towards interoperable archives: the Universal Preprint Service initiative

herbert van de sompel
University of Gent - Central Library

Interconnectivity 2000
November 1-2, Stellenbosch, South Africa

acknowledgements
the Council on Library and Information Resources
the Digital Library Federation
SPARC & the American Research Libraries
Los Alamos National Laboratory - Research Library
5 input considerations lead to a conclusion (a personal interpretation of the discussions on the transformation of scholarly communication)

the Universal Preprint Service Initiative (an initiative to promote preprint solutions)
i1. the information chain

- consideration: disintermediation
- value chain: only links that add value survive
- endangered: equity of access
- opportunities: create optimal communication mechanism
<table>
<thead>
<tr>
<th>theory</th>
<th>i2. the established journal system</th>
</tr>
</thead>
<tbody>
<tr>
<td>accessibility</td>
<td>serials crisis</td>
</tr>
<tr>
<td>registration</td>
<td>no registration only</td>
</tr>
<tr>
<td>certification</td>
<td>delay &amp; supressing of ideas</td>
</tr>
<tr>
<td>awareness</td>
<td>who does the digital archiving?</td>
</tr>
<tr>
<td>archiving</td>
<td>stabilizer</td>
</tr>
</tbody>
</table>
• deconstructed journal (Smith - 1993)
  • decoupling of registration and certification
  • institutions as collectors/distributors of their author’s uncertified writings
• subversive proposal (Harnad - 1994)
  • author self-archiving
• xxx e-print archive (Physics - 1991 - Los Alamos - Ginsparg)
• NCSTRL (Computer Science - Cornell U - Lagoze)
• CogPrints (Cognitive Sciences - Southampton U - Harnad)
• NDLTD (Theses - Virginia Tech - Fox)
• RePEc (Economy - Surrey U - Krichel)

high accessibility - registration - no certification
• The Innovator’s dilemma (Christensen - 1997)
  • sustaining versus disruptive technologies;
  • disruptive technologies somehow perform worse than established ones, but they are convenient, cheap, …
  • disruptive technologies can create competition in an existing value network by creating a new one first;
01. main conclusions

• decouple registration and certification:
  • create free information layer of non-certified information
  • promote commercial overlay services, amongst others addressing certification

• free layer via institutional involvement through libraries; libraries can add value to the digital information chain by
  • repositioning in the information chain,
  • becoming archivers of non-certified information authored by institutional authors
A R

theory o2. another information chain

company commercial information layer

library free information layer
03. other ways to address the functions

- Accessibility:
  - high availability: free layer
  - retrievability: free & commercial layer

- Registration:
  - author via an institutional process

- Certification:
  - commercial layer

- Awareness:
  - fast & no suppressing of ideas

- Archiving:
  - free layer: institutional
  - commercial layer: (?)

- Rewarding:
  - mechanisms based on free & commercial
The Universal Preprint Service initiative has been set up to create a forum to discuss and solve matters of interoperability between preprint solutions, as a way to promote their global acceptance.

Paul Ginsparg, Rick Luce & Herbert Van de Sompel

http://vole.lanl.gov/ups/
The transformation of scholarly communication is not only a political/sociological/economical matter. Technology is important:

• How should preprint archives be architected?
• How to achieve a level of compatibility between archives, both discipline-oriented and institution-based?
• How to integrate the free layer and the commercial layer?
• Technical means to introduce alternatives for the rewarding function?
• ...
practice the Santa Fe UPS/1 meeting 10/21-22/99

• Supported by:
  • ARL, CLIR, DLF, LANL Library, SPARC

• Goals:
  • enable the creation of a cross-archive end-user service
  • make recommendations on the architectural design of archive solutions in order to facilitate the creation of such services

• Discussion stimulated by the creation of an experimental cross-archive prototype
Caroline Arms  Library of Congress
Les Carr  University of Southampton & CogPrints
Eric Celeste  MIT
Mark Doyle  the American Physical Society
Dale Flecker  Harvard University
Edward Fox  Virginia Tech & NDLTD
Mike Friedman  HighWire Press & Stanford University
Paul Ginsparg  Los Alamos National Laboratory & xxx
Paul Gherman  Vanderbilt University & SPARC
Stevan Harnad  University of Southampton & CogPrints
Thomas Krichel  University of Surrey & RePEc
Carl Lagoze  Cornell University
Rick Luce  Los Alamos National Laboratory

