

Creating useful digital repositories

a vision based on scientific tradition and needs

Towards a European e-Infrastructure for
e-Science Digital Repositories

Final Workshop, 4th September 2007, Lisbon, Portugal
Closing speech, Jens Vigen, CERN

“Standing on ye shoulders of Giants”

e-Science requires digital shoulders

Scientific progress

- New findings were always built on previous results
- Adequate access to information is as needed in e-Science as it is in science, but quicker, deeper, and more accurate

Open Access revolutionized the access to information

- e-Prints are the main vehicle of OA information exchange
- Disciplines start to move towards Open Access publishing

A long-awaited watershed

- More than 15 years after the invention of the Web scientific information remains an electronic clone of the paper era
- The EC can play a pivot role in preparing the route towards e-Science through funding of development of infrastructures

Scientific information provision in the era of e-Science

About to be achieved

- Full text and data-mining applications
- Detection of *relations* between articles
- Treatment of large datasets for statistical and citation analyses

In progress

- Identification of popular and influential articles and authors with complementary ranking criteria; alternative metrics to ISI
- Access to numerical information from figures and tables within published articles

Still a long way to go for most disciplines

- Offer integrated access to primary scientific data

The “Digital” shoulders of giants

High-Energy Physics as an example

- Infrastructure for repository of scientific information
 - There is urgent need for an integrated repository for the High-Energy Physics (HEP) community employing state-of-the-art technology for storage, retrieval and information analysis
- Entire corpus of the HEP information in one place
 - *E.g.* the CERN Document Server hosts today 915 000 entries; half of which are catalogue records (just metadata) and the other half are objects freely available for download: full text articles but also slides, videos, photos, etc.
- Current priority
 - Merge data and services with HEP-SPIRES, ensure interoperability with related services, *e.g.* arXiv and NASA ADS. Empower the repository with new technology and content; enabling researchers to explore information matching the emerging expectations of the e-Science era.

Information discovery (I)

The tradition - keeping to the example of physics

Abstract journals

- Science Abstracts, 1898-1902
 - Multidisciplinary coverage. Joint collaboration between the Institution of Electrical Engineers and The Physical Society of London who set up a publishing committee to organise and administer the abstracting service
- Physics Abstracts, 1903-1966
 - The publication was split into the parts A and B, this allowed the subject scope, particularly in physics, to widen. From 1967 the service was computerized, under the name INSPEC, and does since then re-embrace the current six publications of Science Abstracts

Information discovery (II)

The tradition - keeping to the example of physics

The first repositories

- CERN and SLAC Library started cataloguing ~1960
 - Mme. Luisella Goldschmidt-Clermont, probably the first ever preprint librarian
 - Sustained Open Access collections, including references to the corresponding published literature, running since 50 years!
 - arXiv.org acting as the main submission interface since 1991

Why did these services develop?

- New needs surfaced within the community as the communication pattern was changing
 - Include the grey literature, index all authors (highly important for scientists working large collaborations) with their corresponding affiliation
 - ... and free access is an issue

Community building - the key to success

Defining and developing the repositories

- Survey user perception of present information systems
- Assess user requirements and preferences
- Learn nitty-gritty details for short-term (easy and feasible) improvements of current systems
- Look for the killer application(s) of the next years

Who should run the business of subject repositories?

- Authors, libraries, learned societies, publishers or any company simply motivated by the information business?
- Each community will have to find their own model
- ..., but do not forget, a centralized service can be easily monopolized
- mechanisms must be in place to avoid misuse of market power

Transforming our library web sites

Lara Cohen, Dec.15 2006

- I wish I could show you examples of exemplary academic library Web sites, but I can't. There aren't any. Yet.
- Get ourselves moving in order to stay relevant with today's users
- The service is about our users, not about us
- Library web masters will be replaced by blogs, wikis and RSS

