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

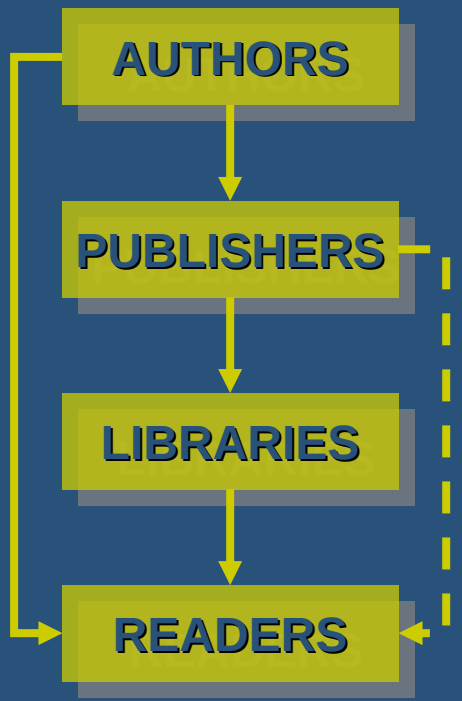
A seminal shift on June 28, 2007:
Information Management vs. Social Computing

Changes in Publishing, Library Services, and Scientific and Technical Information Management

- The Context
- The ‘Shifting Sands’
- What Way Forward

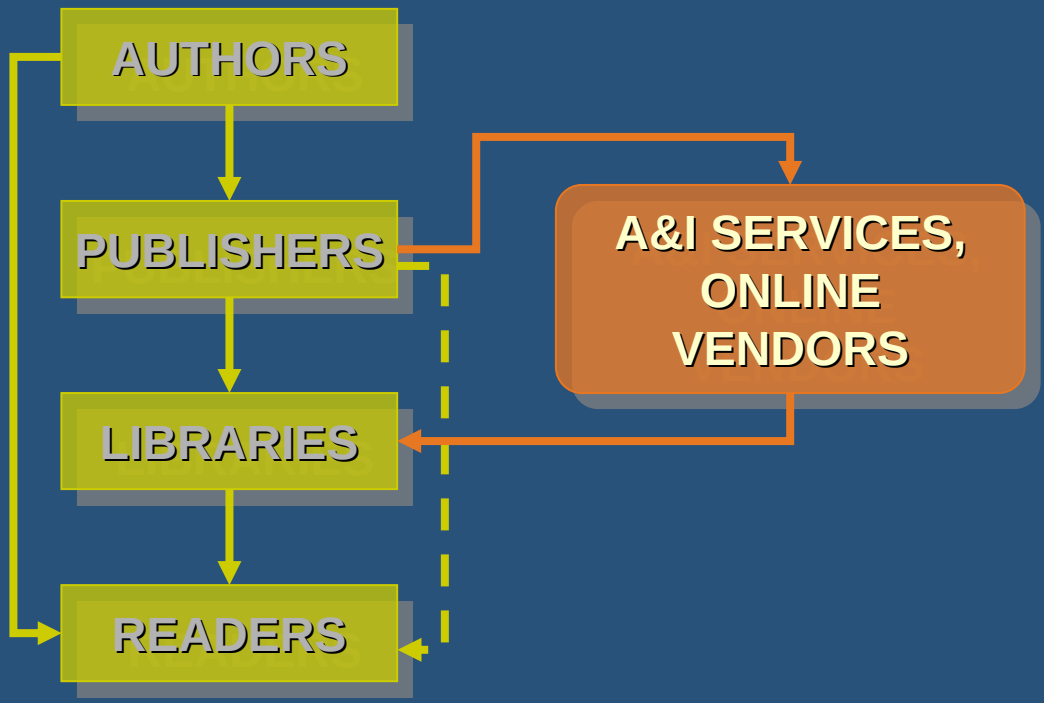
Information Supply Chain

Information Supply Chain
Historical



Information Supply Chain

Information Supply Chain
Early Online Impact

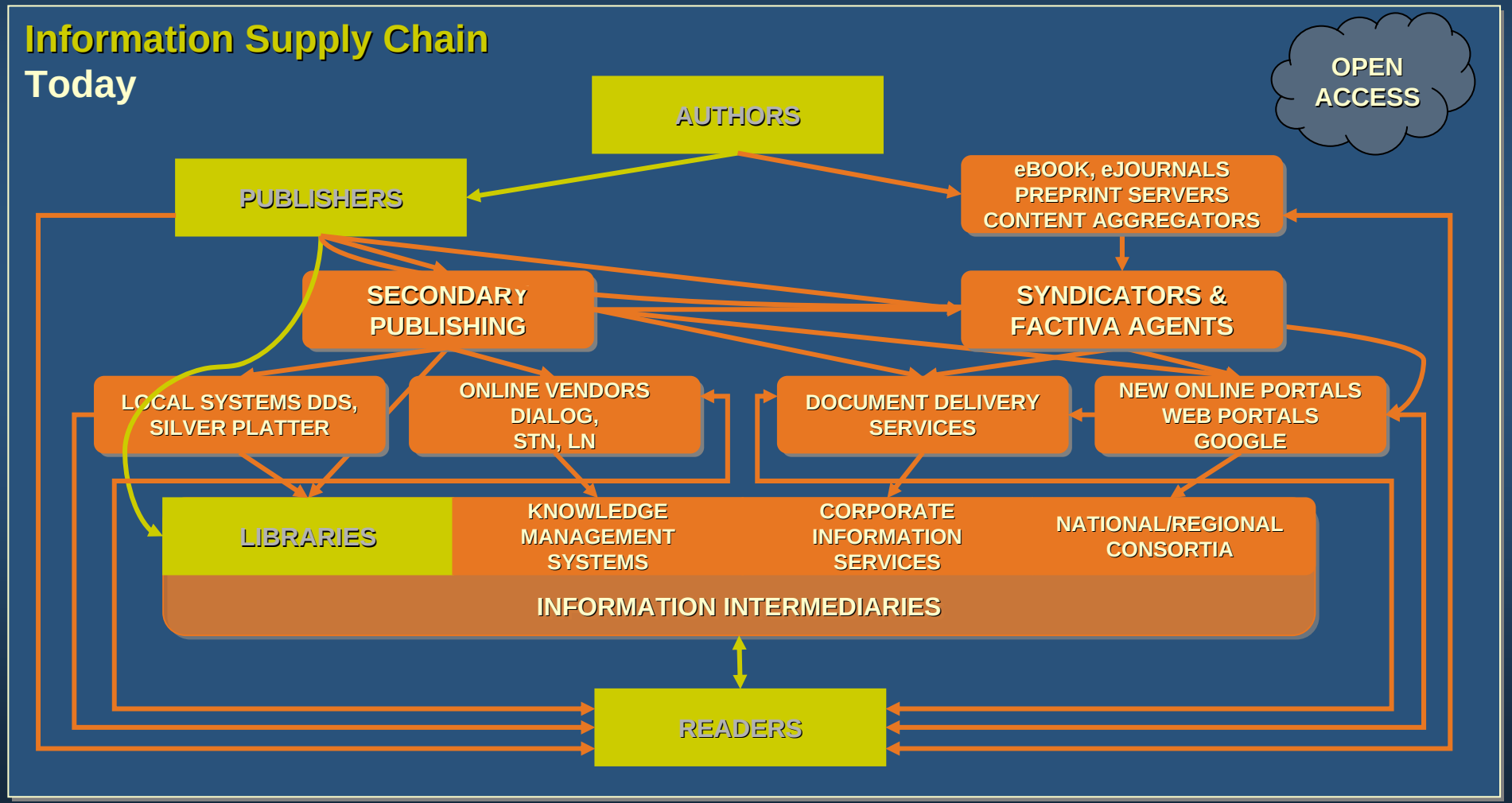


Information Supply Chain

A simple information “supply chain”

has been transformed into a complex information network…

Information Supply Chain



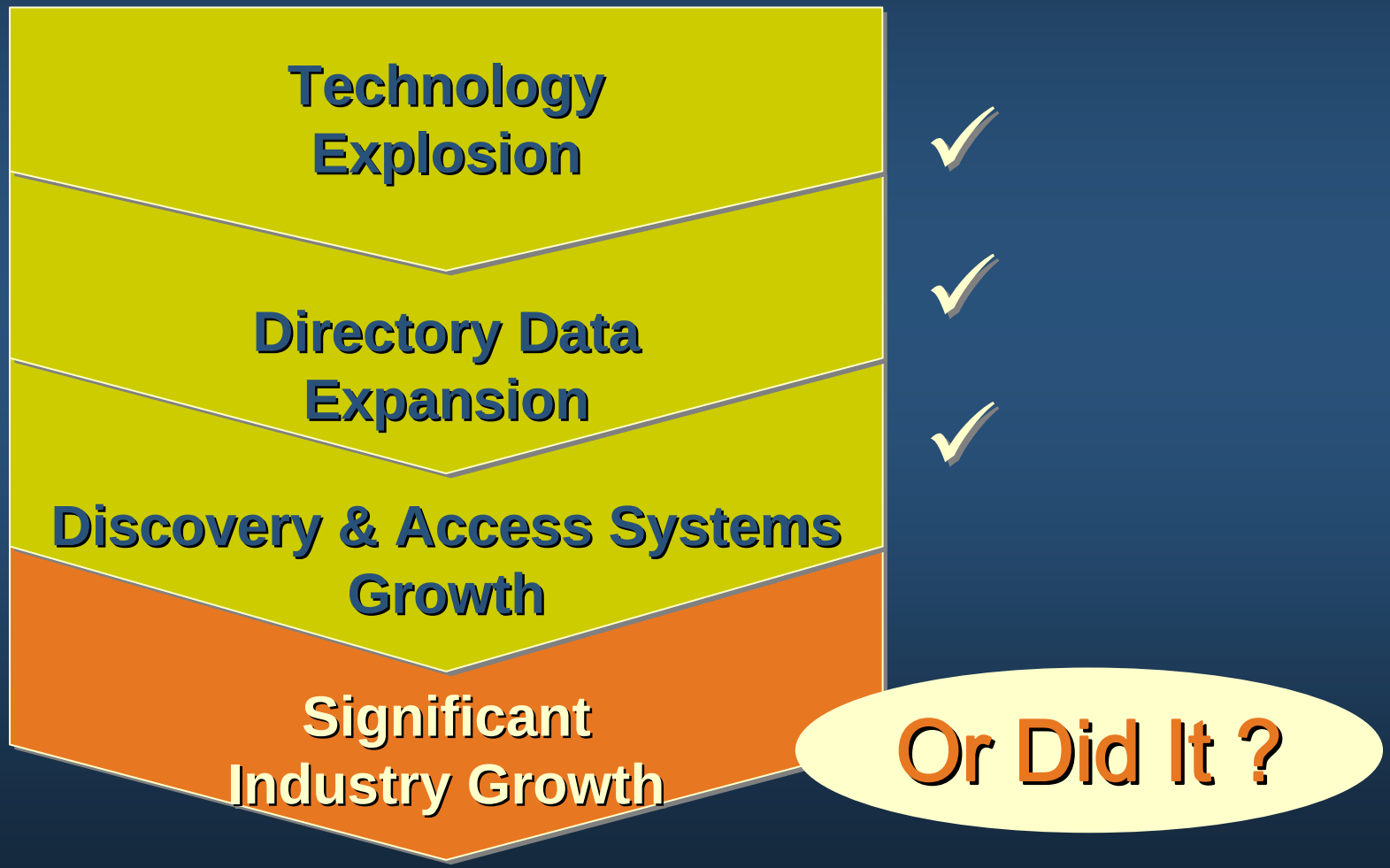
Information Supply Chain

A simple information “supply chain”

