Managing Scientific and Technical Information: A Changing Landscape

A Symposium on Information Management

...In a changing world

24 – 26 October

Hacettepe University, ANKARA

John J. Regazzi
Dean Emeritus, College of Information and Computer Science

Long Island University
October 2007
A seminal shift on June 28, 2007: Information Management vs. Social Computing
Changes in Publishing, Library Services, and Scientific and Technical Information Management

- The Context
- The ‘Shifting Sands’
- What Way Forward
Information Supply Chain

THE CONTEXT

AUTHORS

PUBLISHERS

LIBRARIES

READERS

Historical
Information Supply Chain

THE CONTEXT

Information Supply Chain
Early Online Impact

AUTHORS

PUBLISHERS

LIBRARIES

READERS

A&I SERVICES, ONLINE VENDORS
Information Supply Chain

A simple information “supply chain” has been transformed into a complex information network…
Information Supply Chain

THE CONTEXT

Information Supply Chain

Today

AUTHORS

eBOOK, eJOURNALS
PREPRINT SERVERS
CONTENT AGGREGATORS

SECONDARY PUBLISHING

LOCAL SYSTEMS DDS,
SILVER PLATTER

ONLINE VENDORS
DIALOG, STN, LN

SYNDICATORS & FACTIVA AGENTS

DOCUMENT DELIVERY
SERVICES

NEW ONLINE PORTALS
WEB PORTALS
GOOGLE

LIBRARIES

KNOWLEDGE
MANAGEMENT
SYSTEMS

CORPORATE
INFORMATION
SERVICES

NATIONAL/REGIONAL
CONSORTIA

INFORMATION INTERMEDIARIES

READERS

OPEN ACCESS

OPEN ACCESS
A simple information “supply chain” has been transformed into a (perhaps too) complex information network …

How did this happen?
What Happened?

- Significant Industry Growth
- Discovery & Access Systems Growth
- Directory Data Expansion
- Technology Explosion

Or Did It?
Explosion of Technology

Information Technology Costs
1972 vs. 2003
$ per million transactions

- Storage Characters
- Telecommunications Characters NY - LA
- CPU Power Instructions

Graph showing the decrease in costs from 1972 to 2003 for various types of technology.
Growth of Data

Refereed Academic Journal Growth
1900-2000 Number of Journals

- 1900: 3.3%
- 1910: 3.3%
- 1920: 3.3%
- 1930: 3.3%
- 1940: 3.3%
- 1950: 4.7%
- 1960: 4.7%
- 1970: 4.7%
- 1980: 4.7%
- 1990: 4.7%
- 2000: 4.7%
Growth of Databases

**Number of Databases**
- 1972: 7
- 1992: 8,000
- 2003: 20,000 est.
- Growth Rate: 150% 10YR

**Records in Databases**
- 1972: 2 M
- 1982: 5.2 M
- 1992: 4.5 BN
- 2003: 10 BN est.
- Growth Rate: 122% 10YR
Adjusted figures show that the A&I market is shrinking in overall size.

**Growth of the A&I Market**

- 5.3% Annual Growth
- $830M
- $200M

**STM A&I Global Revenues**

1972-2003

Note: 0-1% growth in constant dollar

- 2% Market Growth since 1999
Shifting Sands

- What Shifts?
  - Publisher Pricing & Publishing Trends
  - University Infrastructure Spending
  - Production, Coverage and Focus
  - Scientists’ and Researchers’ Search Patterns

- And why ...?

- Why is the SciTech Access Market **not** growing but in decline?
Eight-year increases in the CPI vs. tuition and journal prices

- CPI (U.S.): 
  - Average annual increase of 2.3%

- Tuition (US 4 yr universities): 
  - Average annual increase of 6.0%

- Industry journal price: 
  - Average annual increase of 6.5%
Tuition Driven by University Spending not lack of federal funding

Total Educational Revenues

- Net Tuition
- Education Appropriations
Perceived Open Access growth

There is a common perception that open access journals have grown rapidly.

