Open Archives Initiative (OAI)

A low-barrier interoperable standard for the dissemination of content

- In priniciple, not tied to a specific purpose
- Note: open in terms of open architecture, not necessarily) free

- **P**rotocol for **M**etadata **H**arvesting
  - Defines standard for advertising metadata in a repository.
  - Standard packages for harvesting have been defined.
- DP 9
  - A standard for exposing metadata to web crawlers as web pages.

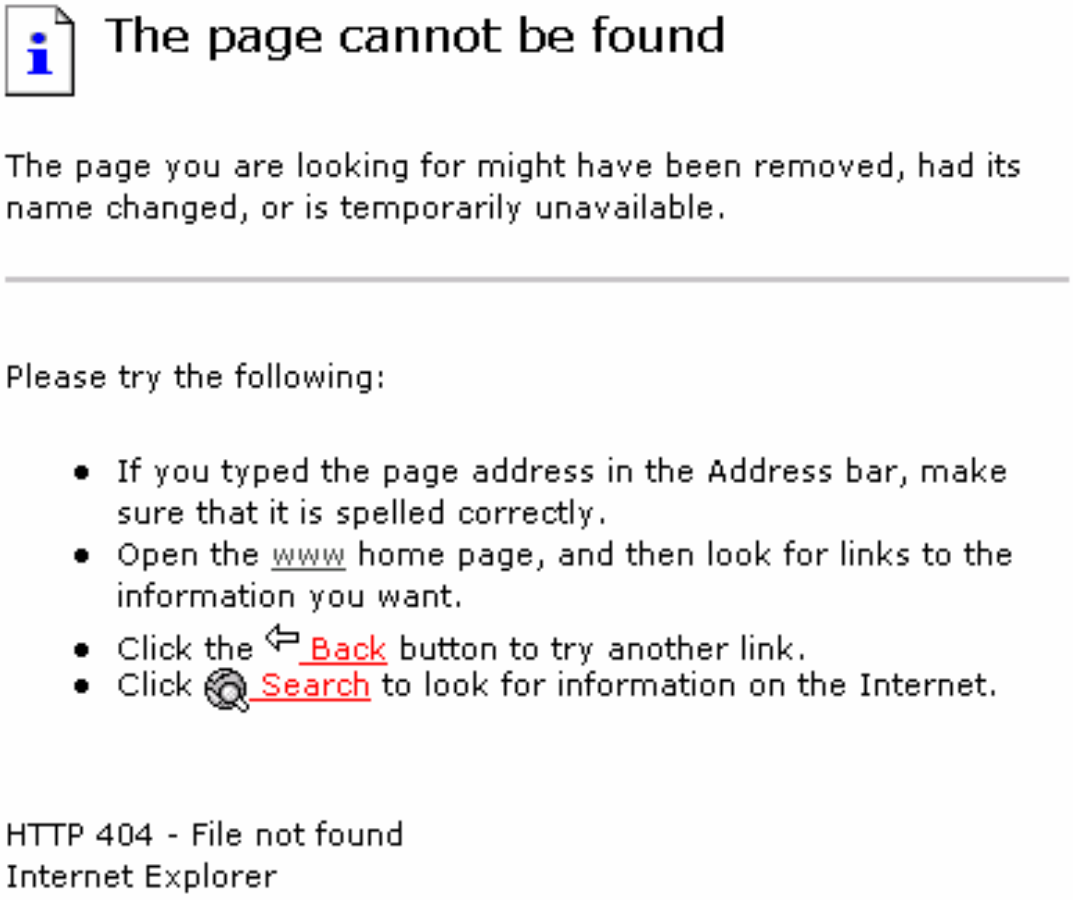
Identifiers

Week 4 Min-Yen KAN

*Partially based on William Arms presentation at Cornell University

Modified by permission

You see this everyday…



The page cannot be found

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the www home page, and then look for links to the information you want.
- Click the Back button to try another link.
- Click Search to look for information on the Internet.

HTTP 404 - File not found
Internet Explorer

What's the solution?

What's the 404 for anyways?
• 4xx – Error codes for web servers
• 404 – File not found or permission errors

Desirable Properties of Identifiers

- Location independent name
- Globally unique
- Persistent across time
- Choice of human generated or automatic generation
- Fast resolution
- Decentralized administration
- Supported from standard user interfaces

Identifier systems

We'll look at several different systems today

- URN
- PURL
- DOI
- OpenURL

Uniform Resource Names (URN)

- Globally unique, persistent, and accessible over the network
  - Persistence: That is, the URN will be globally unique forever.
  - Scalability: URNs can be assigned to any resource
  - Legacy / Extensible: Backward and forward compatible

Some Examples:

urn:hdl:cnri.dlib/august95

urn:lifn:some.domain:anything-goes-here

urn:path:/A/B/C/doc.html

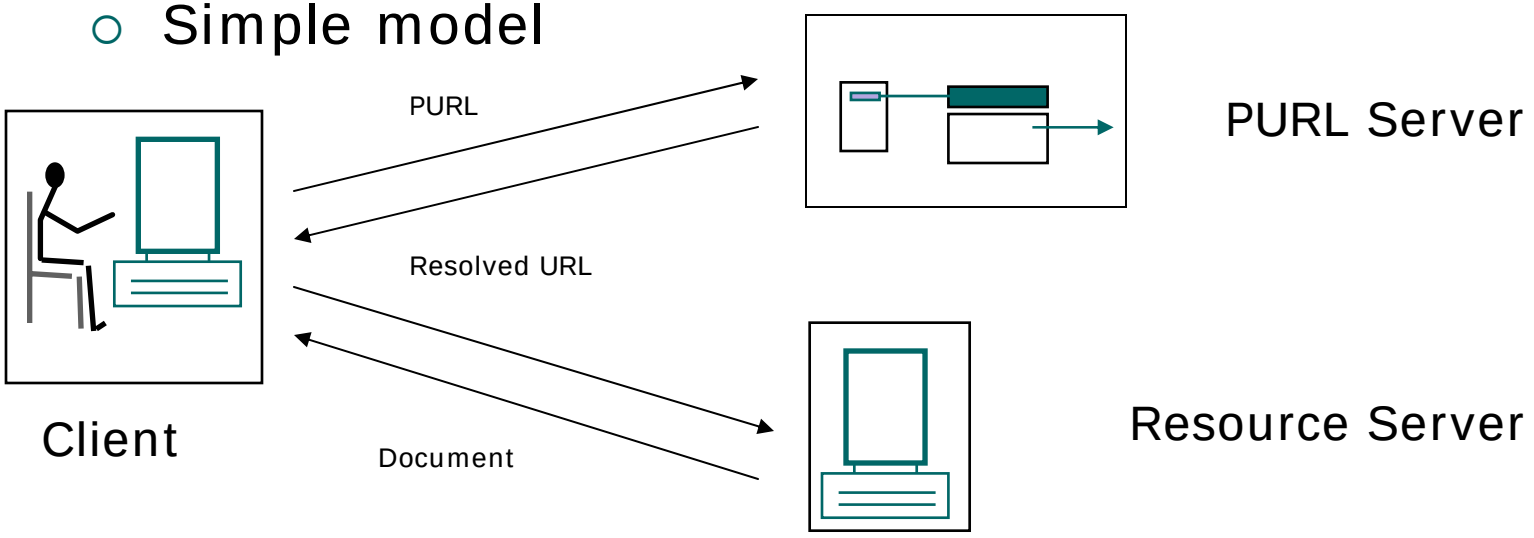
urn:inet:library.bigstate.edu:aj17-mcc

Scheme **Parameters**

Persistent URLs

http://purl.org/

- PURL is a normal URL
- Implement a layer of *in*direction
- Uses standard HTTP *re*direct
- Simple model



More details on PURL

- Partial redirection
  1. http://purl.org/kanmy/pictures/nus.jpg
  2. http://www.comp.nus.edu.sg/~kanmy/pictures/nus.jpg

- A PURL with no associated indirection causes the PURL resolver to generate a history page