practice the UPS/1 participants (1)
<table>
<thead>
<tr>
<th>Name</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clifford Lynch</td>
<td>Coalition for Networked Information</td>
</tr>
<tr>
<td>Kurt Maly</td>
<td>Old Dominion University</td>
</tr>
<tr>
<td>Deanna Marcum</td>
<td>CLIR</td>
</tr>
<tr>
<td>Michael Nelson</td>
<td>NASA Langley</td>
</tr>
<tr>
<td>Heath O’Conell</td>
<td>Stanford SLAC/SPIRES</td>
</tr>
<tr>
<td>John Ober</td>
<td>University of California</td>
</tr>
<tr>
<td>Bill Parks</td>
<td>Washington University &amp; EconWPA</td>
</tr>
<tr>
<td>Herbert Van de Sompel</td>
<td>University of Ghent</td>
</tr>
<tr>
<td>Eric Van de Velde</td>
<td>Caltech</td>
</tr>
<tr>
<td>Don Waters</td>
<td>The Andrew W. Mellon Foundation</td>
</tr>
<tr>
<td>Ken Weiss</td>
<td>University of California</td>
</tr>
</tbody>
</table>
• coordination: herbert van de sompel, michael nelson, thomas krichel

• involvement of:
  – Old Dominion U & NASA Langley
  – U of Surrey
  – U of Ghent
  – Los Alamos National Laboratory - Library
  – Russian Academy of Science - Siberian branch
• Los Alamos National Laboratory - Research Library
• JISC eLib WoPEc project
project datasets

- metadata only
- full text remains at archives
- static dumps obtained ca. July 99

<table>
<thead>
<tr>
<th>Source</th>
<th>Objects</th>
<th>Full-text</th>
</tr>
</thead>
<tbody>
<tr>
<td>the arXiv</td>
<td>85,223</td>
<td>85,223</td>
</tr>
<tr>
<td>CogPrints</td>
<td>742</td>
<td>659</td>
</tr>
<tr>
<td>NACA</td>
<td>3,036</td>
<td>3,036</td>
</tr>
<tr>
<td>NCSTRL</td>
<td>29,184</td>
<td>9,084</td>
</tr>
<tr>
<td>NDLTD</td>
<td>1,590</td>
<td>951</td>
</tr>
<tr>
<td>RePEc</td>
<td>73,367</td>
<td>13,582</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>193,142</strong></td>
<td><strong>112,535</strong></td>
</tr>
</tbody>
</table>
• conversion of metadata to ReDIF
• re-creation of archives using intelligent data-objects called buckets
• creation of NCSTRL+ end-user service
• addition of SFX linking services
fundamental distinction between:
• data providers take care of archive functions: submission, maintenance
• data implementers provide end-user functions: search interfaces, …
results
open archives

submission mechanism

archive

machine interface

native end-user interface
open, harvestable archives
results  harvest protocol & criteria

subset of Dienst protocol:
• ask an archive for its subdivisions
• harvest data from these subdivisions

harvest criteria
• bulk
• accession date
• subject
• author affiliation
• material type
results: metadata formats resulting from harvest

• Internal format of the archive (RFC1807, ReDIF, Refer, MARC, …)
• Santa Fe Set: Dublin Core compliant minimal set
  • identifier (archive, id)
  • author, author affiliation
  • title
  • accession date
  • abstract
  • subject
  • date for user
  • comment
  • refereed
results
Santa Fe Convention

• Architectural concept
• Technical recommendations
• Gentlemen’s agreement between providers and implementers:
  • providers describe conditions for usage of the data
  • implementers describe their usage of the data

• Will be implemented by all existing and future archives represented in Santa Fe
• Already very positive reactions from other archives
• Short story in Science Magazine, Friday October 29
thank you

Find up-to-date information on UPS at
http://vole.lanl.gov/ups/

Check out the experimental proto at
http://ups.cs.odu.edu
UPS protoproto

Initiated by Herbert Van de Sompel, Michael Nelson & Thomas Krichel

- Supported by LANL Library, eLib
- Data from xxx, NCSTRL NDLTD, NASA, CogPrints, RePEc

- Architectural choices:
  - distributed archives
  - N open archives
  - M digital library services
  - intelligent metadata
  - integration with established system via SFX

A discussion model for UPS 1
commericial overlay

integrate with commercial overlay

C O M M F R E E

peer review | journal paper | citation searches
S6 | S7 | S8

peer comment | full text search | citation searches
S1 | S2 | S3

meta
institution based ASA

• Institution:
  • institutional adaptation of rewarding function
  • natural extension of the prepublishing habit beyond early adopter disciplines
  • the addition of a level of author-authentication to submissions
  • concern about the institutional brand which might encourage institutions to impose a certain level of quality control to its output

• Library:
  • ethical incentive to ensure the right for information
  • existing international infrastructure and cooperation
  • should be best motivated to archive the institutional output
  • know-how that guarantees a formal quality of the output
  • extension of the ongoing creation of the institutional bibliography