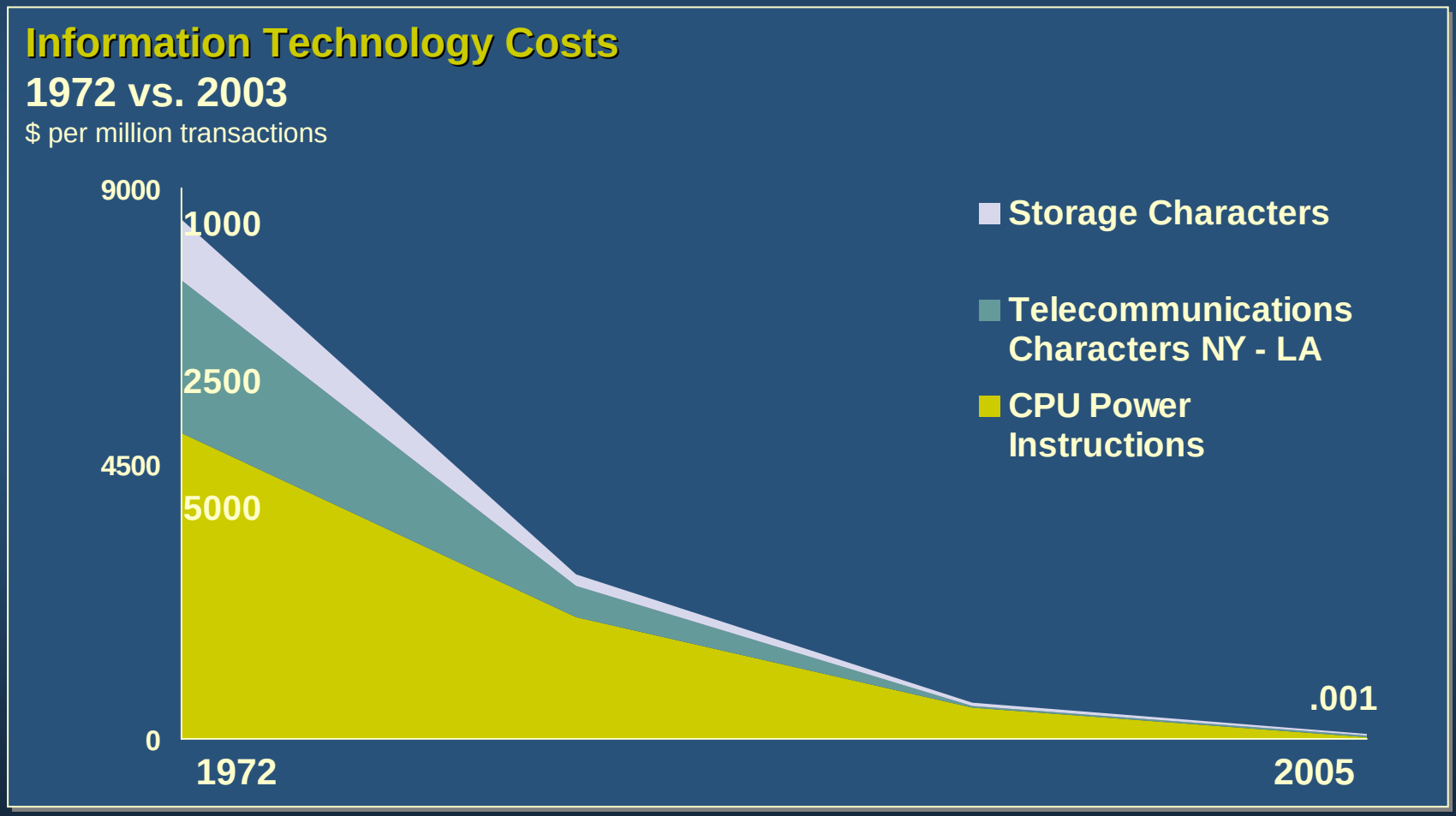
has been transformed into a (perhaps too) complex information network …

How did this happen?

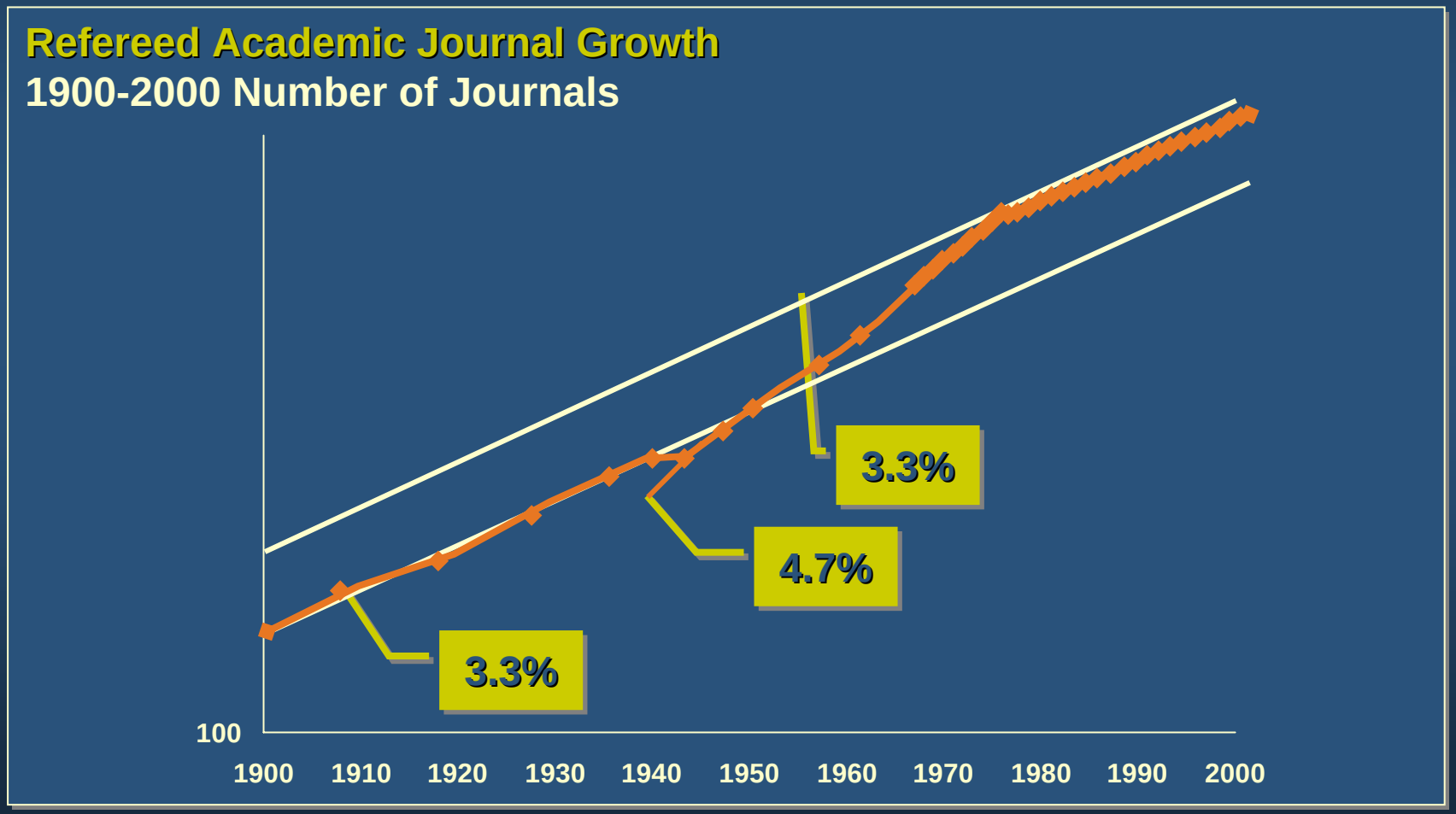
What Happened?