- Private and universal indirection with access control

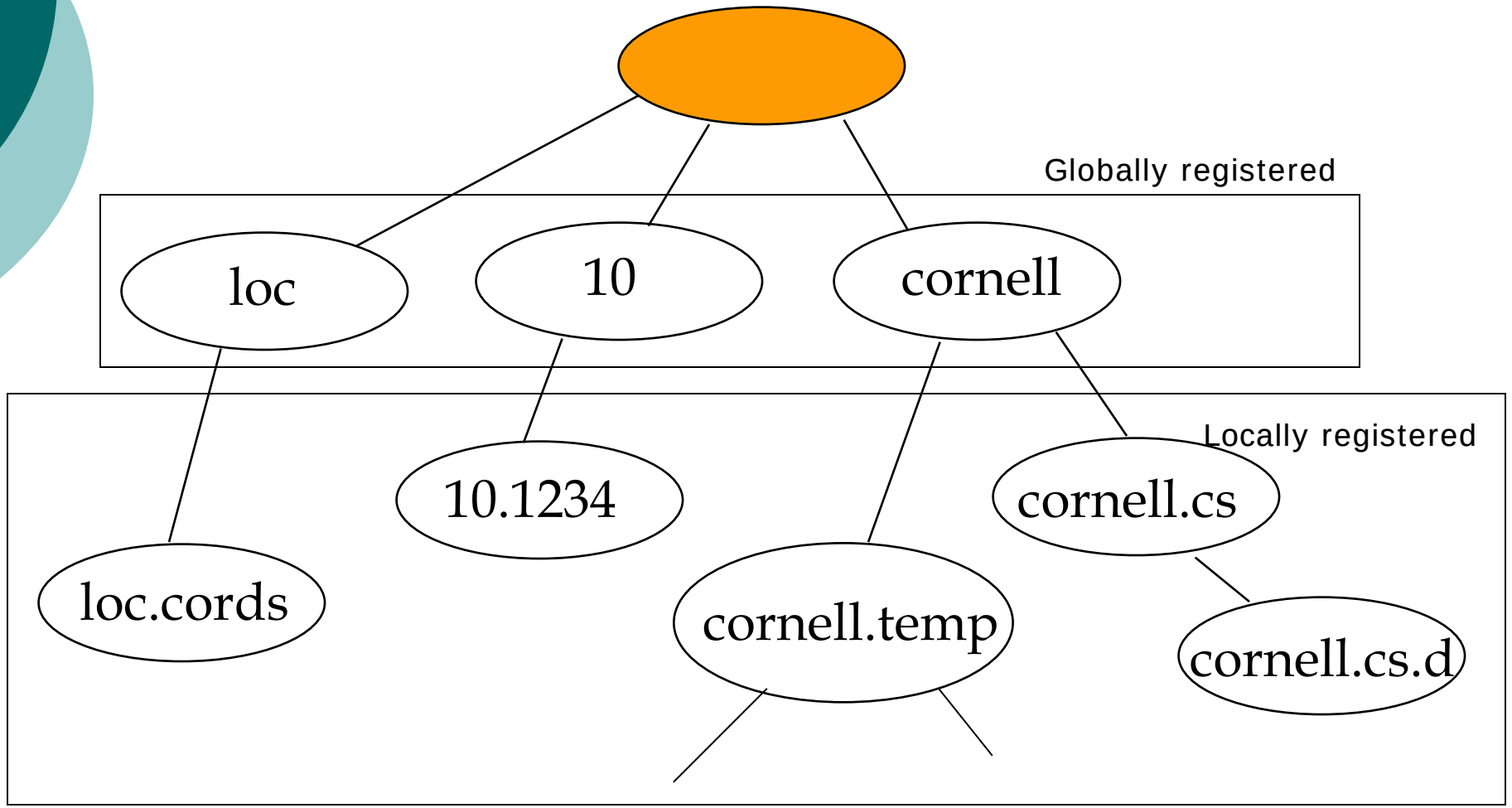
PURL Issues

- Places the burden of resolution on the manager of information
- PURL resolvers don't know about each other: federated, no centralized registry
- If URL goes down, doesn't force or notify maintainer
- Doesn't guarantee that document will be available, indirection can lead to a 404

Examples of DOIs

Publisher ID assigned by DOI Agency	**Item ID** assigned by Publisher
10.1048	/ 872
10.156	/ catalog-96
10.1532	/ *PII*
10.18698	/ *SICI*

Hierarchy of Naming Authorities



Address Rules

The Global Handle Service stores:

- a record for each naming authority
- a record for each local handle service

The record for each naming authority includes:

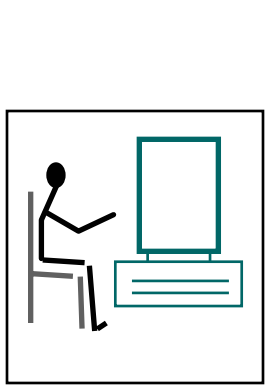
- the home handle service for that naming authority

For each handle, the home handle service stores:

