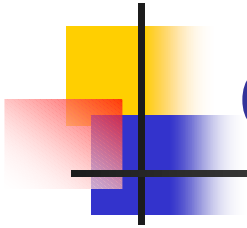
authority, he said

- authorities are groups that will get together to do the registration work
- model authority is RePEc
- emerging authorities
  - rclis
  - Physnet
- slow process…



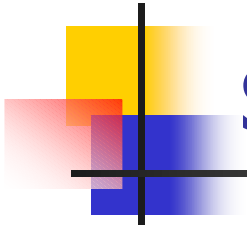
beyond basics

- to get scholars on board, we need to present them with measures on how well they are doing.
- simple repeat email saying "your work is more accessed in the open" is not enough.
- evaluative ranking is what works best.



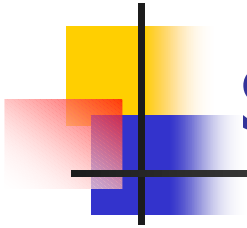
evaluative data

- evaluative data need to be gathered
  - abstract views
  - downloads
  - citation counts
  - collection membership evaluation
- LogEc does some of this in a crude way for RePEc
- even though crude, scholars pay attention



summary: status quo

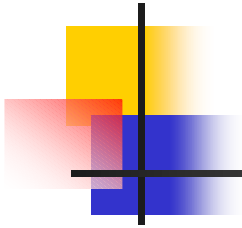
- two tear of archive and services
- services built on isolated metadata instance
- metadata is too broad to be "useful beyond Google"
- benefit of archives essentially digital preservation



status in spe

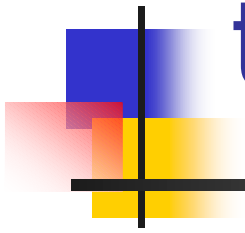
- three tier structure of archives, authorities, and services
- more specific metadata
- more *relational* metadata
- evaluative data will get scholars to come along

rallying cry



Krichel's call

- if a bunch of volunteer geeks can offer a whole computer O/S on the Internet then
- there is no reason to say that a bunch of librarian can not offer complete catalogs of scholarly data on the Internet
- tools building is in progress, ask me for details
- create authorities, go forth and collect!



thank you for your attention

http://openlib.org/home/krichel

http://repec.org

http://rclis.org