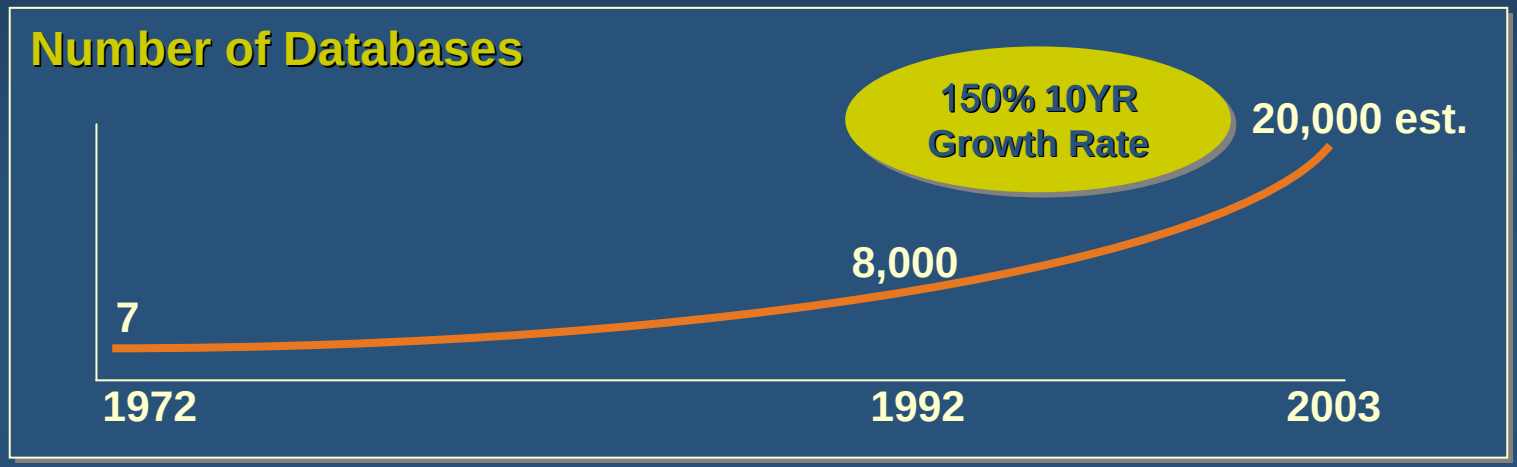
Explosion of Technology

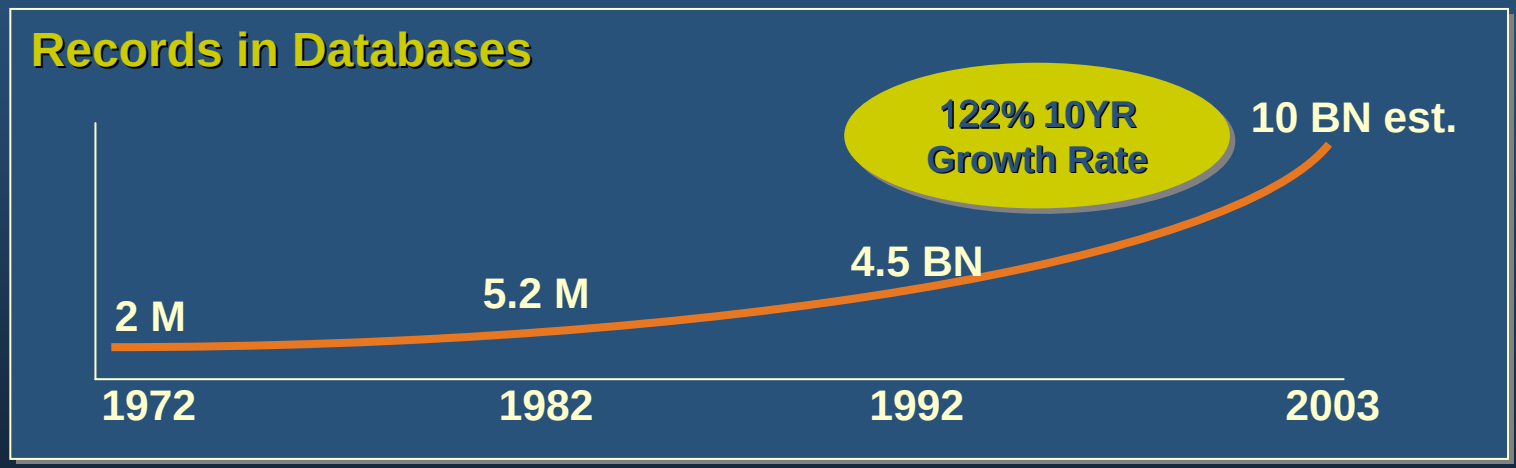


Growth of Data



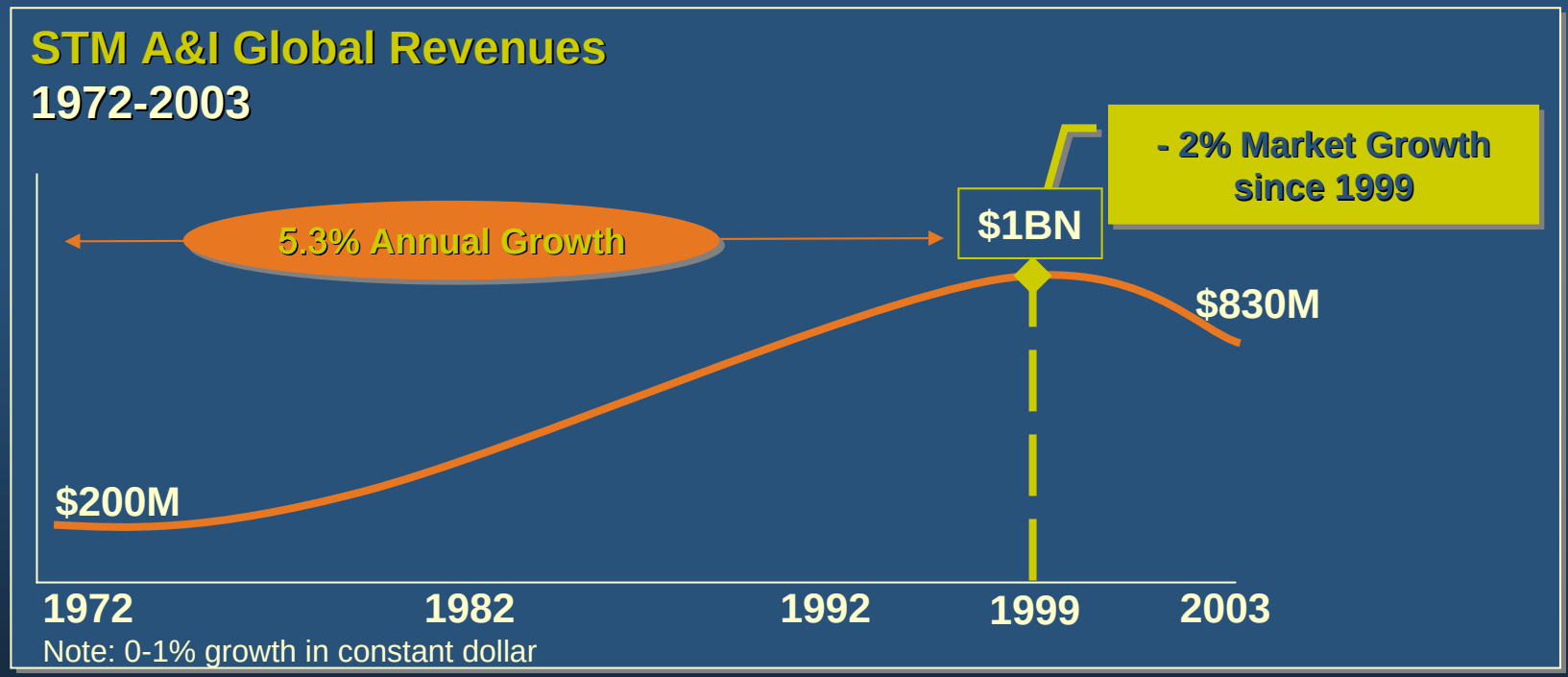
Growth of Databases





Growth of the A&I Market

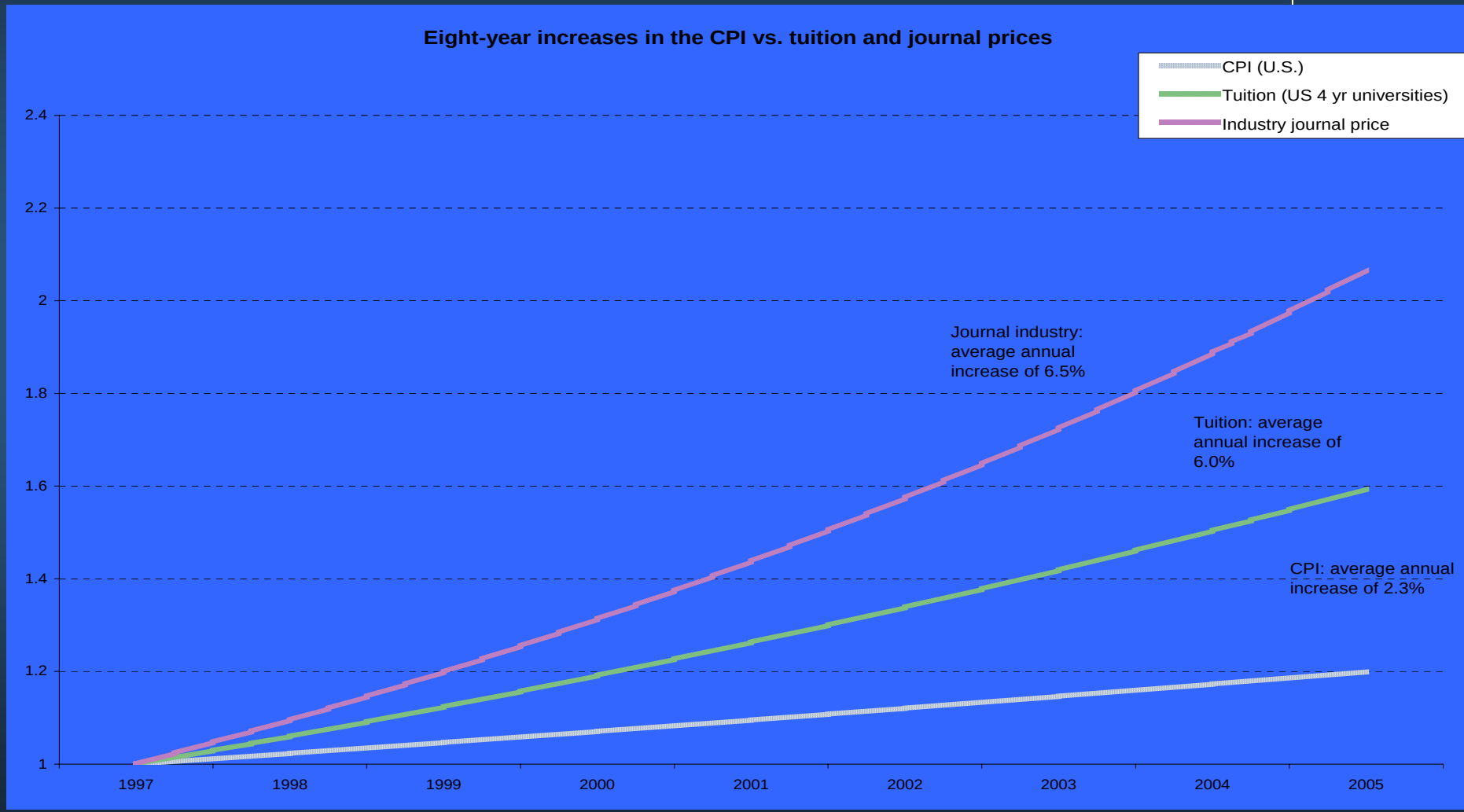
- Adjusted figures show that the A&I market is **shrinking** in overall size.



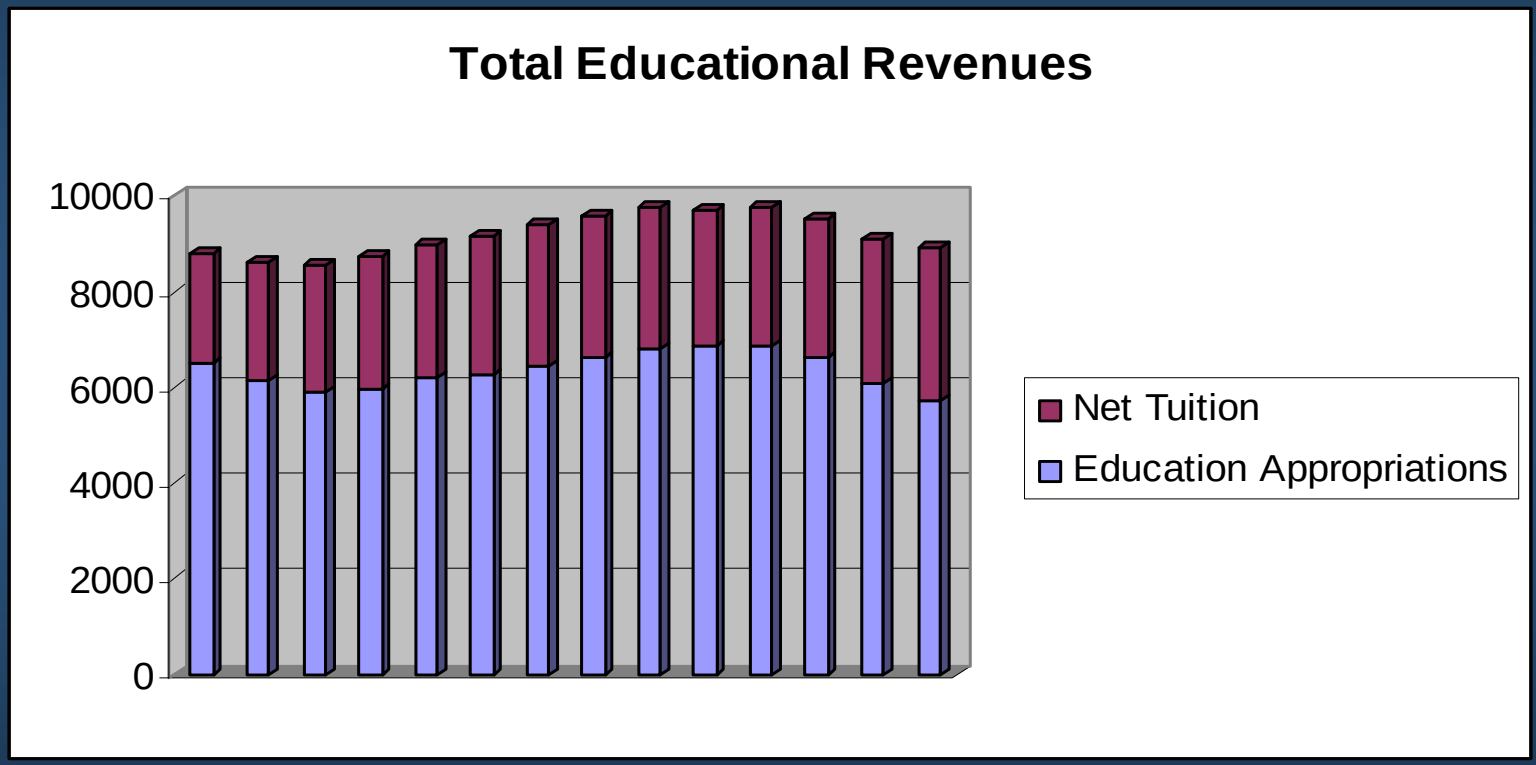
Shifting Sands

- 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?