- the handle record

Multiple Resolution

- Leave the resolution up to the client
- Return all DOI data to the client



Client

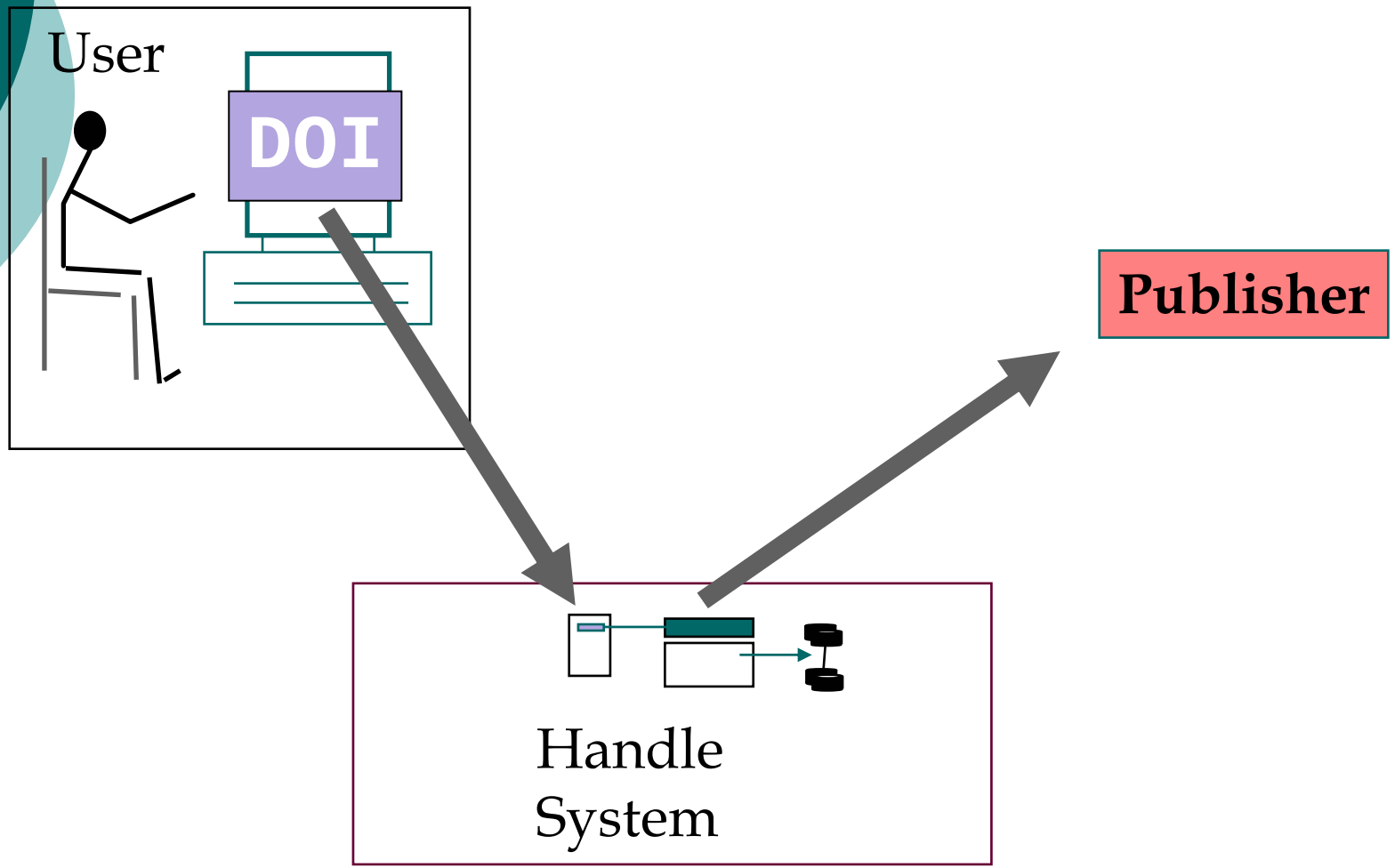
handle record

DOI>

resolution

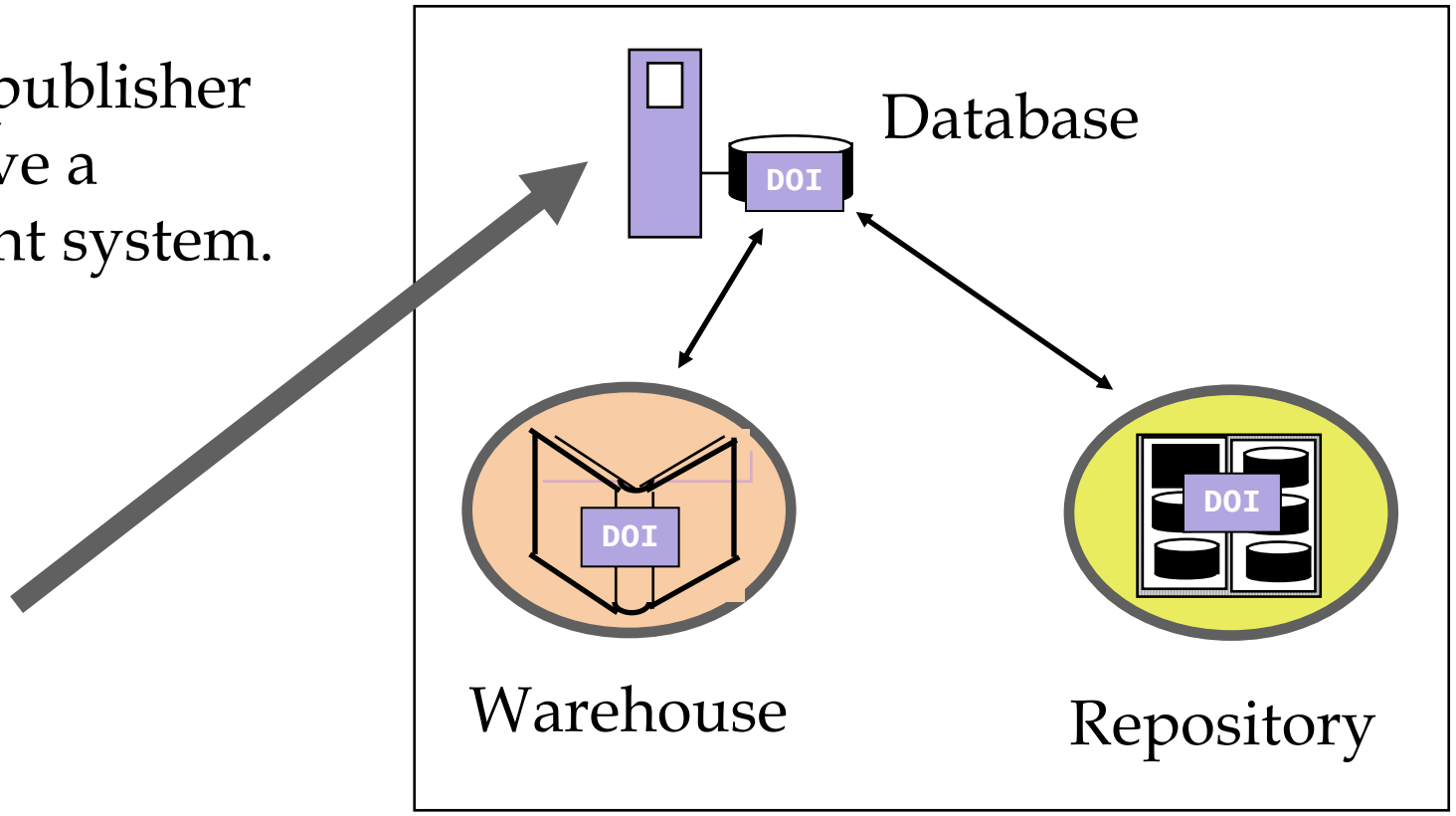
Adm	Admin Data
Adm	Admin Data
URL	http://www.cnri/xyz
RAP	merlin.dlib.org
CORBA	orb:#cornell[]norb

DOIs in action



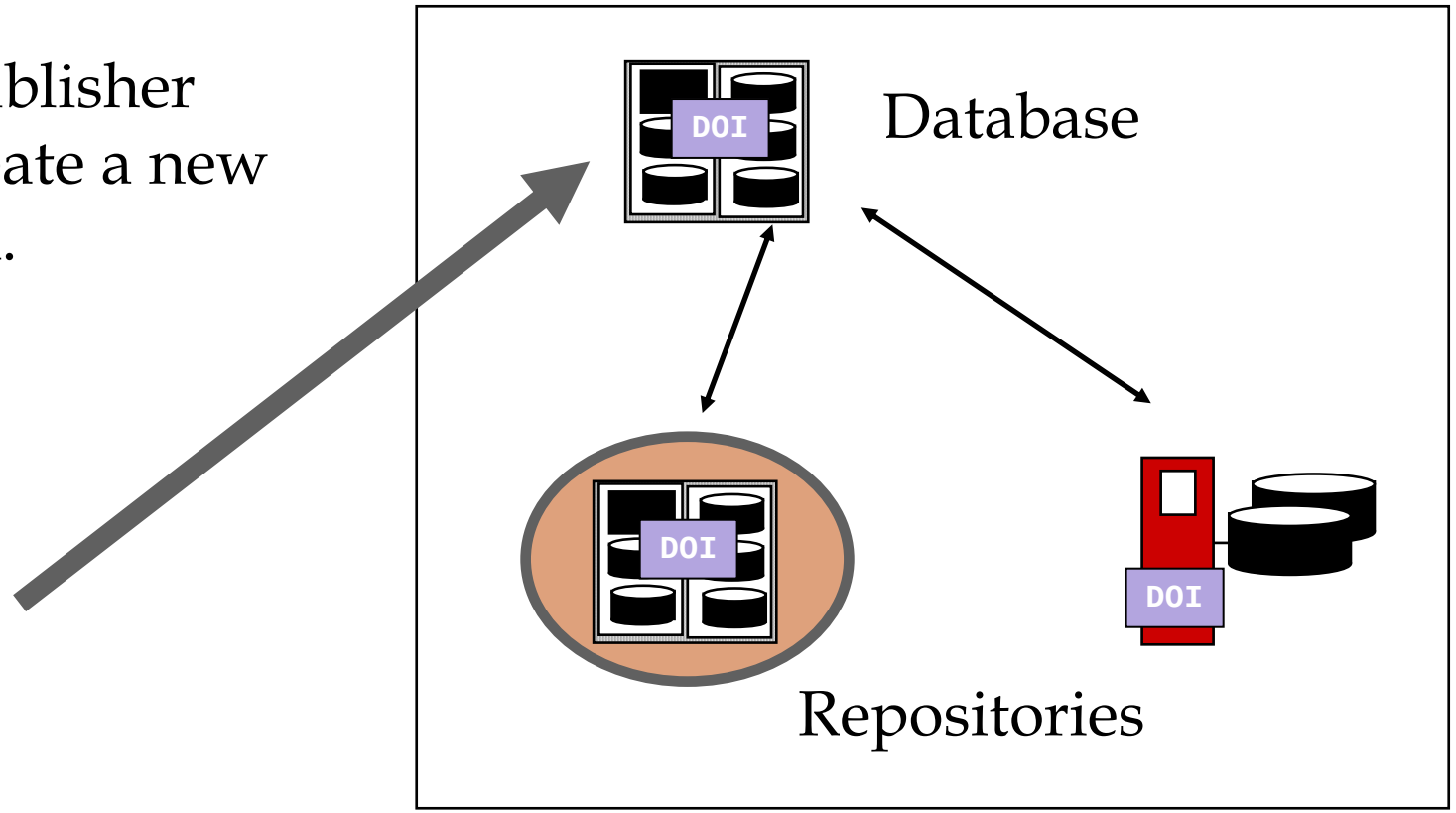
Flexibility

Every publisher can have a different system.