Library 2.0

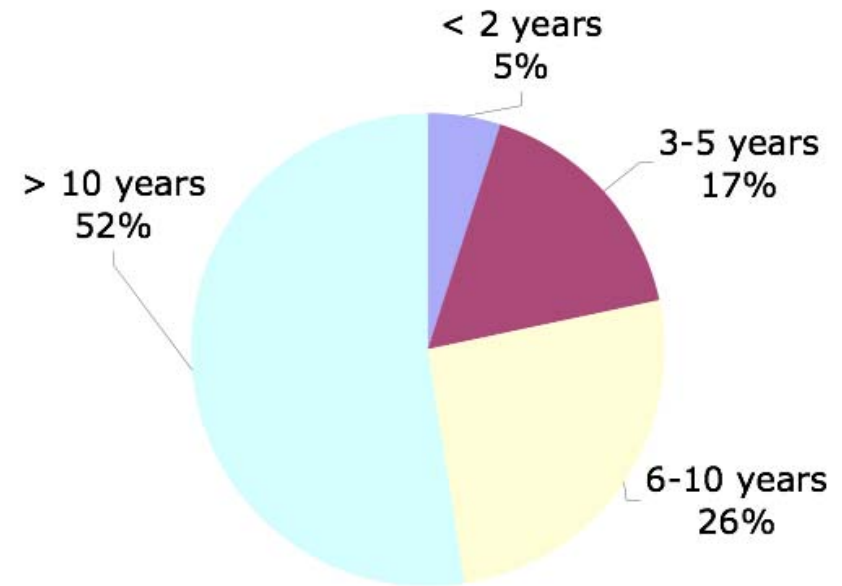
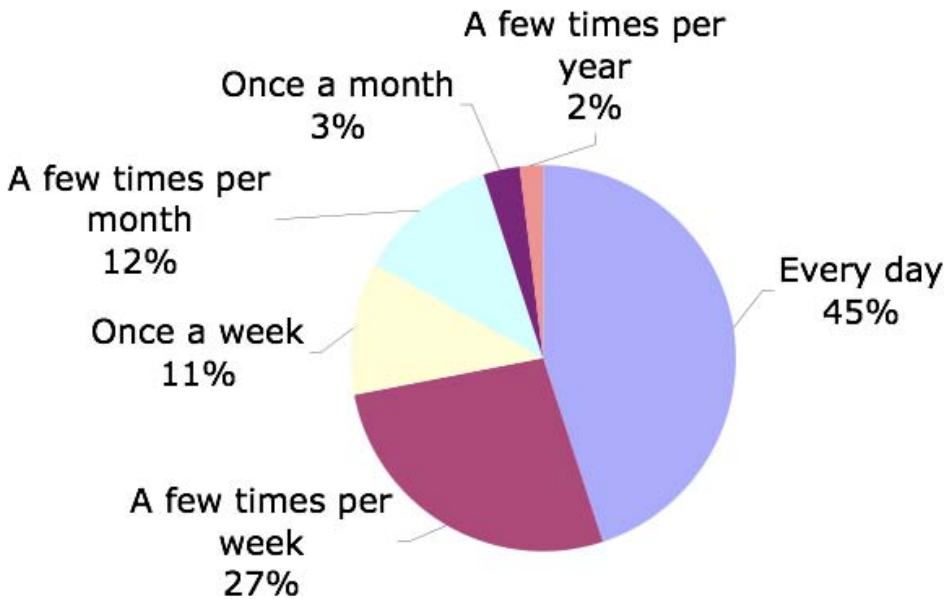
An academic's perspective →



The use of repositories ... given that they are discovered

Frequency of use of HEP Information Services

Experience with HEP Information Services



- In HEP usage of repositories is not an issue
- The survey collected more than 2000 replies
- Users are highly concerned
 - 43% wish to receive the results via e-mail
 - 89% answered to the “free-text” questions

Which system do you use the most?

3% Commercial services

- ~ 0% pay databases
- 3% publisher portals



11% Internet search engines

- 11% Google

86 % Community services

- 28 % Subject repositories
- 58 % Specialized libraries



What's on a user's mind today?

Showing top 40 of 3960 possible tags

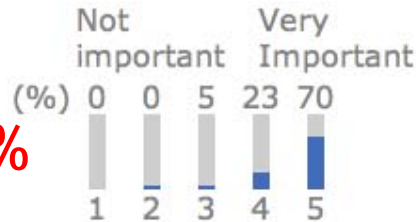
access ads analysis **articles** arxiv author
available best better cds **citation** complete contains convenient
coverage daily database **easy** engine fast field free **full-text**
google hep information interface journals links
physics **preprints** published references results
search spires subject system user work

TagCrowd
ALPHA

How important are these features of an information system?