Publisher Pricing vs. other trends

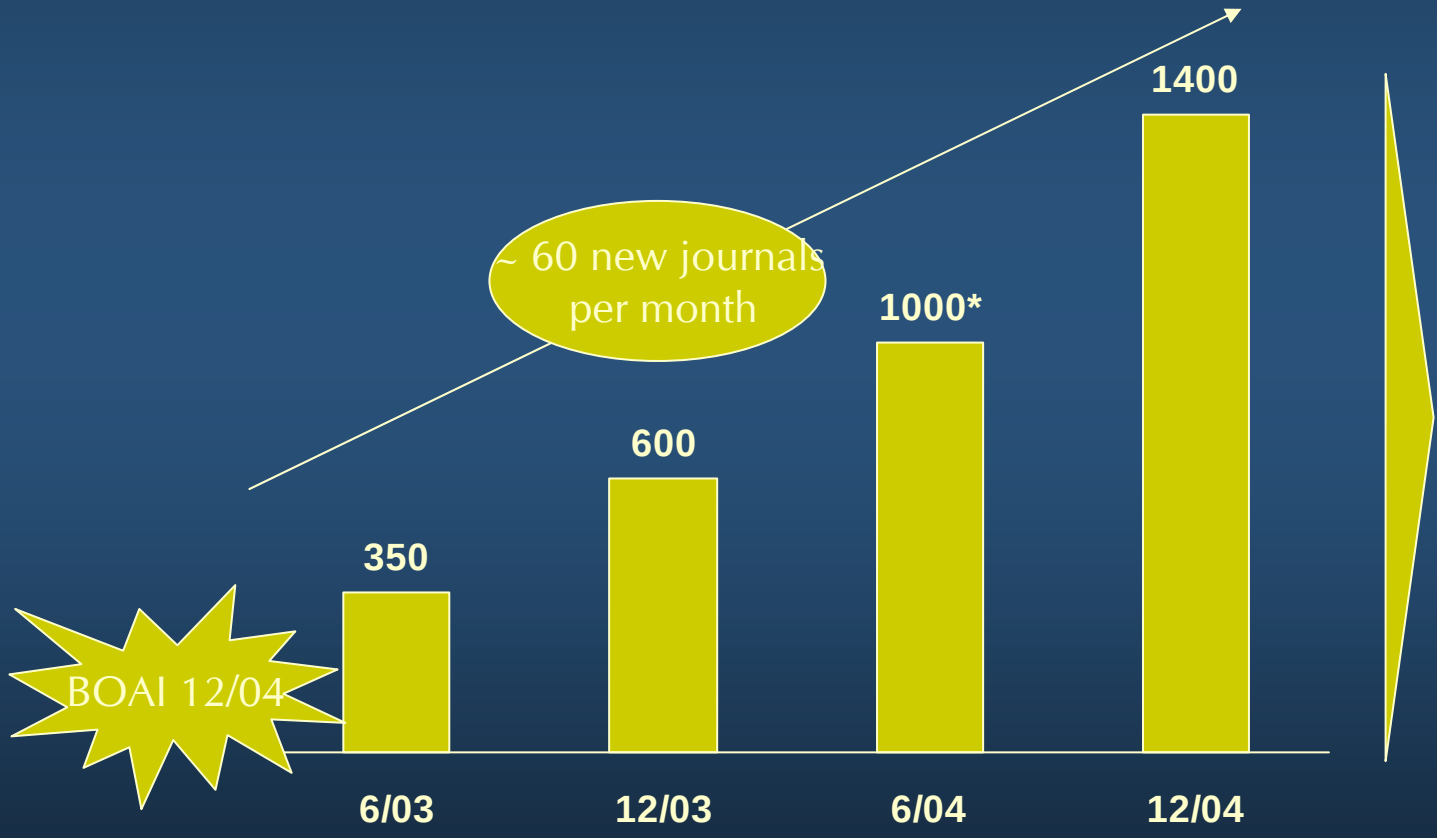


Tuition Driven by University Spending not lack of federal funding



Perceived Open Access growth

Titles accessible through the Directory of Open Access Journals

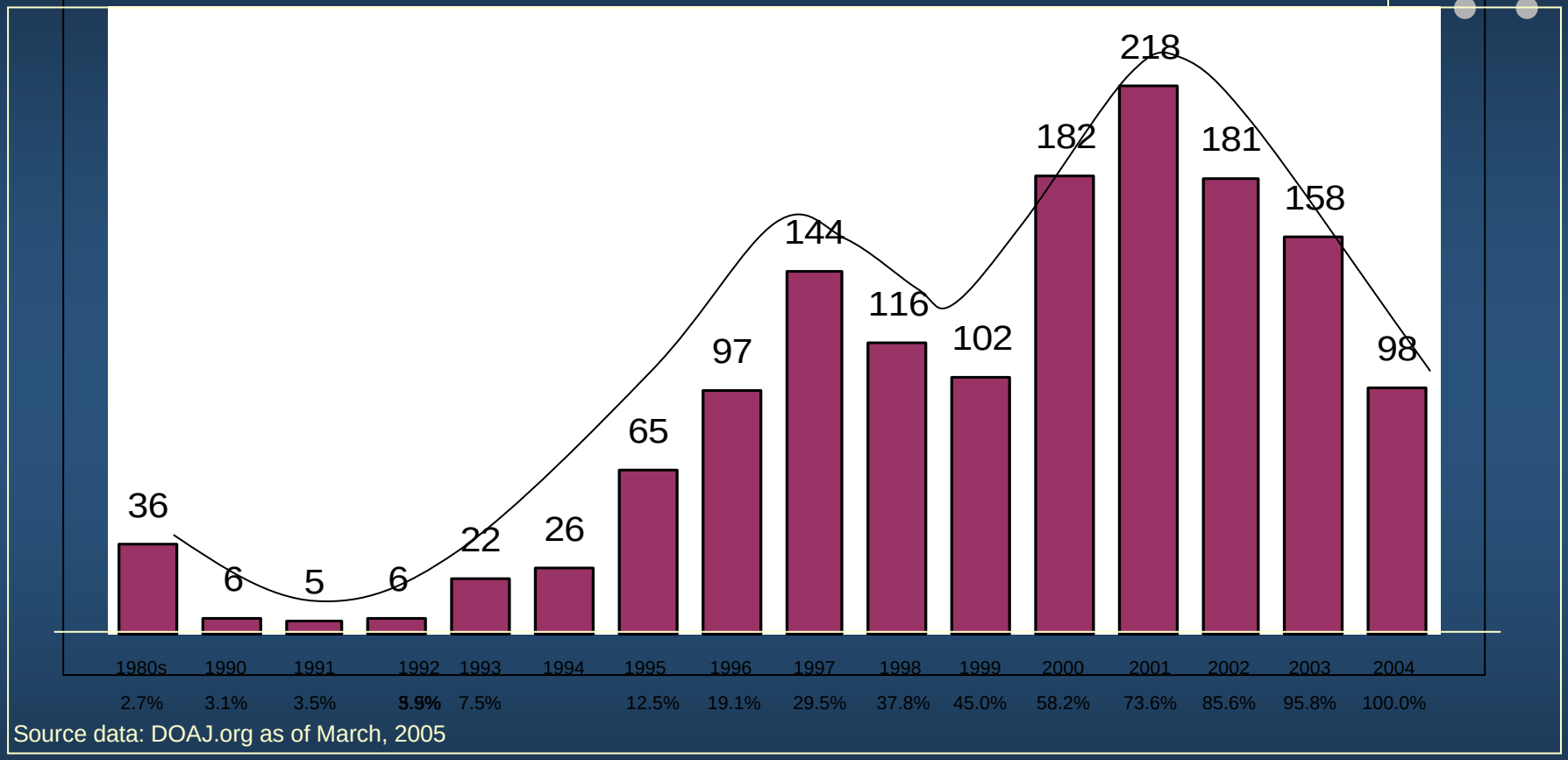


There is a common perception that open access journals have grown rapidly

* Estimate
Source: DOAJ press announcements/websites

Actual OA growth

Number of Open Access journals by year of origin (not necessarily year of foundation)



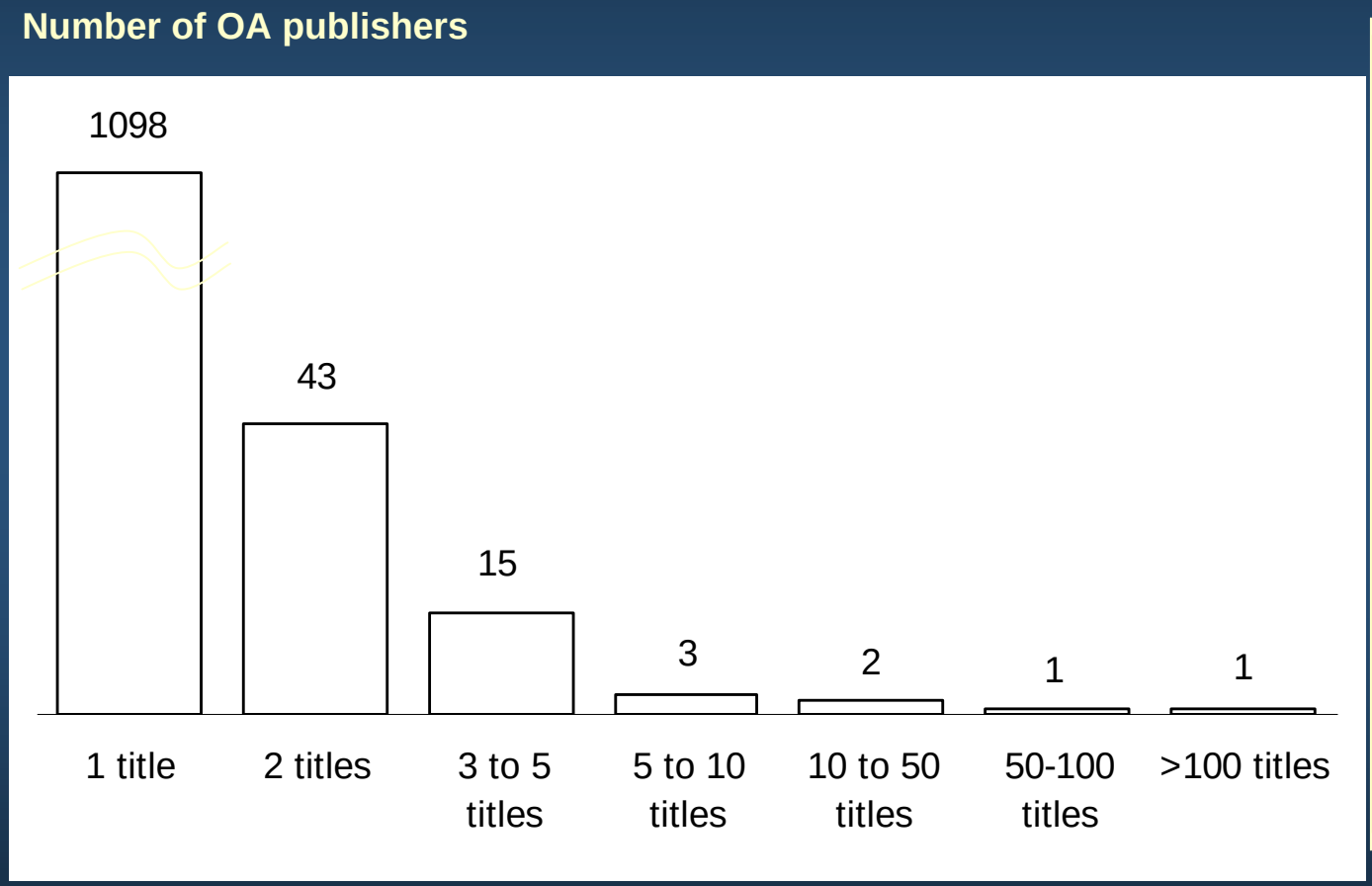
Source data: DOAJ.org as of March, 2005

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

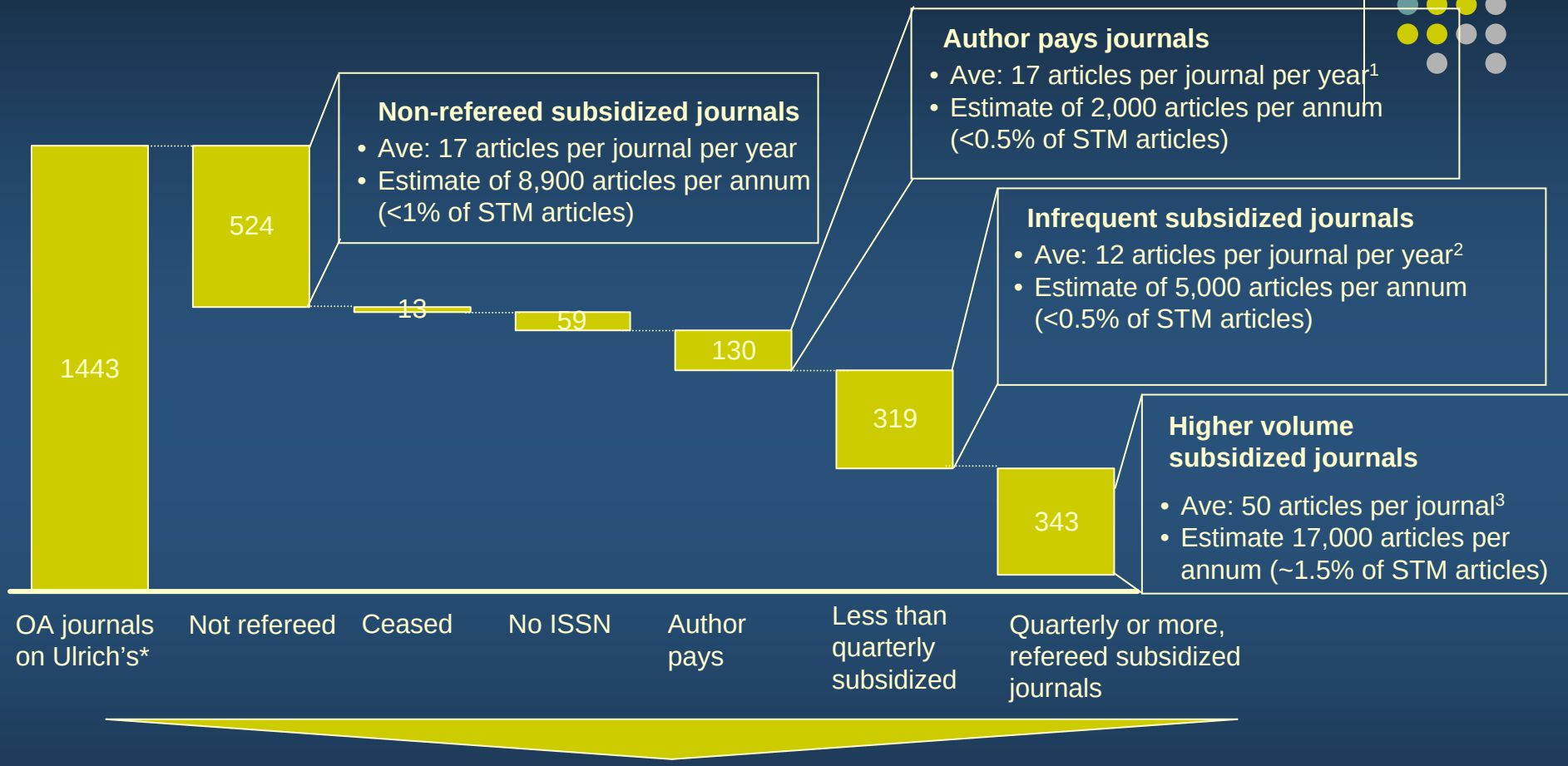
Distribution of titles held per publisher



- 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



- 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)