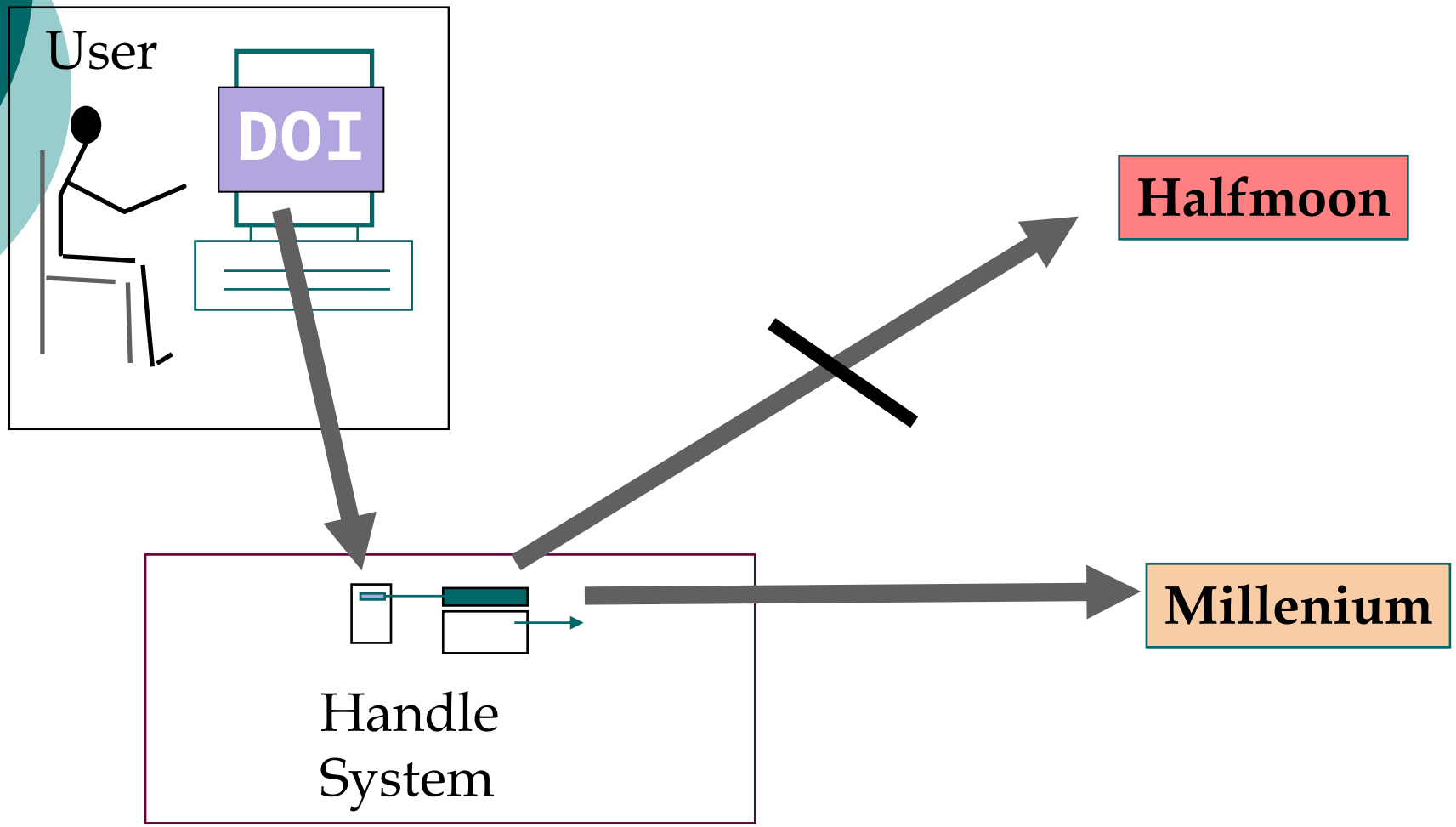


Reorganization by Publisher

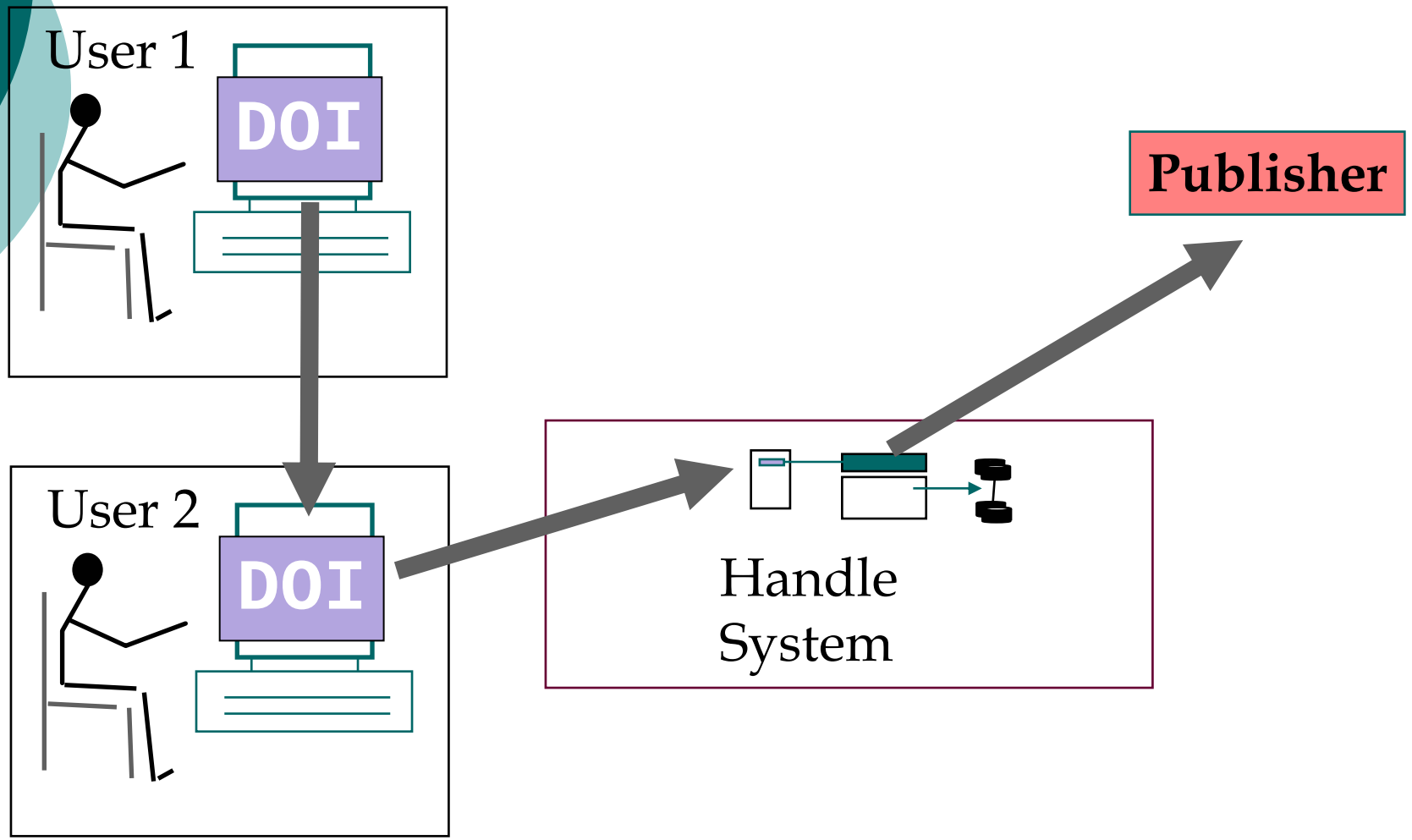
The publisher can create a new system.