60 new journals per month

* Estimate
Source: DOAJ press announcements/websites
Actual OA growth
Number of Open Access journals by year of origin (not necessarily year of foundation)

Source data: DOAJ.org as of March, 2005

Highlights:
• Open Access journal growth has been slowed down since 2001. There are 98 OA journals started in 2004, down from 218 in 2001, among which 18 are by BMC, one by PLoS and the other author pays is Advances in Electronics Manufacturing Technology, published by Vertilog.
• Only 9% of ~1,400 journals classified as Open Access by DOAJ are author pays.
• Almost all of the (currently) known author-pays titles are published by BioMedCentral

Source: DOAJ
Distribution of titles held per publisher

Number of OA publishers

1098
43
15
3
2
1
1

1 title 2 titles 3 to 5 titles 5 to 10 titles 10 to 50 titles 50-100 titles >100 titles

Number of OA titles in publisher portfolio

• 94% of OA publishers have only 1 title
• 99.6% of OA publishers have under 10 titles
• Only BioMed Central (121 titles) and Internet Scientific Publications (61 titles) have over 50 titles

Source: DOAJ as of March 1, 2005
Journal and article distribution by business model

OA journals on Ulrich’s*  1443
  Not refereed  524
  Ceased  13
  No ISSN  59
  Author pays  130
  Less than quarterly subsidized  319
  Quarterly or more, refereed subsidized journals  343

Non-refereed subsidized journals
  • Ave: 17 articles per journal per year
  • Estimate of 8,900 articles per annum (<1% of STM articles)

Infrequent subsidized journals
  • Ave: 12 articles per journal per year
  • Estimate of 5,000 articles per annum (<0.5% of STM articles)

Higher volume subsidized journals
  • Ave: 50 articles per journal
  • Estimate 17,000 articles per annum (~1.5% of STM articles)

Author pays journals
  • Ave: 17 articles per journal per year
  • Estimate 2,000 articles per annum (<0.5% of STM articles)

• Only a small portion of OA titles are comparable to typical commercial journals
• Articles published in subsidized and AP journals remains a small portion of overall STM content

* Estimated (based on sampling 100 journals) to have 95% overlap with DOAJ 1455 titles on 03/02/05
1 Average of PLoS and BMC 2004 publications (121 out of 130 journals)
2 Based on sample of 80 randomly selected journals in set
3 Based on sample of 70 randomly selected journals in set

Source: Ulrich’s database
Total growth of OA articles - 2004
Articles published in OA journals

2003: 24516*
2004: 34538**

41% increase

* Based on sampling of 821 DOAJ journals appearing on DOAJ in Spring 2004
** Based on sampling of 1443 OA journals catalogued on Ulrich’s database on 03/01/05 as detailed in “Author Pays and subsidized OA journals”
Source: DOAJ, Ulrich’s database, Market Development
Records hosted in institutional repositories

Distribution of records hosted on repository categories: 1-4

- **Subject Area repositories**
  - 84% of records hosted on Pub Med Central and ArXiv
  - >90% is STM articles/content

- **Institution wide repositories**
  - Hosts wide range of articles, images, working papers, memoranda, etc.
  - 76% have under 100 records
  - Current focus of OA movement

- **Department repositories**
  - Small repositories capturing output of dept

- **Aggregators**
  - "Point" to articles hosted elsewhere (e.g. in other repositories or on websites)
  - CiteSeer (comp. sci) and RePEc (economics) have subject area focus
  - Some mirror other sites, e.g. Citebase links to ArXiv, PMC and BioMed Central

- Subject area repositories and aggregators link to the largest proportion of STM content
- Institution wide, and dept. repositories remain nascent, and link to articles as well as a range of other content

Source: OAister 03/09/05;
Growth of IRs – Example: Installation and usage of EPrints IR software*

Installation of, and uploading content into, EPrints institutional repositories

- The establishing of EPrints institutional repositories increased 40% over 2004
- Uploading of content has similarly followed a steep trajectory
- Growth of other IR platforms (e.g. DSpace) has followed suit

* EPrints offers open source IR software for installing and managing an institutional repository
Source: http://archives.eprints.org (generated through ‘analyses’ link)
Subject area institutional repositories