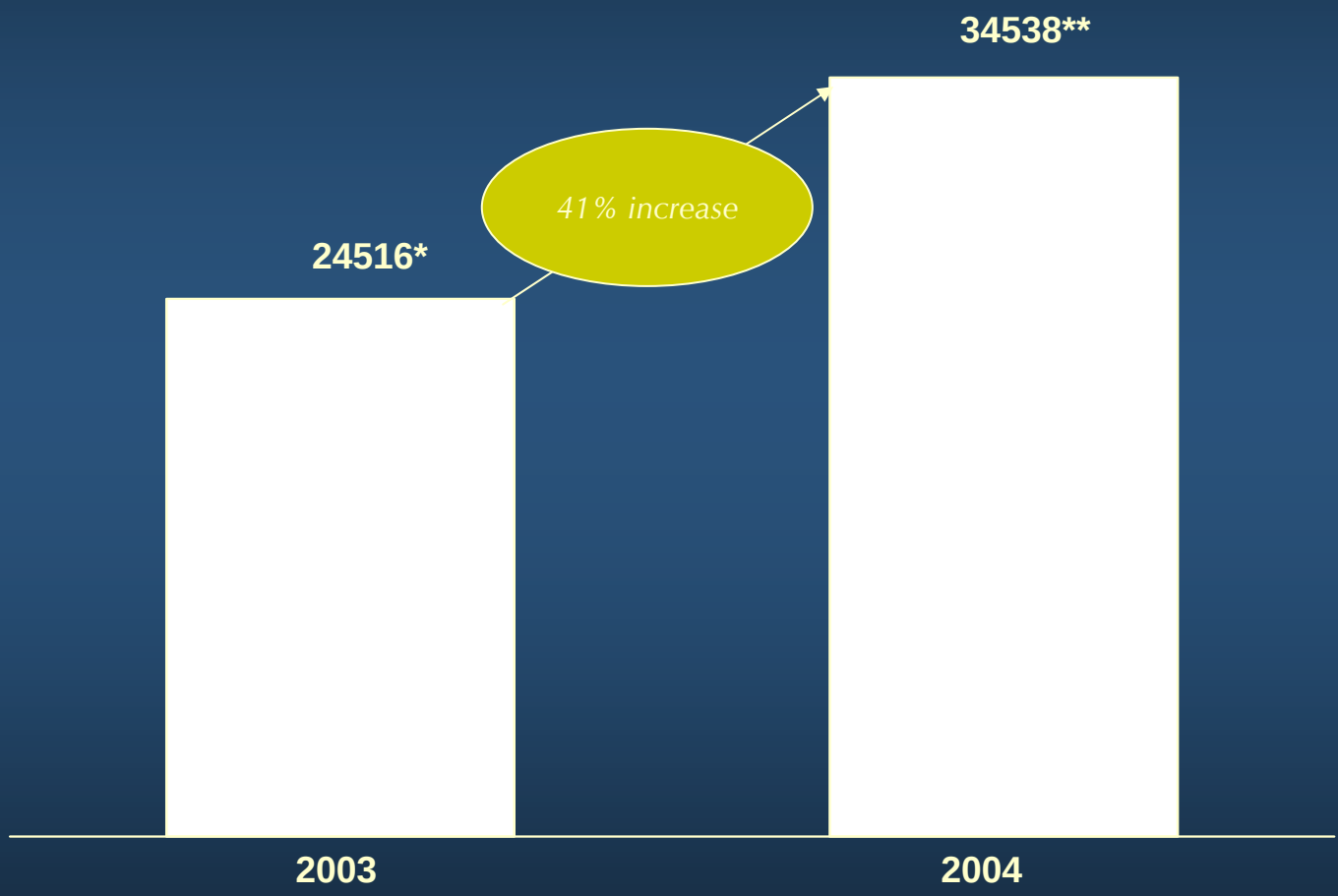
2 Based on sample of 80 randomly selected journals in set

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



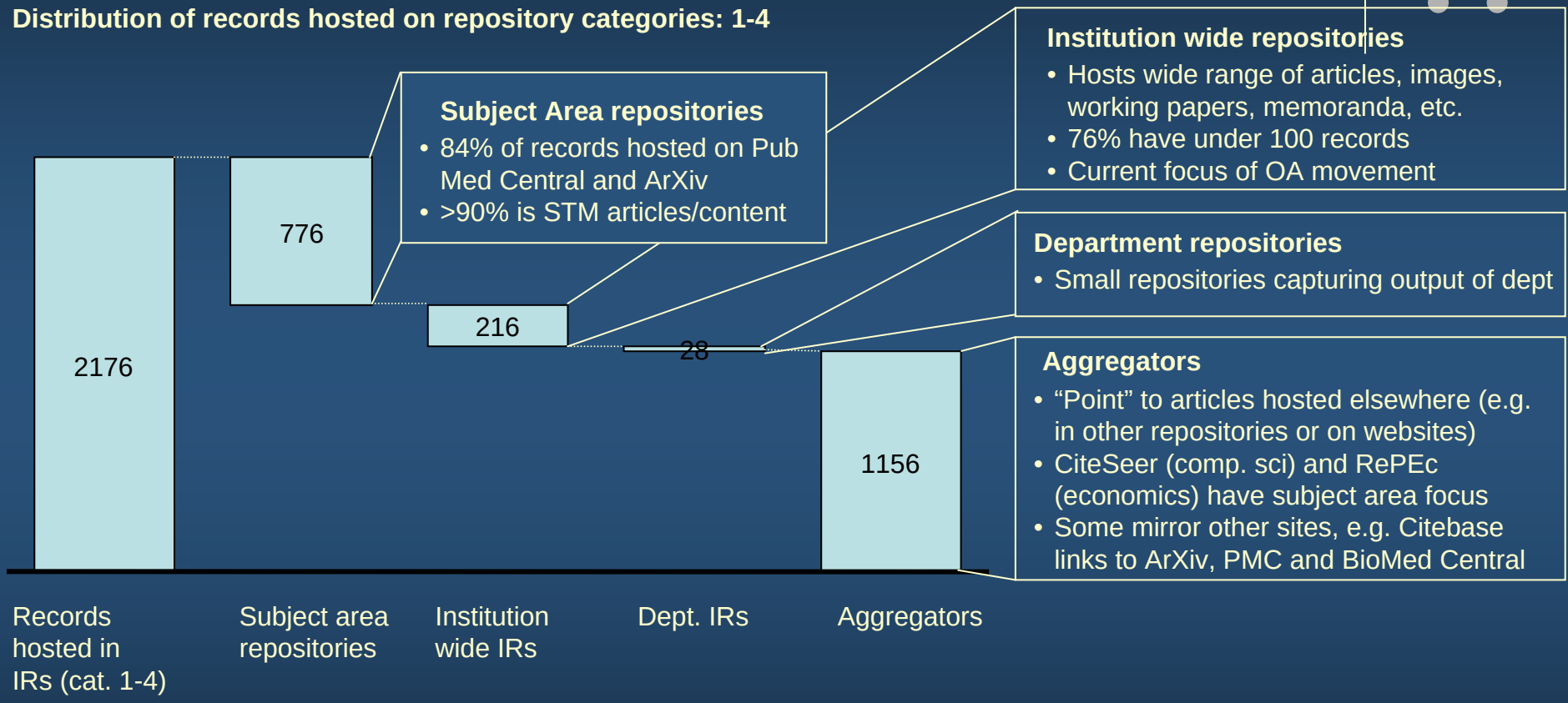
* Based on sampling of 821 DOAJ journals appearing on DOAJ in Spring 2004

** 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

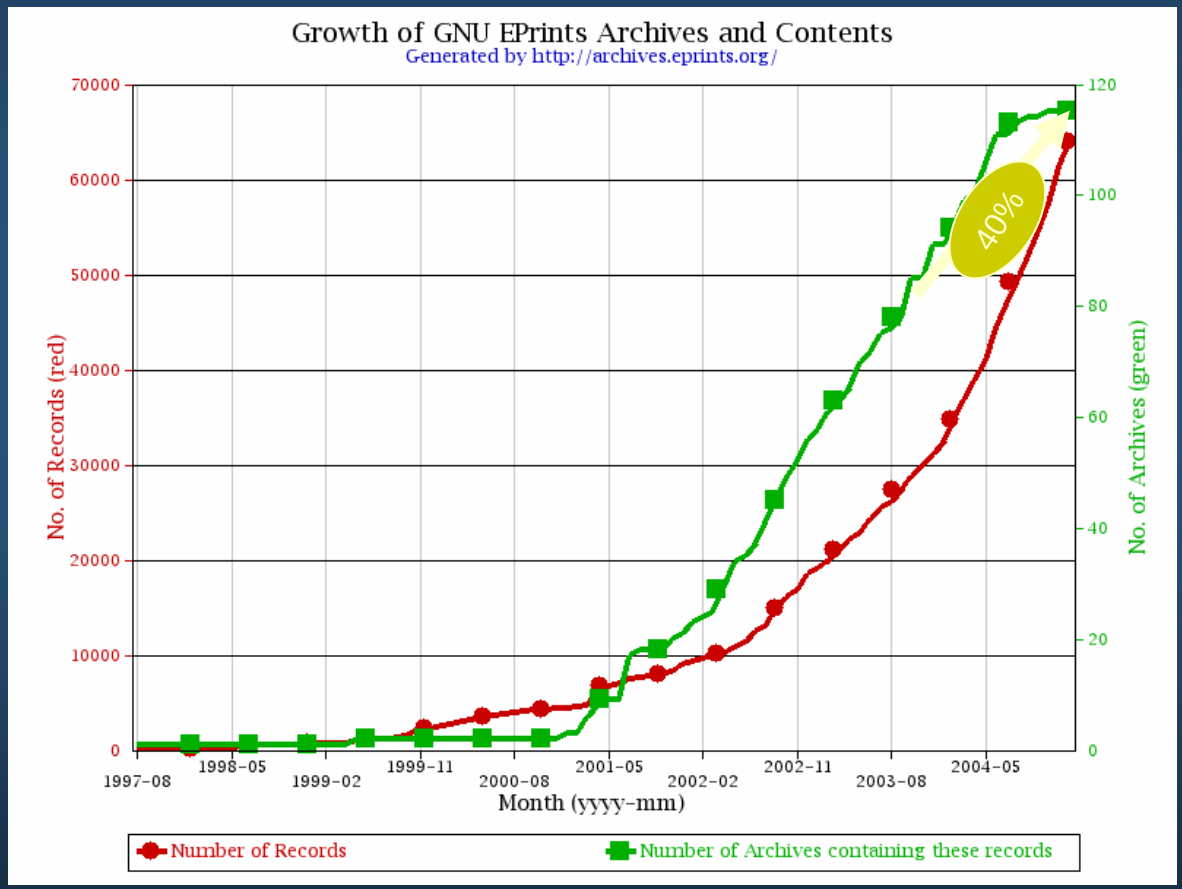
Aggregators
- "Point" to articles hosted elsewhere (e.g. in other repositories or on websites)
- CiteSeer (comp. sci) and RePEc (economics) have subject area focus
- Some mirror other sites, e.g. Citebase links to ArXiv, PMC and BioMed Central

- 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

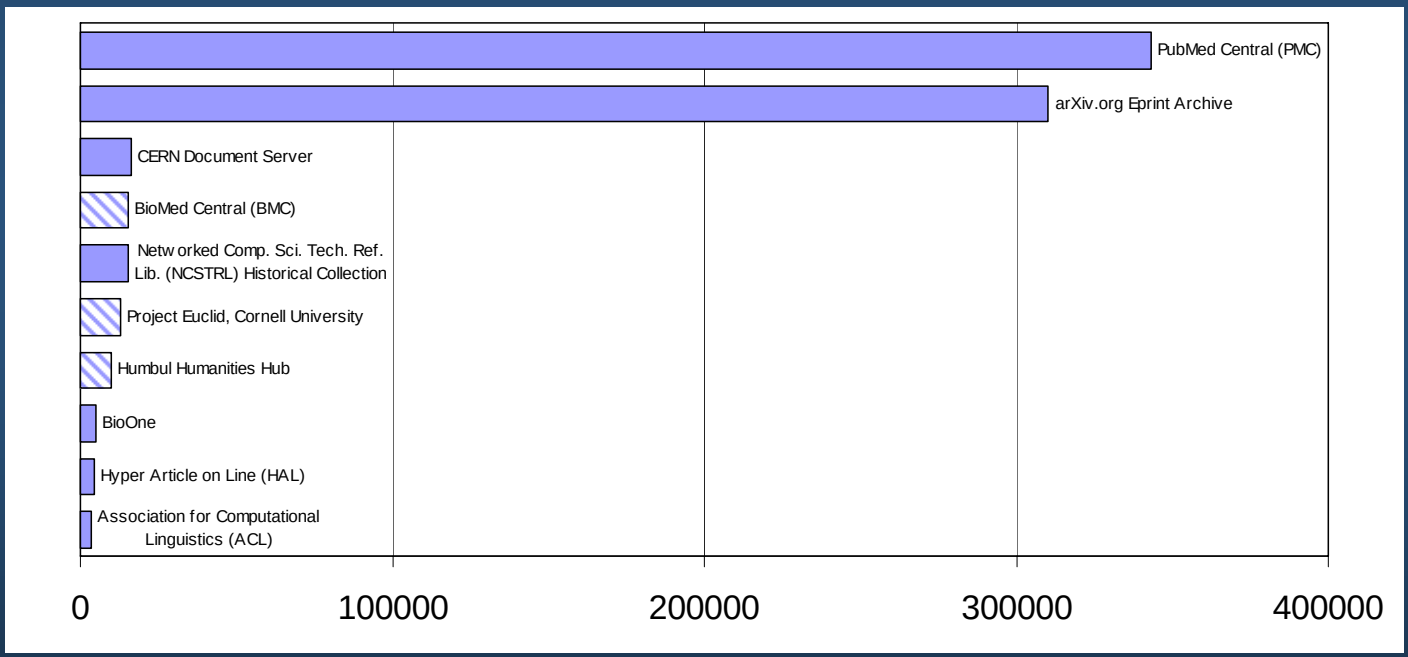


- The establishing of EPrints institutional repositories increased 40% over 2004
- Uploading of content has similarly followed a steep trajectory
- Growth of other IR platforms (e.g. DSpace) has followed suit