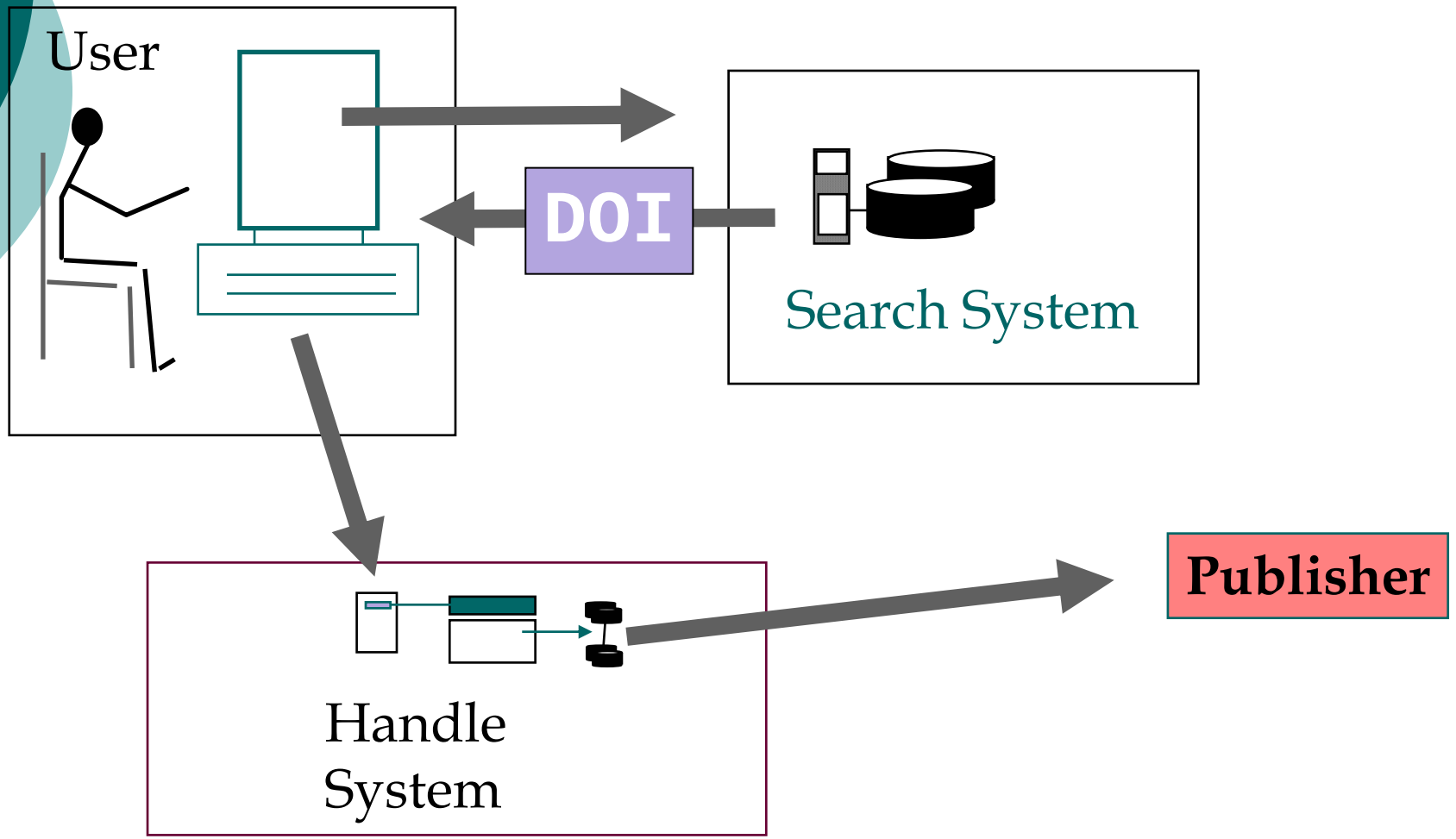
Change of Publisher



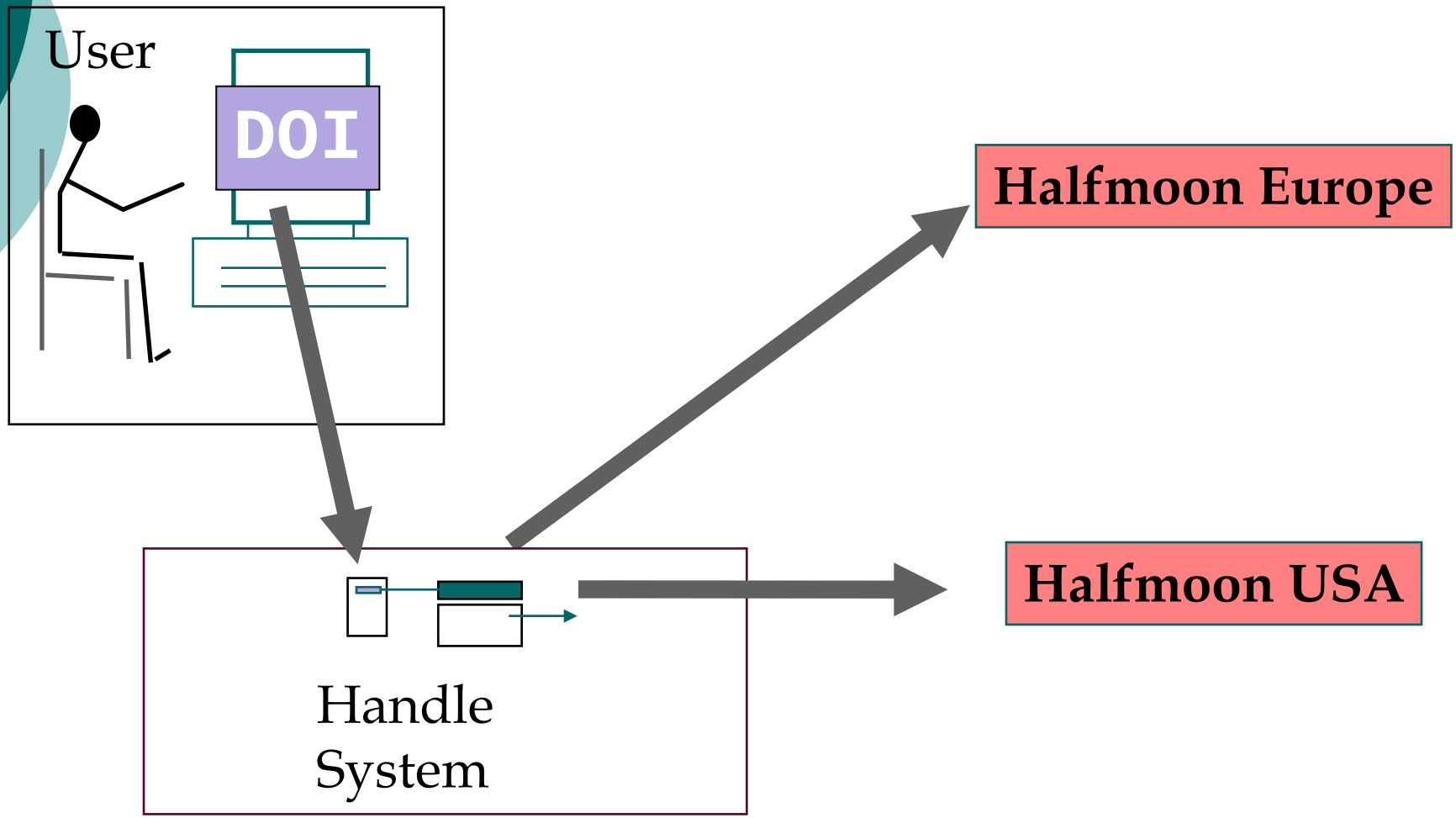
Citation



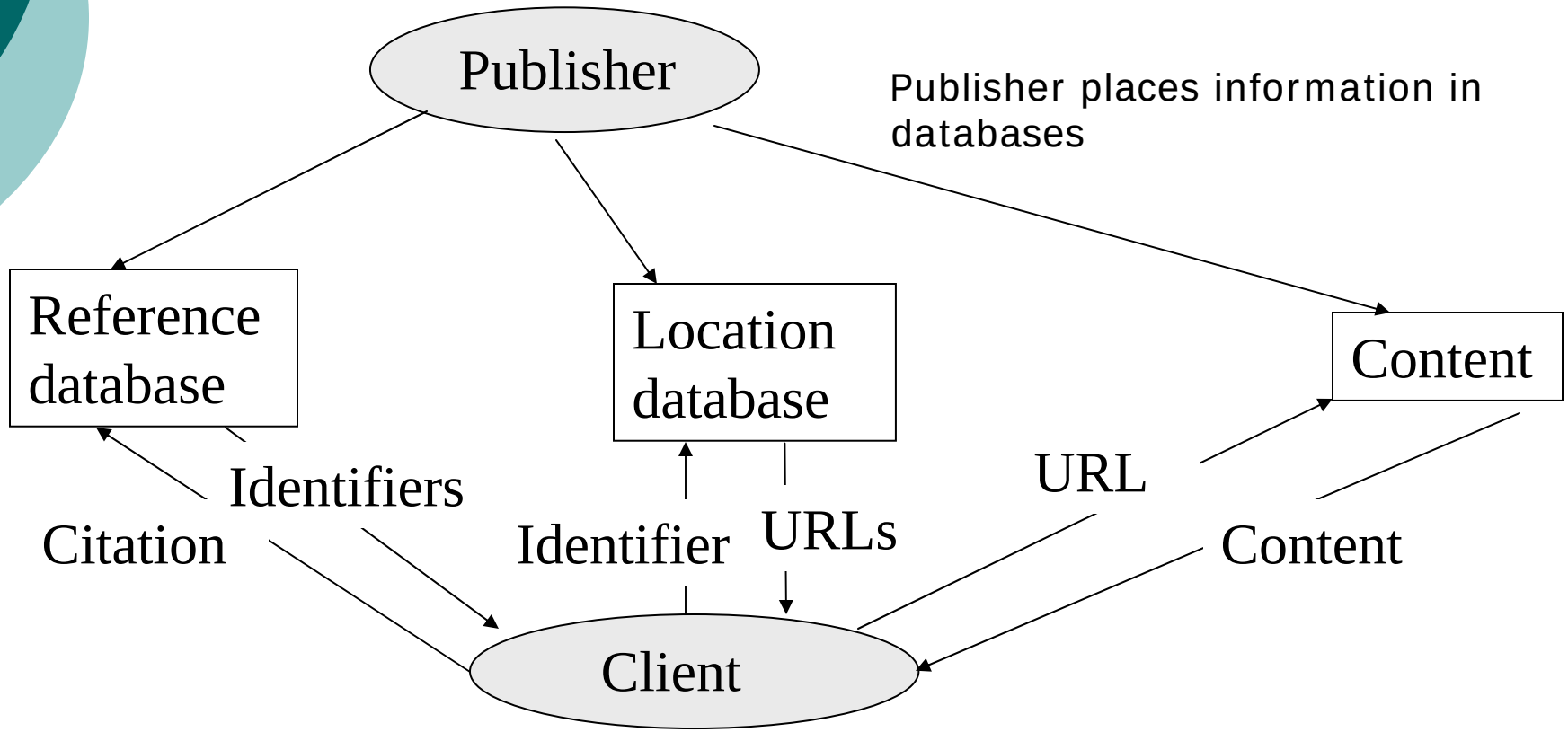
Catalogs and Indices



Multiple Copies



The General Model

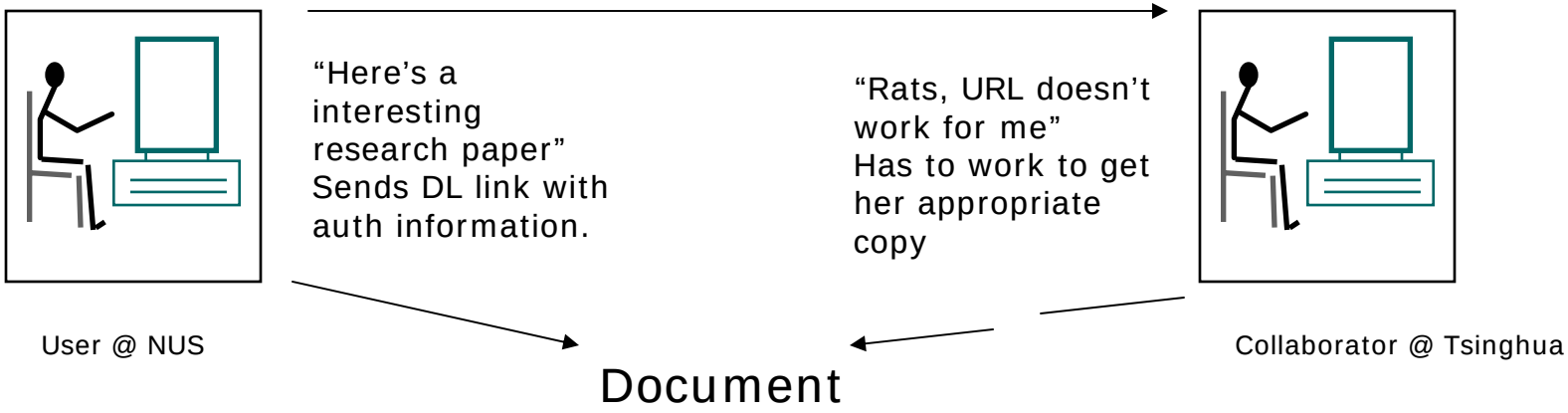


DOI Summary

- Uses multiple levels of indirection
- More robust than PURL
- But also more complicated, relies on central authority
- Supported by consortium of publishers (big and small)