Top ten subject area repositories (accounts for 95% of subject area records)
Archived records* accessible through OAIster

- **All articles also archived in arXiv.org**
- **All articles also archived in PubMed Central**

Source: OAIster, March 9, 2005; Market development analysis

- PMC and arXiv account for 84% of all records in subject area repositories
- Most of remaining major repositories are similar in size to BMC

* Records may include author manuscripts, conference proceedings, dissertations, and other text documents; not all records hosted on IRs may be harvestable by OAIster; not all records accessible through OAIster are freely available

** All articles also archived in arXiv.org
*** All articles also archived in PubMed Central
Institution-wide repositories

Top ten general institution repositories (accounts for 69% of general institution records)
Archived records* accessible through OAIster

- University of Cambridge DSpace Repository
- Australian National University (ANU) DSpace Repository
- NASA Technical Report Server (NTRS)
- University of Illinois Archives
- University of California eScholarship Repository
- DSpace at MIT
- ETH (Eidgenössische Technische Hochschule Zürich) E-Collection
- SMARTech: Scholarly Materials and Research at Georgia Tech
- e-Prints Soton, University of Southampton

At least 80% of contents are image files

- Content hosted in ‘general’ IRs is spread across a number of institutions
- Several ‘large’ repositories (e.g. Cambridge) actually host small amount of potential journal content

* Records may include author manuscripts, conference proceedings, dissertations, images, and other media; not all records hosted on IRs may be harvestable by OAIster

Source: OAIster, March 9, 2005; Market development analysis
Leading aggregators are CiteSeer and Citebase (these aggregators utilize meta-data to provide analysis of, and access to, online documents)

* Records may include author manuscripts, conference proceedings, dissertations, reviews and other text files; not all records hosted on IRs may be harvestable by OAIster; analysis excludes OAIster itself with access to over 5.1M records (not all of which are freely accessible)

** Citebase harvests metadata from ArXiv, CogPrints and BMC

Source: OAIster, March 9, 2005; Market development analysis
OA movement and institutional repositories

Recommendations from the recent Berlin3 OA Conference (Southampton, February 28-March 1):

In order to implement the Berlin Declaration institutions should:

1. Implement a policy to **require** their researchers to deposit a copy of all their published articles in an open access repository

2. **Encourage** their researchers to publish their research articles in open access journals where a suitable journal exists and provide the support [i.e. pay the processing fees] to enable that to happen.
Authors as readers: views on journal access (Ciber Study 2004) of authors expressing an opinion, \( n=3,754 \)

Access: now vs 5 years ago

- lot worse
- worse
- same
- better
- much better
After Six Months, An Article has Delivered Only 30% of its Lifetime Value

- After 6 months, an article has delivered only 30% of its full lifetime value, as measured by total lifetime “readings”.
- By mandating articles to be posted online after 6 months, NIH would make 70% of an article’s value freely available.
- By requiring deposit in institutional repositories for all articles immediately upon publication, publisher revenues are put at risk – the question is how much.

Source data:
Heer, “Article Aging”.
Some Preliminary Indications . . .

- Open Access journals are not growing, and does not seem to be meeting the fundamental market needs of current authors.

- “Traditional” society and commercial publishers seem to be experimenting with hybrid models, offering pay to publish options for authors, and these experiments seem to be driving open access articles.

- Subject institutional repositories are showing good growth, but thus far this has been driven by the availability of enabling technologies.

- It is not clear how the key elements of scholarly communication, i.e., verification, registration, dissemination, and preservation are financed adequately over the life of an article if articles are simply transferred by authors or mandate, from commercial or non-profit sectors to the public domain.
The Shifting Sands

University Infrastructure Spending

Research & Library Spending Growth
1975-1995

Index


US Total Academic R&D

Avg ARL Library Expenditure

4.6% CAGR

2.2% CAGR
University Infrastructure Spending

Expenditures at (Private) US University A

- Academic Support: 63.0%
- Library: 1.3%
- Other: 37.5%
A&I Production & Coverage