* 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



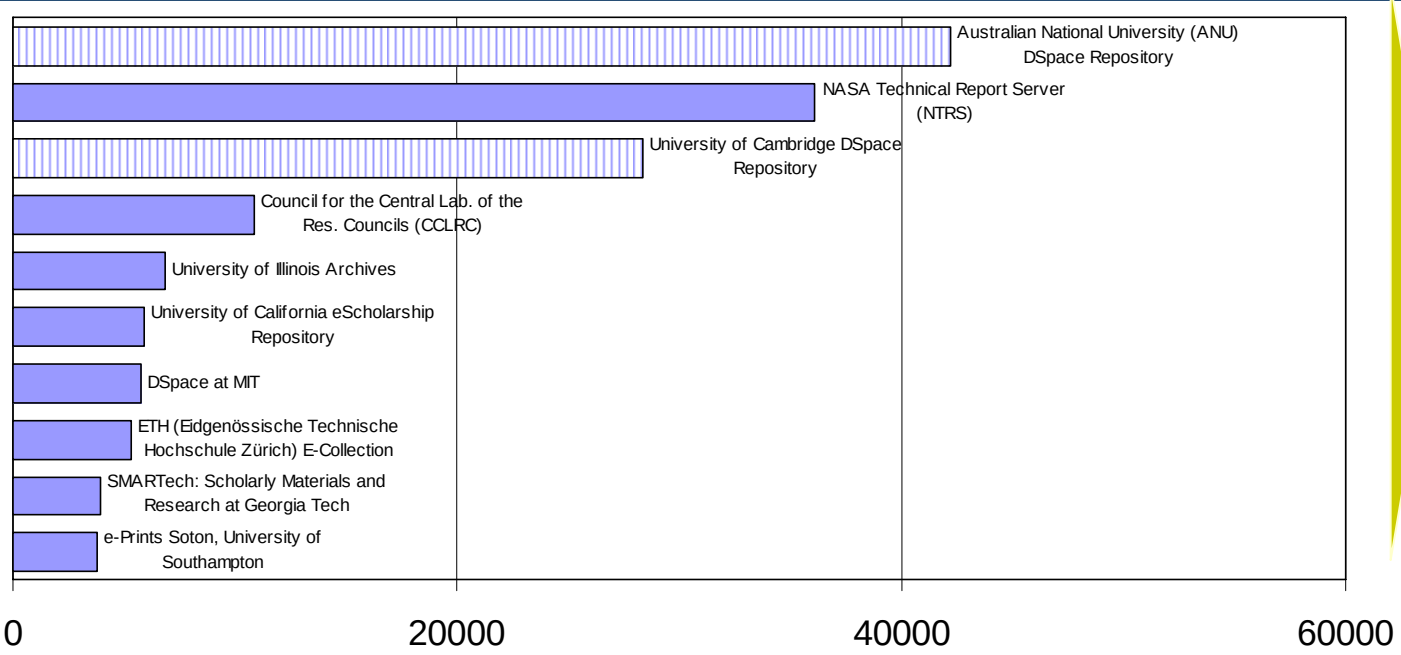
- PMC and arXiv account for 84% of all records in subject area repositories
- Most of remaining major repositories are similar in size to BMC

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

Institution-wide repositories

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

At least 80% of contents are image files



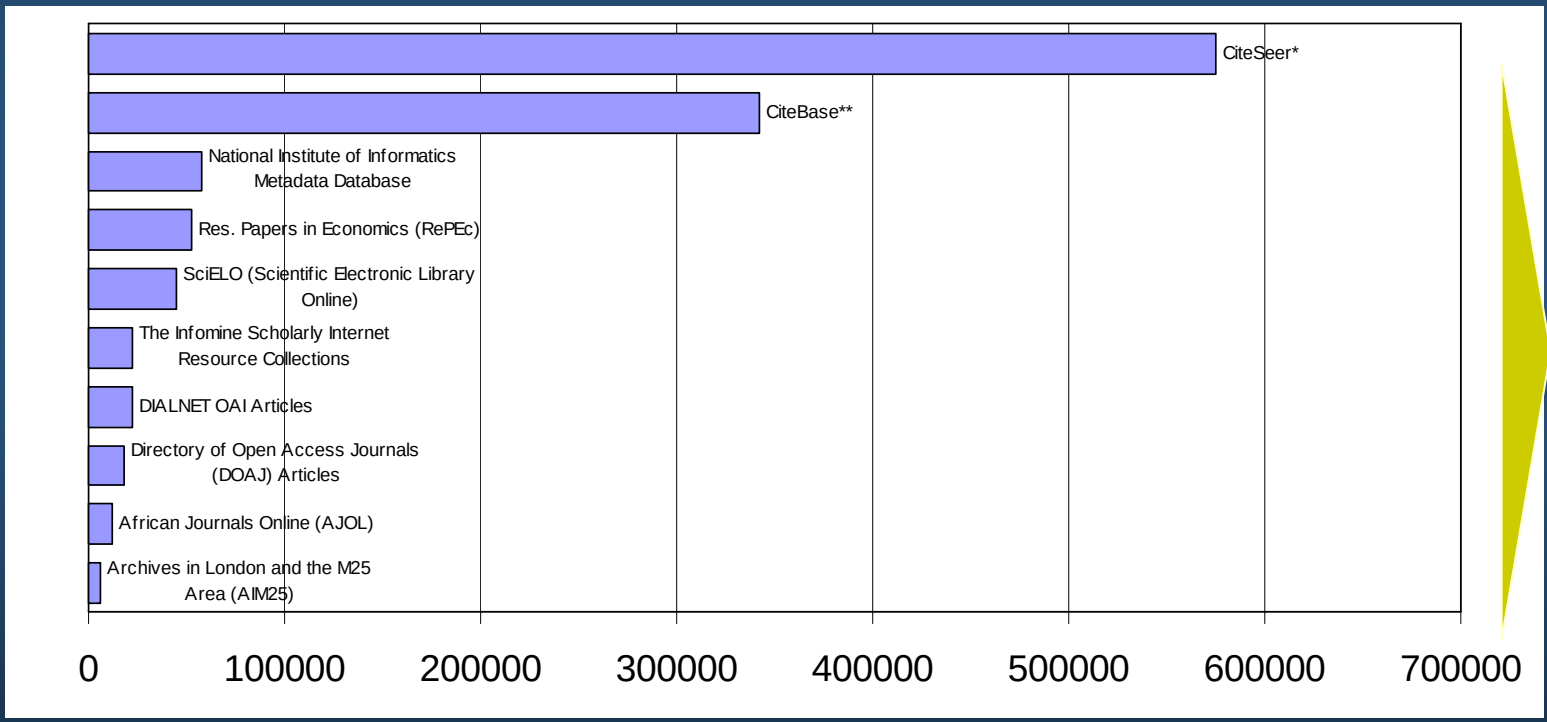
- Content hosted in 'general' IRs is spread across a number of institutions
- Several 'large' repositories (e.g. Cambridge) actually host small amount of potential journal content

* 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

Meta-data institutional repositories

Top ten aggregators (accounts for 99% of records of aggregators registered with OAIster)
Archived records* accessible through OAIster (excludes OAIster itself)



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

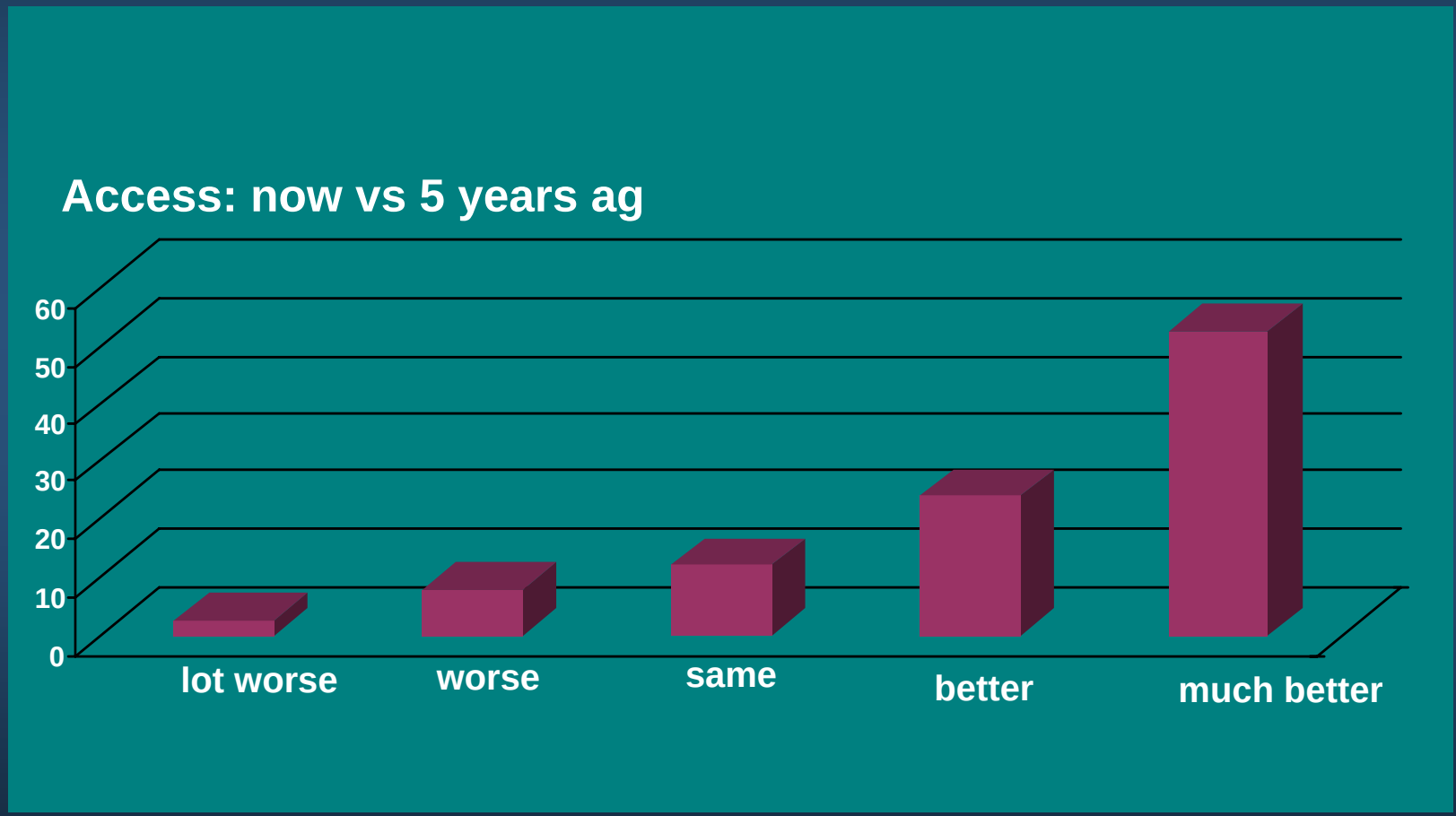
New in OA

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.

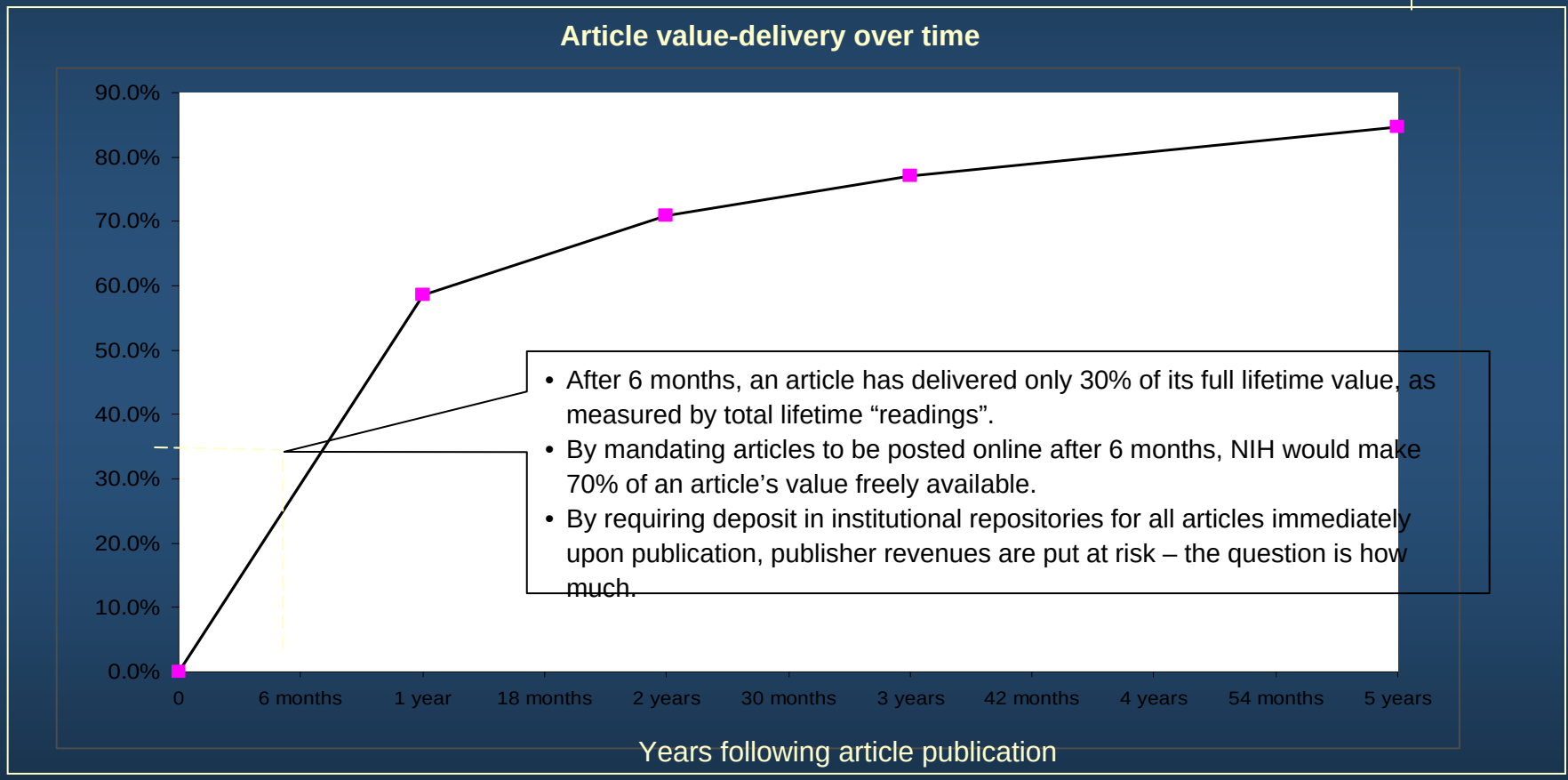
Source: EPrints OA News Blog: 03/03/05

Authors as readers: views on journal access (Ciber Study 2004)