OpenURL

- A identifier system that takes user's context into account
- Created to solve the *appropriate copy problem*



Different providers use different URL and points of access to the data

Citation Databases

Full text

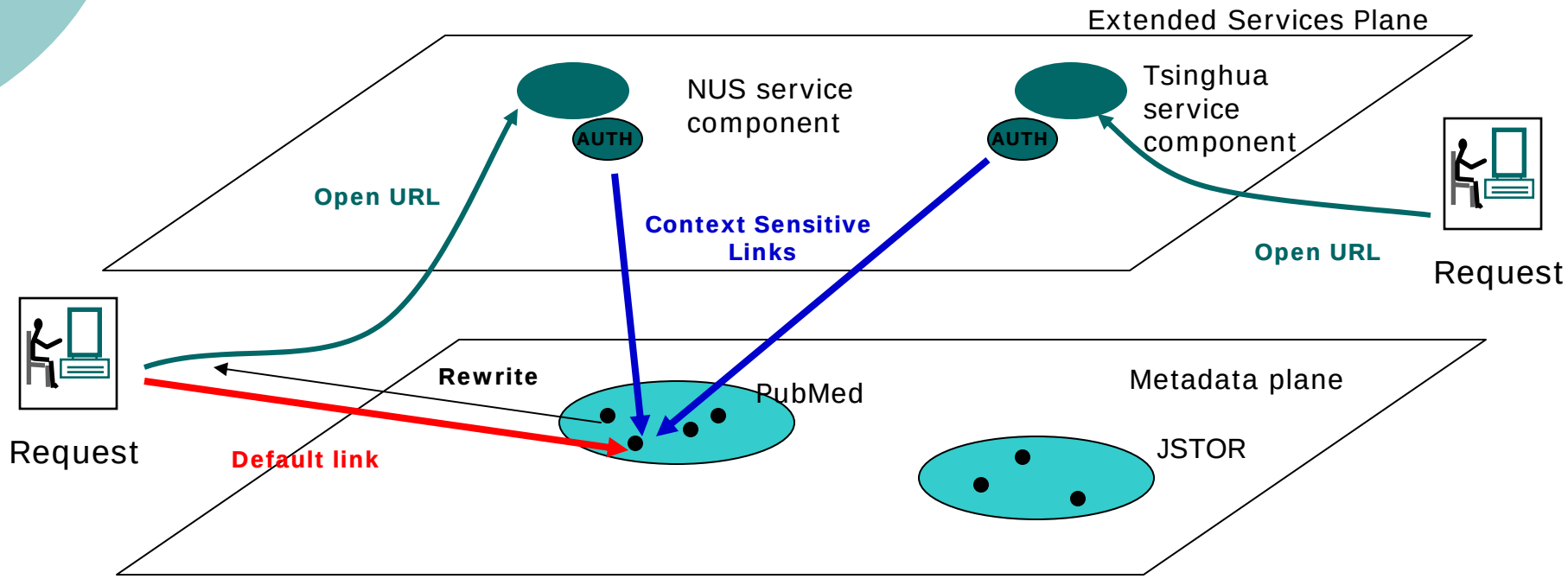
Ulrich's

☐ **Record 1 of 2358 in MEDLINE EXPRESS (R) 1999/01-1999/05**

TI Intraventricular concentration times time (C x T) methotrexate and cytarabine for patients with recurrent meningeal leukemia and lymphoma.
AU Moser-AM; Adamson-PC; Gillespie-AJ; Poplack-DG; Balis-FM
AD Pediatric Oncology Branch, National Cancer Institute, Bethesda, Maryland 20892, USA.
SO Cancer. 1999 Jan 15; 85(2): 511-6.
ISSN 0008-543X
PY 1999
LA ENGLISH
CP UNITED-STATES
MIME
Adolescence-; Adult-; Antimetabolites,-Antineoplastic-administration-and-dosage; Antimetabolites,-Antineoplastic-adverse-effects; Antineoplastic-Agents,-Combined-adverse-effects; Child-; Child,-Preschool; Cytarabine-administration-and-dosage; Cytarabine-adverse-effects; Injections,-Intraventricular; Methotrexate-administration-and-dosage; Methotrexate-adverse-effects; Recurrence-; Remission-Induction; Treatment-Outcome
MJME
*Antineoplastic-Agents,-Combined-therapeutic-use; *Burkitt-Lymphoma-drug-therapy; *Leukemia,-Lymphocytic,-Acute-drug-therapy; *Meningeal-Neoplasms-drug-therapy
TG Female; Human; Male
PT JOURNAL-ARTICLE
RN 0; 0; 147-94-4; 59-05-2
NM Antimetabolites,-Antineoplastic; Antineoplastic-Agents,-Combined; Cytarabine; Methotrexate

Indirection in OpenURL

- Dissociate document from vendor-, library-specific provisions
- OpenURL lists access metadata only



Input: OpenURL Example

Moll JR, Olive & M, Vinson C. Attractive interhelical electrostatic interactions in the proline- and acidic-rich region (PAR) leucine zipper subfamily preclude heterodimerization with other basic leucine zipper subfamilies. J Biol Chem. 2000 Nov 3 ; 275(44):34826-32. doi:10.1074/jbc.M004545200

http://sfx1.exlibris-usa.com/demo?sid=ebsco:medline&aulast=Moll&auinit=JR&date=2000-11-03&stitle=J%20Biol%20Chem&volume=275&issue=44&spage=34826

http://sfxserv.rug.ac.be:8888/rug?id=doi:10.1074/jbc.M004545200

- **Legend:**
  - red - BASE-URL of service component
  - blue - identifier of the resource where the user clicks the OpenURL, added by publisher's rewrite
  - grey - metadata and identifiers
- DOI can be used to resolve the actual content

OpenURL Issues

- Service component gets metadata query information
  - Access and use information goes to library, not to publisher
- Not just user-to-user, but for generalized *dynamic* linking
  - Web page to journal article full-text
  - Abstract to library catalog collection

Demo:
http://www.ukoln.ac.uk/distributed-systems/openurl/

Summary

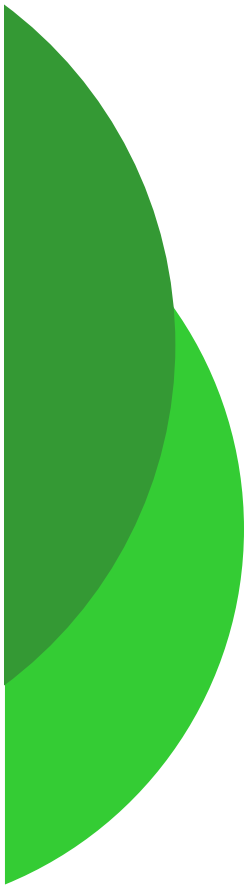
- PURLs
  - Good for small, local solutions
  - Single level indirection
- DOI
  - Multiple, hierarchical layers of indirection
  - Purpose:
    - Actionable identifiers to content
    - Resolution to multiple items of current state data
      - Notably including location(s) and metadata
- OpenURL
  - Purpose: solves appropriate copy problem
  - Selects between multiple items returned by DOI

References

- URN: http://www.w3.org/Addressing/
- PURL: http://www.purl.org/
- DOI: http://www.doi.org/
- openURL: http://www.sfxit.com/open/index.html

Tea break!

- See ya!



Digital Library Policy

Week 4 Min-Yen KAN

Legal, Economical, and Social Aspects

Outline

- Intellectual property rights
- Economics of the (digital) library
- Social Policy with respect to the DL

Jerome's translation of the Bible

- Perhaps the first copyright dispute
- In 521, the Irish missionary Columba secretly copied a very treasured translation of the Bible. When his master Finnian found out, he demanded that Columba turn over the copy. Columba refused and the matter went to the High King of Ireland, Diarmit.

- What do you guess the ruling was?

Two worlds: digital and print media

Print	Digital
Fair Use – for individual purposes or research ○ Fair use was first introduced as an "exception" to copyright	No fair use – "any digital transmission" is considered copying. ○ That means by viewing the web page you have copied it.

Rights Management

- In general,
- “Rights” can mean many things:
  - Access rights – can I see/use/copy it?
  - Intellectual Property Rights (IPR) – who owns it? Where do I go to get access rights?

Access Policy

- We have been mostly concentrating on making the distribution of materials as easy and quick as possible.
- But that's not always the case.