STM Database Expansion – Duplication of Effort

Shifting Sands
Creation of inefficient market where library pays multiple times for same record, plus (often) increase distribution markups

Current Indexes Overlap
2001 Study

Significant Redundancies in Public & Private Sectors

Records in Databases

<table>
<thead>
<tr>
<th>Year</th>
<th>Science</th>
<th>Education</th>
<th>Social Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>1972</td>
<td>2 M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1982</td>
<td>5.2 M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>4.5 BN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>10 BN est.</td>
<td>55%</td>
<td>53%</td>
</tr>
</tbody>
</table>

122% 10 YR Growth Rate

10 BN est.
Scientists’ and Researchers’ Search Patterns

Online Experience
> 3 years using internet%

<table>
<thead>
<tr>
<th>Category</th>
<th>Total</th>
<th>MD’s</th>
<th>PhD’s</th>
<th>Over 55</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>70%</td>
<td>82%</td>
<td>86%</td>
<td>91%</td>
</tr>
</tbody>
</table>

Work-related Internet Use
Days per Week

<table>
<thead>
<tr>
<th>Days per Week</th>
<th>Weekly</th>
<th>1-3 Days</th>
<th>4-5 Days</th>
<th>6-7 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>95</td>
<td>15</td>
<td>33</td>
<td>47</td>
</tr>
</tbody>
</table>

Increase to 65% in over 55 group

Work-related Search Engine Use
Days per Week

<table>
<thead>
<tr>
<th>Days per Week</th>
<th>Weekly</th>
<th>1 Day</th>
<th>2-6 Days</th>
<th>7 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>86</td>
<td>12</td>
<td>41</td>
<td>33</td>
</tr>
</tbody>
</table>
Scientists’ and Researchers’ Search Patterns

Early shift in Access Models to desktop

**Shifting Sands**

Experimentation by primary publishers with free access to metadata

- IEEE+
- Google
- Cross Search

Information Supply Chain Historical
Shifts in Mindshare

Who has the mindshare lead?

<table>
<thead>
<tr>
<th>Top 3 Online Scientific Search Resources</th>
<th>Librarians VS Scientists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science Direct</td>
<td>42%</td>
</tr>
<tr>
<td>ISI Web of Science</td>
<td>37%</td>
</tr>
<tr>
<td>Medline</td>
<td>31%</td>
</tr>
<tr>
<td>Google</td>
<td>39%</td>
</tr>
<tr>
<td>Yahoo</td>
<td>10%</td>
</tr>
<tr>
<td>PubMed</td>
<td>11%</td>
</tr>
</tbody>
</table>

Who has the mindshare lead?
A Sea of Change

But what will researchers need and will search engines alone be able to deliver

- Scientists and Researchers have needed to become much more pragmatic
- Researchers increasingly work in teams – less emphasis on research and more on development as seen in top five information sources used
- Corporate research teams are involved from design to market launch and need information that is traditional (journals, reports, etc.) but also business and regulatory
- Data Integration/Data Mining
Scientists still have to integrate different information sources in a related context, be they sources of internal & external information, or scientific and business information.

BCG estimates $282M (33% of R&D costs) could be saved with an integrated content platform.

Making informed decisions: The top three barriers identified by Outsell for corporate scientists:
- Inability to compare across information sources
- Determining quality, credibility and accuracy
- Knowing what is available

Key driver to improved productivity in biopharmaceutical R&D, recognized by McKinsey, is successful integration of data.
Agricultural Engineering Needs

**Search the database to produce a list of possible pests.**

**Datasheet provides a route to further text, pictures and maps.**

**CABI COMPENDIUM**

**TREATMENT**

**ASSESSMENT**

**CONTEXT**

Compare pest and crop distribution
Summary

Who can deliver these viable information resources?
Well neither content nor search is king here!

Rather it is those organizations which can

- Filter and select
- Structure the content
- Provide the essential information at the right time
- Provide a context – ‘a sense-making’ tool

In short, the future belongs not to those that merely navigate us through cyberspace, nor those who populate it with data. Rather it belongs to those who help us make sense of all that is available to us.