of authors expressing an opinion, *n*=3,754



After Six Months, An Article has Delivered Only 30% of its Lifetime Value

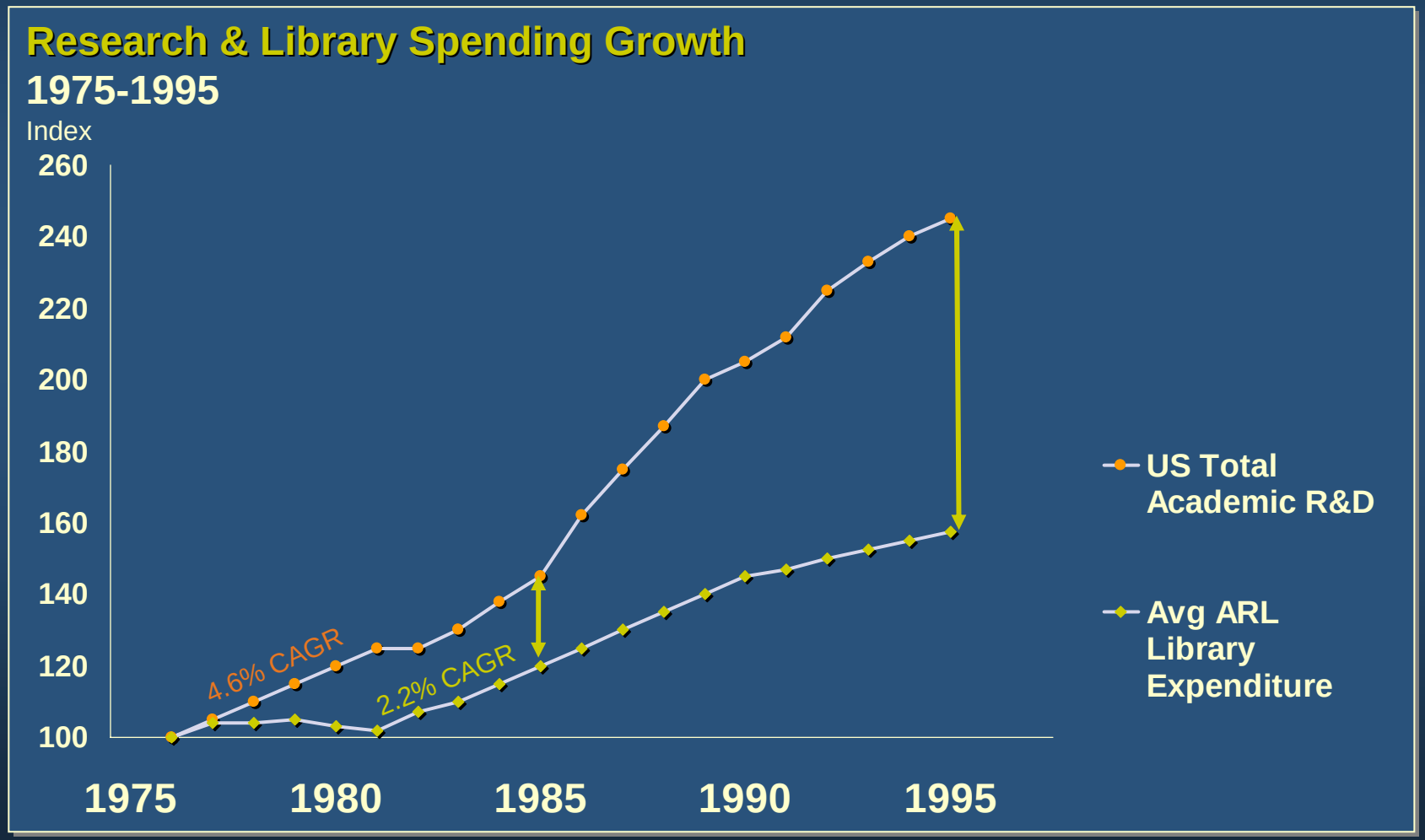


Source data:
Tenopir & King, "Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers", Special Libraries Assn, p 189, 2000.
Heer, "Article Aging".

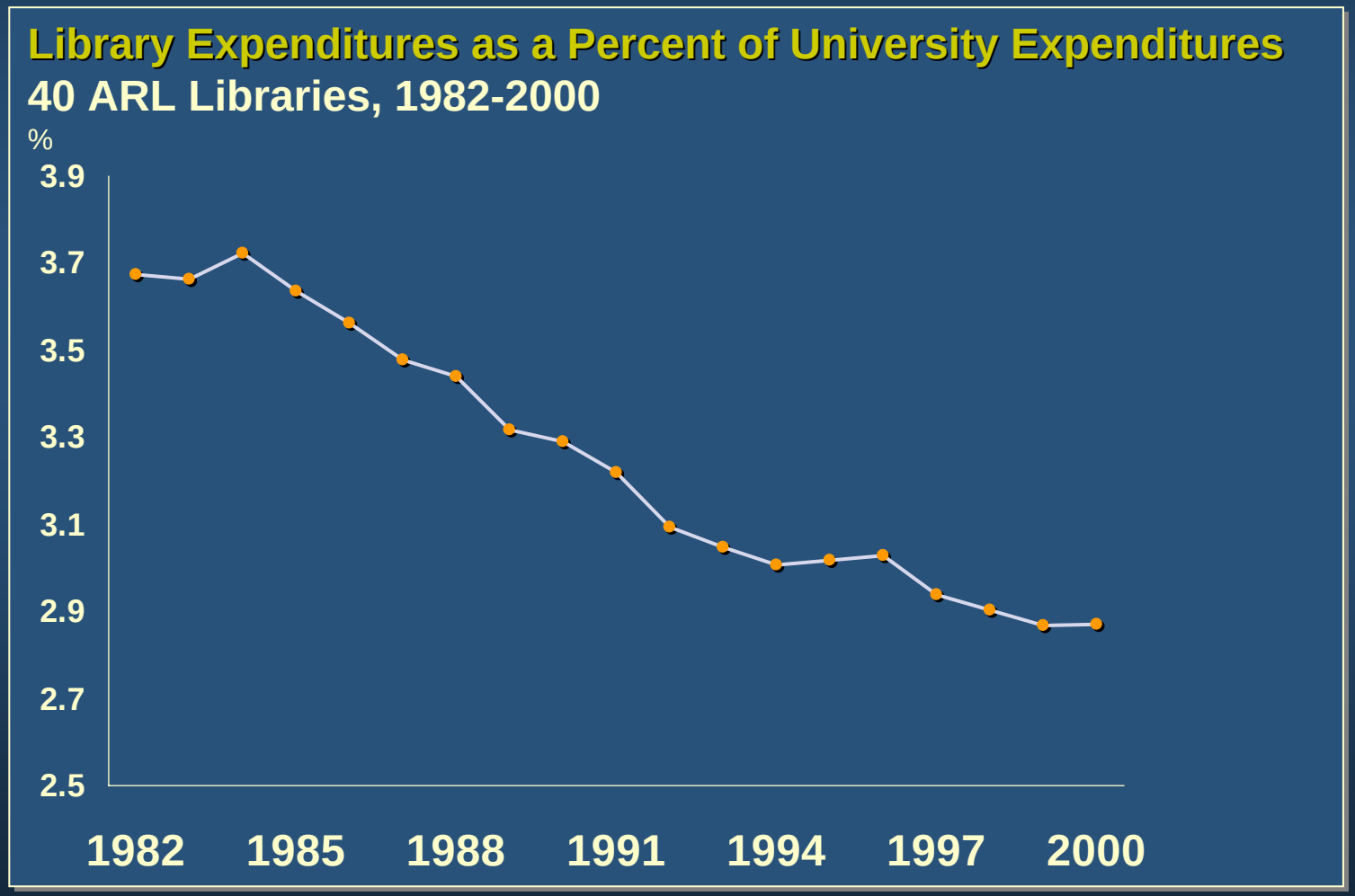
Some Preliminary Indications . . .

- 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 t 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