Restricting Access in DLs

- Integrated with the Warwick Framework
  - Cryptolope
  - Steganography / Document watermarking
  - Hardware solutions
- No copy protection
  - Better than it may seem

Copyrights

- Copyright
- Public domain
- Open source

Open Source Licensing

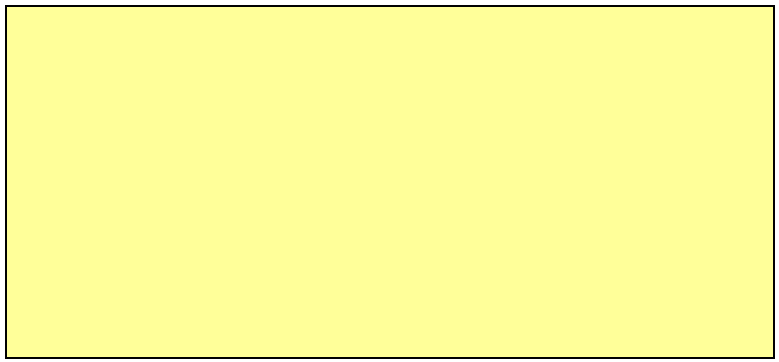
All open source licenses:

- Allow free redistribution,
- Make the source code available
- Allow derived works (modify the code and offer a “new” program)
- Must not discriminate against persons, groups, or fields of endeavor
- Must not be product specific.

- **MIT License** which grants unrestricted rights to copy, modify, and redistribute as long as the original copyright and license terms are retained.
- **BSD License** requires acknowledgements to be made in advertisements and documentation.
- The **Artistic License** allows unrestricted rights to copy, use, and locally modify. It allows the redistribution of modified binary programs, but restricts distribution of modified sources.
- The **GNU General Public License** (GPL) requires that a program that uses portions of GPL'ed source code must also be licensed under the GPL.

Take a quick break: a survey

- How much do you value your library?
- Take a guess! →



- Here's are some ways to do it.
  - What's the cost of buying the sources yourself?
  - What's the opportunity cost if you didn't have access to the information?

A cost model for libraries

- Griffiths & King (93): corporate employees
  - Found that US companies spent about $400-1K per capita on libraries.
  - Reported about 3:1 return on investment

- With library:

$515	Library subscription cost
$95	Library

- No library:

$3300	Cost to access individual materials

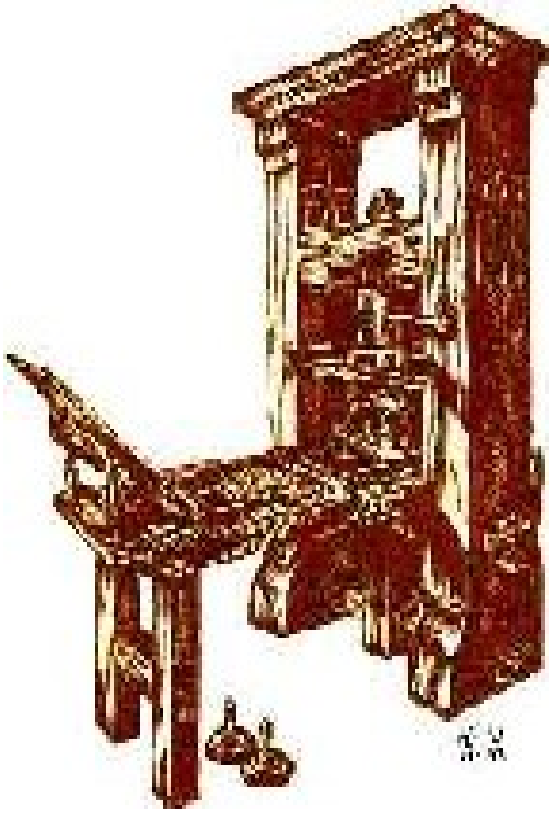
  - These cost only includes buying material, not administrative time in acquiring them.
  - So actual savings is higher.

A brief history of the economics of information

- Ancient Era
  - Public – for religious conversion
  - Private – for knowledge and prestige
- The copying of the Bible by monks in the dark ages
  - To educate them
  - To spread religion

Gutenberg printing press

- Johann Gutenberg (c. 1397-1468):
  - Neither the inventor of moveable type nor printing
  - Paired a wine press with moveable type
- Transformed Europe's spread of information
  - First publication was the Bible
    - Speed allowed mass production and cheaper pricing



The dichotomy today

- Public – ~~for religious conversion~~ government clearinghouse

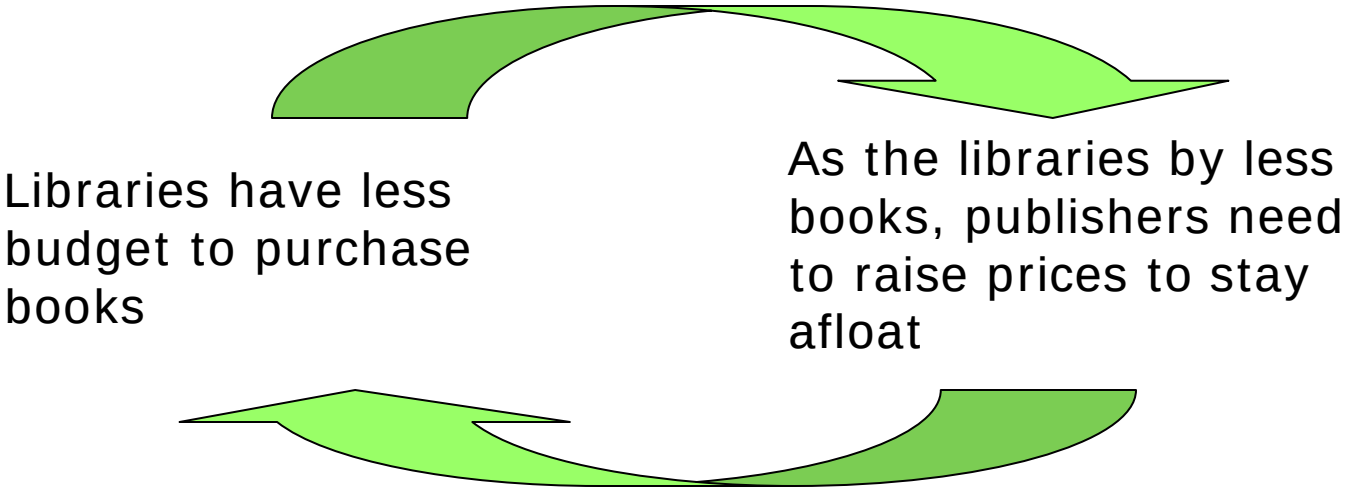
  Make sure the public has:

  - Access to the information
  - Gets authoritative information

- Private – ~~for knowledge and prestige~~ business and entertainment

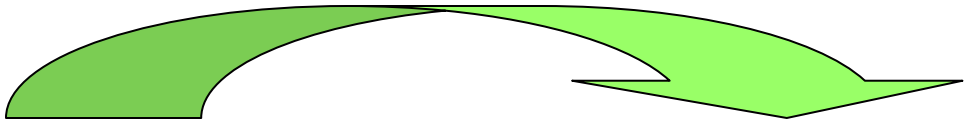
Economics of scholarly media

Will the automated library as we know it survive?



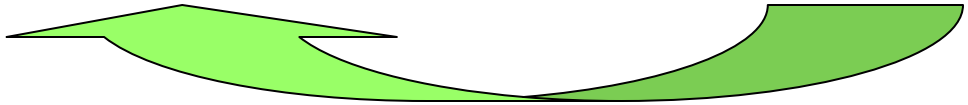
Economics of scholarly media

Will the automated library as we know it survive?



Volume of information published every year grows

But libraries have a fixed budget to purchase materials



Two worlds: digital and print media

Print	Digital
First sale right – you buy it, you can do anything with it (e.g., resell)	Not applicable. ○ How to enforce restricted access to only those who paid?
Have to be in the same place as material	No restrictions
Zero sum distribution (I borrowed it; so you can't)	My borrowing doesn't impede your borrowing
Discrete and self-contained	Continuous and linked

Models for digital economies

- Subscription fees
  - Per month, per year
- Connection time fee
  - Per minute (e.g., Mead Data Central)
- Advertising
  - By an interested party
  - other economic models apply here
- Access fee
  - Per download, may not have profile to remember that you accessed this resource before
- Per-byte fee
  - Typical of connection services (e.g., Broadband)

Access versus ownership

- With DL materials we can't really track ownership, just access
- Trend towards microanalysis
  - Publisher: better targeted marketing
  - Library: better profile of user community

Crisis for publishers

- Ease of publication allows more information to be free
  - And for people to break copyright (perhaps accidentally)
- Ease of accessing (free) information deters users from accessing more cumbersome-to-use sources
- Traditional functions of publishers are taken on by free services
  - Free e-journals do rigorous peer review
  - Search engines act as distributor

Self-archiving

- To deposit a digital document in a publicly accessible website.
  - Preprint: before copyright restrictions have been signed
  - Not a true publication*: hasn't been peer-reviewed, not in prestigious publication.
  - Detractors: accessibility will hurt future revenues of the journal
    - Perhaps 60-80% of a publisher's budget doesn't go towards the direct publication costs

* Debatable, others say it **is** published

E-prints

- Differing acceptance from different fields
  - Physics: accept only if concurrently preprinted
  - Medicine, Business: accept only if not preprinted
- E-journal model: who assumes the cost?
  - Authoring a text
  - Peer review
  - Marketing
  - Editor
  - Publication