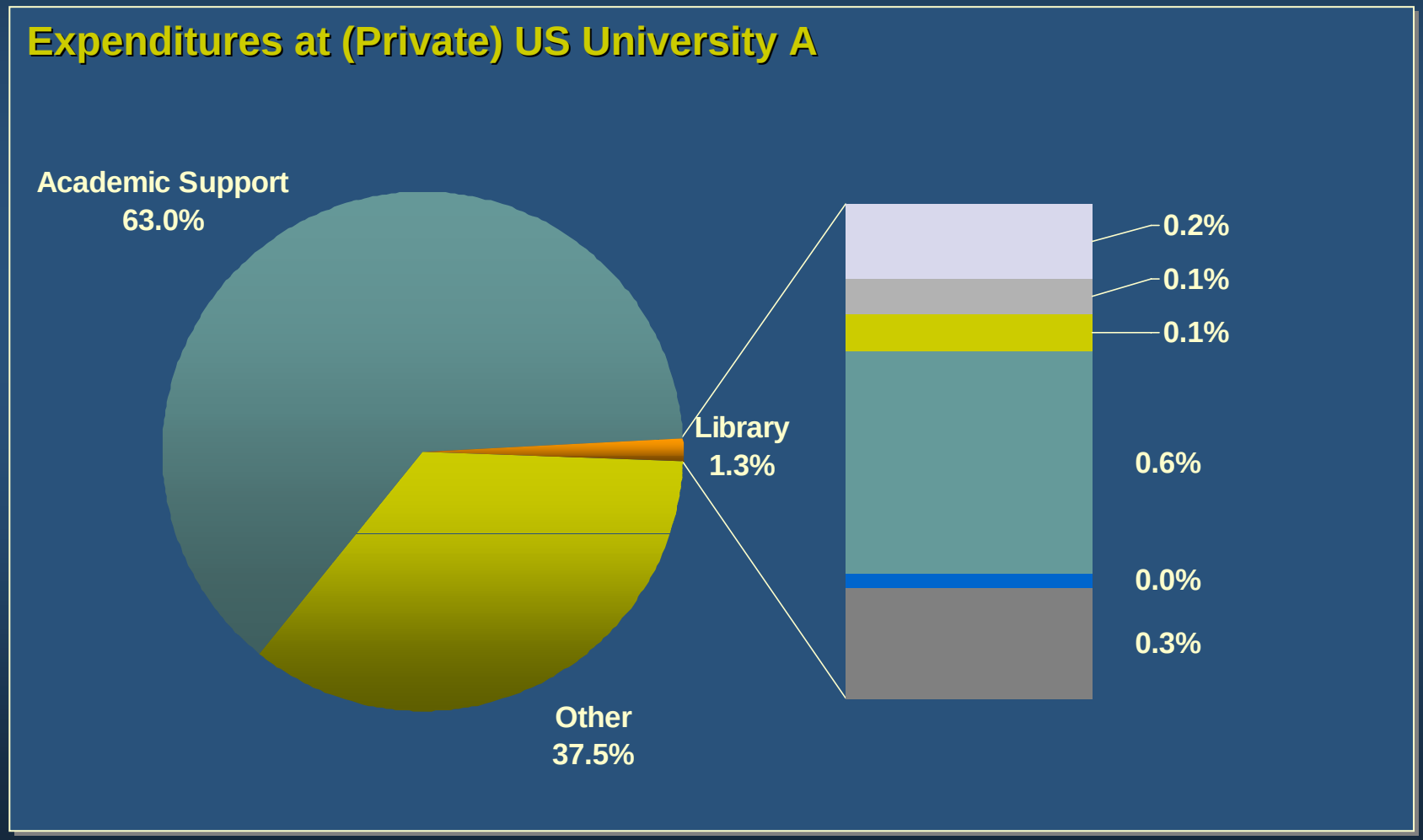
University Infrastructure Spending



University Infrastructure Spending



University Infrastructure Spending



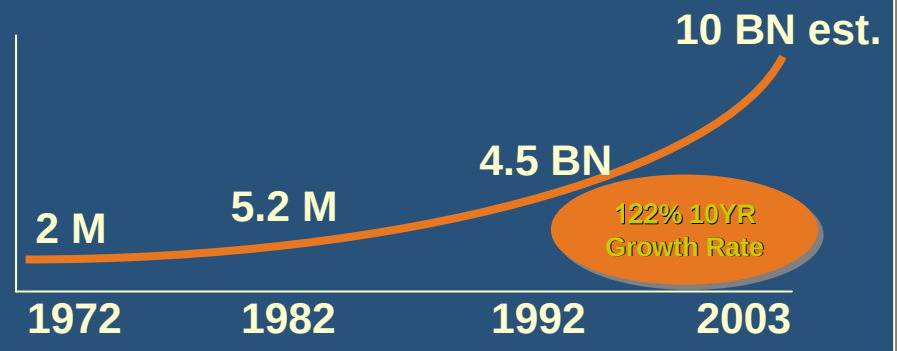
A&I Production & Coverage

STM Database Expansion – Duplication of Effort

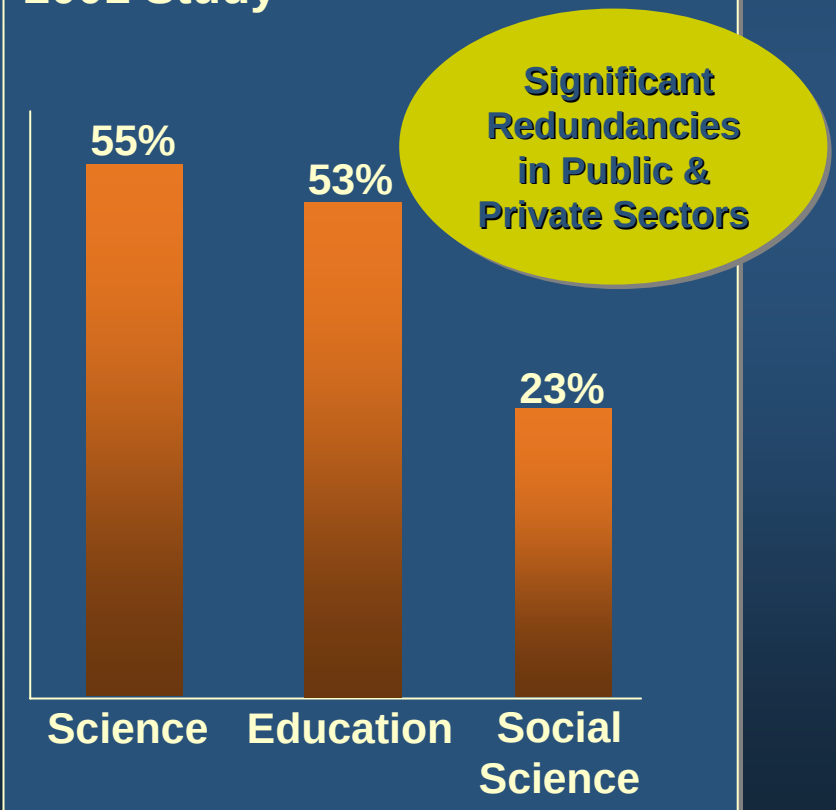
Shifting Sands

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

Records in Databases



Current Indexes Overlap
2001 Study

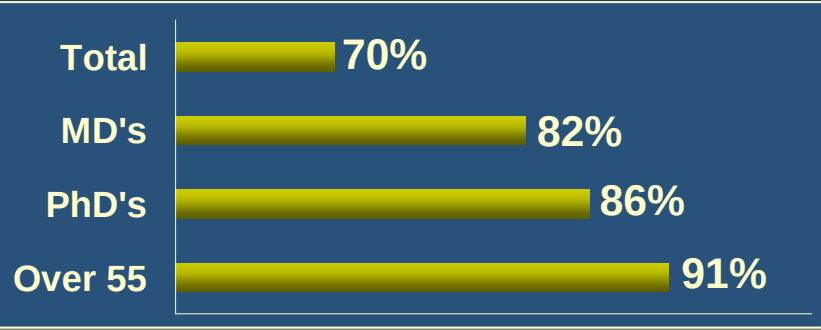


Scientists' and Researchers' Search Patterns

Online Experience

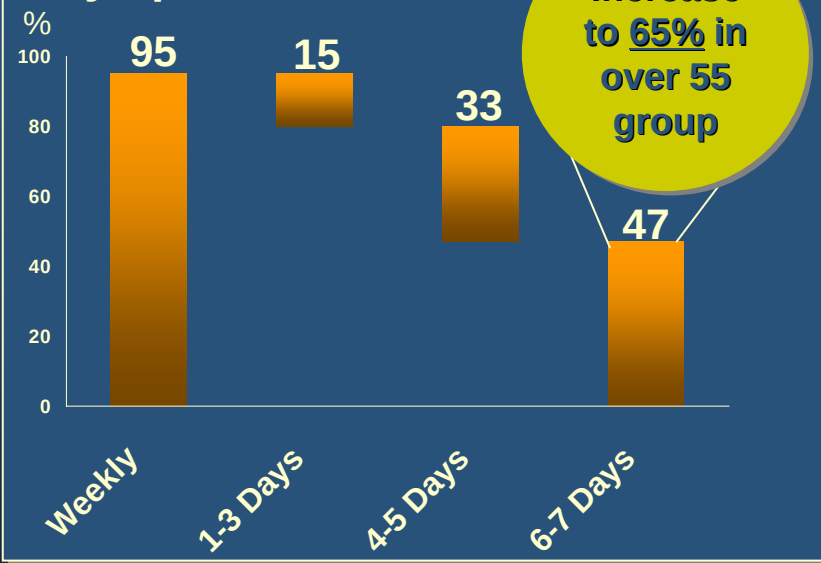
≥ 3 years using internet

%



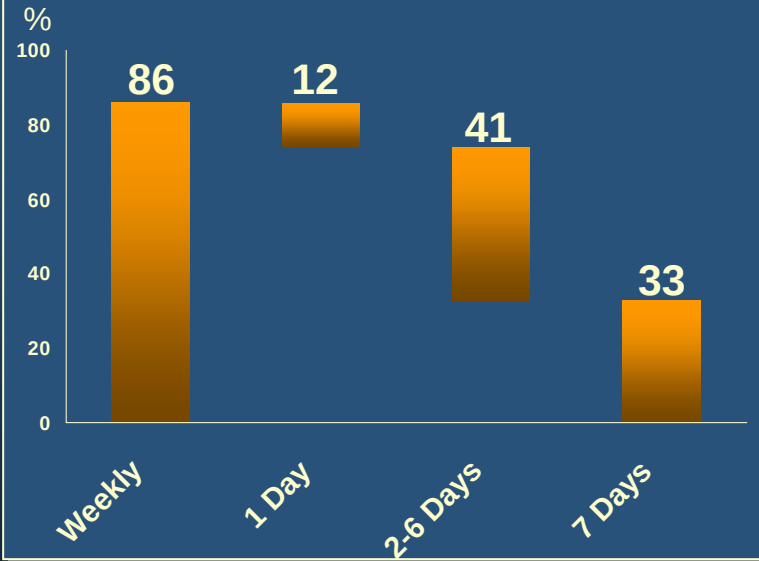
Work-related Internet Use

Days per Week



Work-related Search Engine Use

Days per Week



Scientists' and Researchers' Search Patterns

Early shift in Access Models to desktop

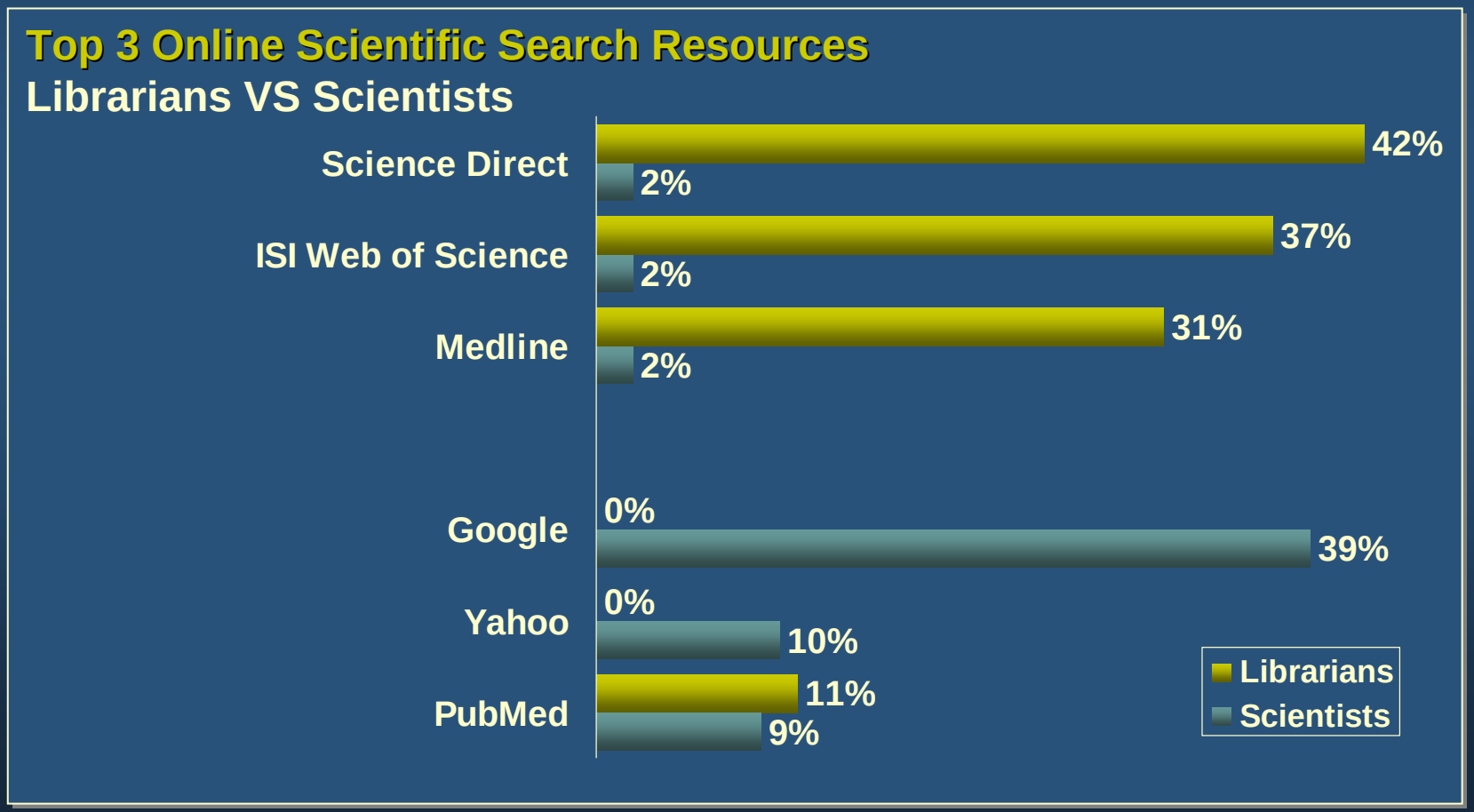
Shifting Sands

Experimentation by primary publishers with free access to metadata

- IEEE+
- Google
- Cross Search

Shifts in Mindshare

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

Pharmaceutical and Biotech Needs

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

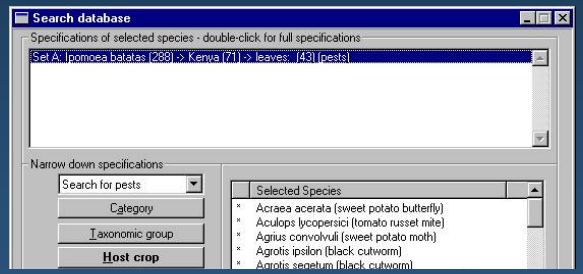
Health Care Needs




Patient Information and Treatment

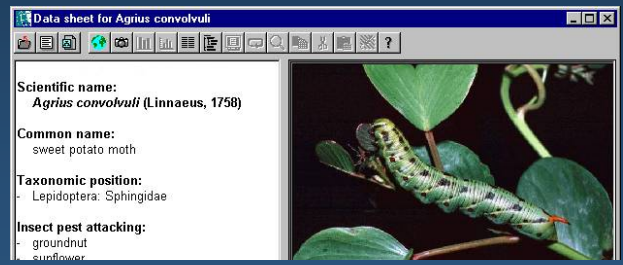
History	Physical Exam Lab	Diagnosis Evaluation Prevention Risk Reduction	Orders • Drugs • Procedures • Tests Referrals Patient Education

Agricultural Engineering Needs



Search the database to produce a list of possible pests.

CONTEXT

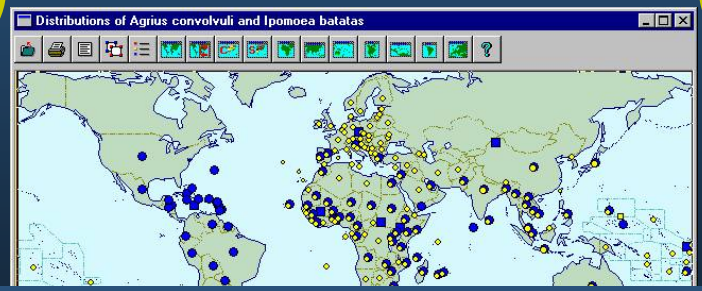


Datasheet provides a route to further text, pictures and maps.

CABI COMPENDIUM

TREATMENT

ASSESSMENT



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.