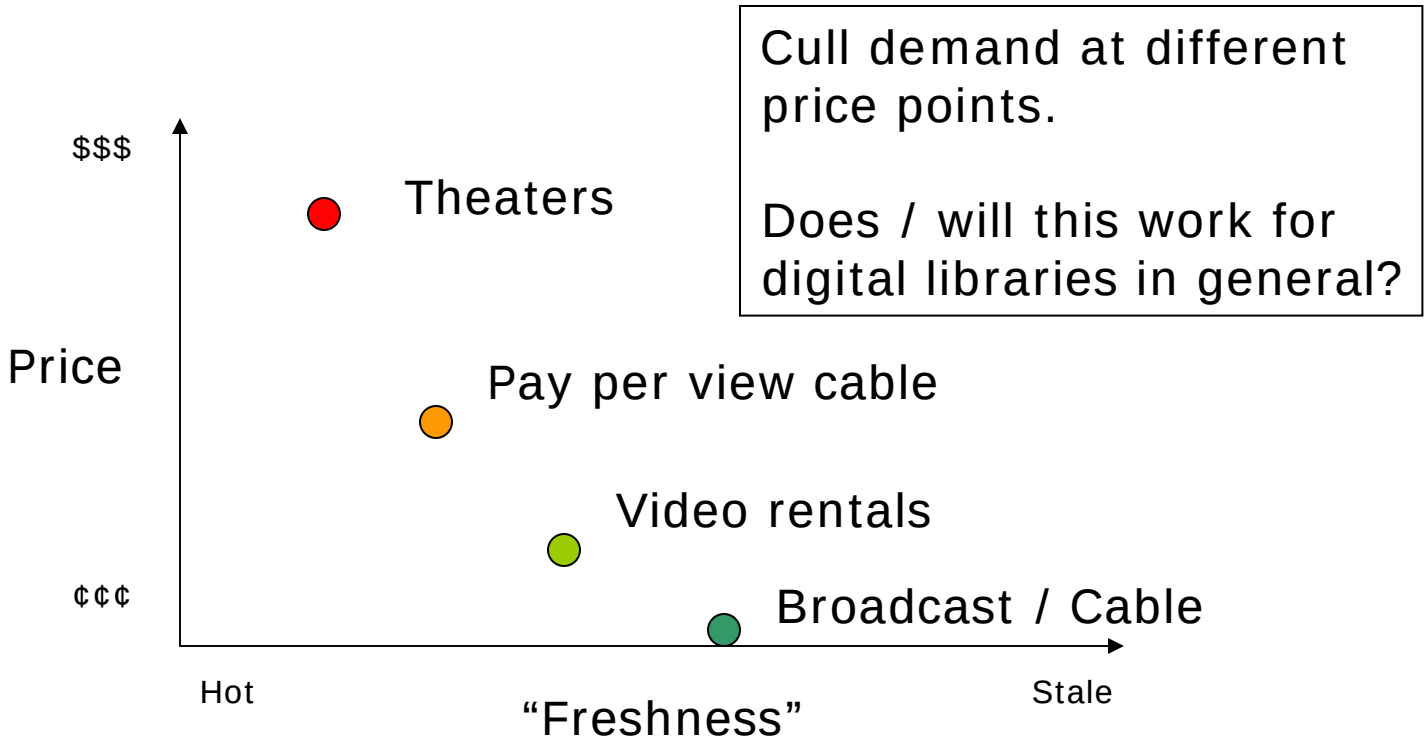
Peer review limitations

- Goal of peer review is to insure:
  - Previous work adequately acknowledged
  - Experimental methodology realistic and reproducible
  - Analysis of data justifies conclusions
- Peters and Ceci (82):
  - Resubmitted 12 psychology articles already published with different author names, 8 of 9 recommended against acceptance and were rejected “serious methodological flaw”, not because of déjà vu.
- Inglefinger study of NEJM reviewers:
  - Concordance of reviews only slightly better than chance
  - Reviewers not skilled in all areas of a study, unable to discern poor writing and have their own biases

Cost structuring

Movie distribution as a possible model

(Lesk, p. 206)



Cull demand at different price points.

Does / will this work for digital libraries in general?

Legal Deposit

Print	Digital
Title IIC gives a copy of every book to the Library of Congress	No legal deposit ○ How to archive the materials of the web? ○ Search engine only covers about 16% of the visible web.

If there's no legal deposit for digital information, how do we archive and preserve website information?

- Especially since search engines only index 16% of the web?

Internet Archive and Bookmobile

Internet Archive

http://www.archive.org

An archive of the www

"The goal of universal access to our cultural heritage is within our grasp."

Are these examples of legal deposit?

Who funds this initiative?

Internet Bookmobile

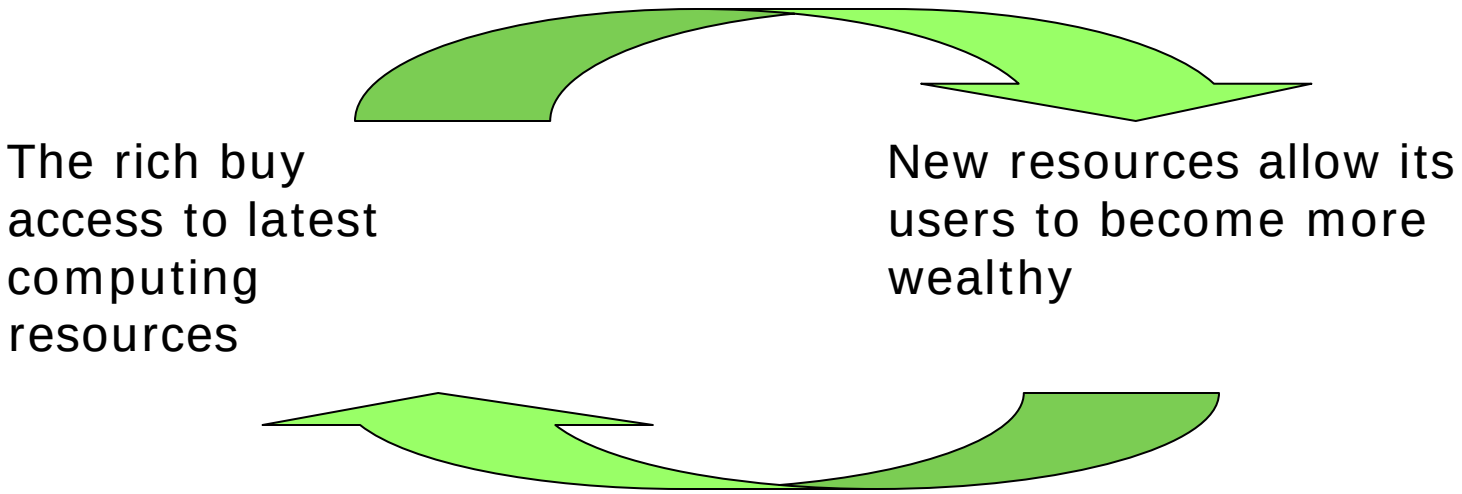
- Prints out of copyright books for reading
- Over 1m books
- $1 USD per book printed

Preservation

- Y2K – two digits to mean four
  - If you knew COBOL, you could get a high paid job.
  - Legacy systems and knowledge need to be preserved
- Use standard formats!
- Media lifetime
  - Tape 15 years
  - CDR 10-50 years
  - HD 30 years
- Software/Hardware lifetime
  - New hardware 3-7 years
  - Software cycles faster
  - How to access old files, applications?

The Digital Divide

- A case of the rich getting richer?



Undoing the Divide

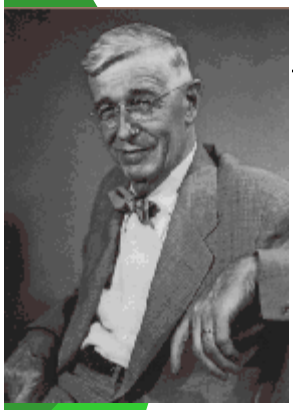
- Can use access rights to impose an unequal payment scheme
  - Blackwell's – all 600 journals made free to the Russian Federation.
  - JSTOR – cost to access its DL depends on the size of the organization.
  - Open source movement – make software available to anyone

JSTOR Logo ® by JSTOR

JSTOR scans and archives past issues of selected journals

• Keeps a moving wall for many publishers to let them produce maximal revenue
• To think about: as an archive repository, what format do you think it keeps its collection in?

Libraries of the Future



- Immediate, random-access to recent knowledge
- May not understand foundation material
- More effort in selection of materials
- Publisher models changing, unifying
- International policy becoming more prominent
- Customized books as the future?

To think about...

- How does the economics of libraries and the information explosion influence publication rates? What about as we make the transition to the digital library?
- Do you think self-archiving and e-journal venues pose a threat to the journal publisher?
- As a single site, the Internet Archives, cannot keep track of all web pages on the web
  - Can you think of a better solution?
  - How would you go about designing a national web page archive for Singapore?

References

- Copyright in Singapore
  http://www.ipos.gov.sg/newdesign/indexpage/inner_frame.html?section=aboutip&sub=4
- Self-Archiving FAQ
  http://www.eprints.org/self-faq/
- JSTOR
  www.jstor.org
- The future of libraries?
  Stephenson, Neal (00) *Diamond Age: A young lady's illustrated primer*, Doubleday