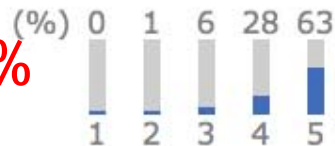
Depth of coverage

93%

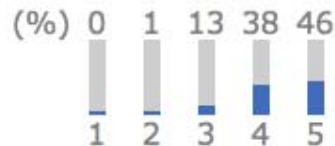


Quality of content

91%



User friendliness



Access to full text

94%



Speed to find what you want

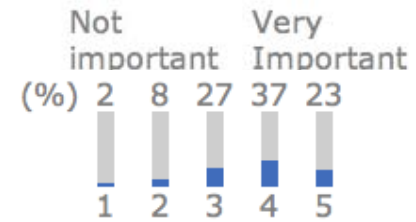


Search accuracy

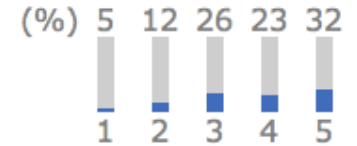
93%



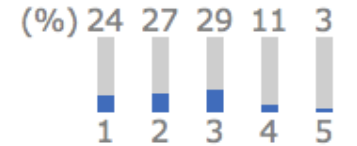
Submission interface



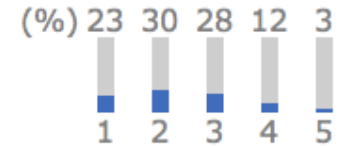
Citation analysis



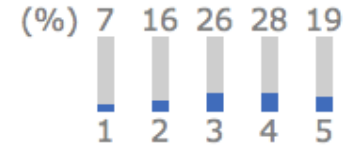
Multimedia content



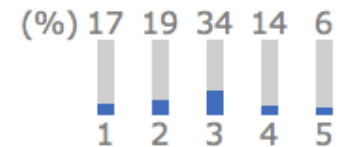
Personalisation



Keywords and classification



Collaborative tools



Which changes do you expect?

Summary of recurrent and inspiring answers

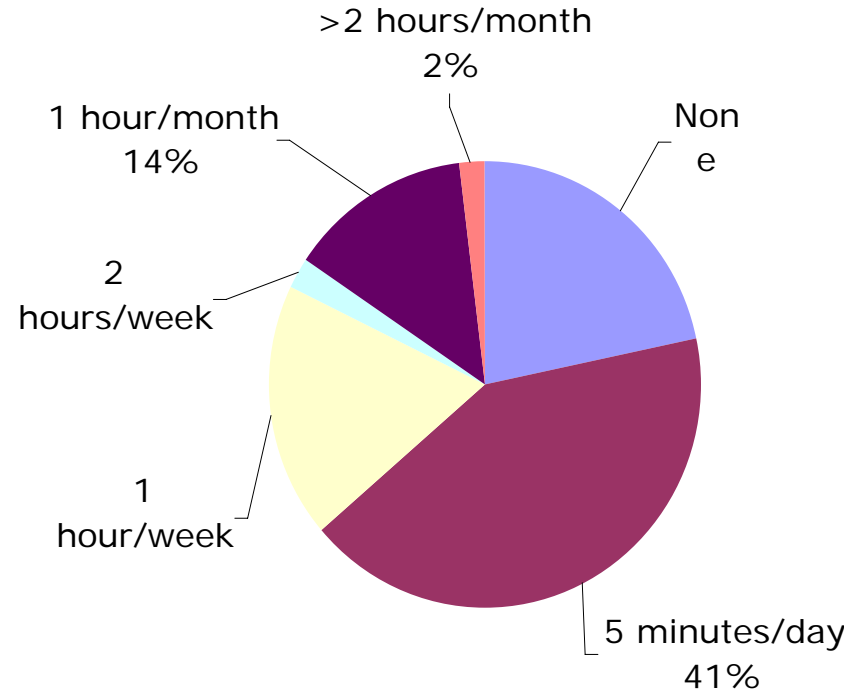
- Seamless (open) access to older articles via a unique portal
- Improved full-text search
- Indexing of conference .ppt slides (interlinked with the corresponding article)
- “Publication” of “ancillary” material:
 - Data in tables, figures
 - Correlation matrices
 - Data (high-level objects)
- (A new kind of) Peer-reviewing overlaid on subject repositories
- “Smarter” search tools

Any features you would dream of?

- **Contents of all famous journals**
- **Flow diagrams showing how certain articles initiated further research**
- **A more clever system of searching for a paper that is "connected" by title, citations, references, ... to a given paper**
- **Follow the 'paper trail'. Citations/References linked to be a single click away inside of the article and open access from anywhere and include peer-reviewed journals...**
- **A weekly alert of all the new preprints and publications**
- **Access to code fragments**
- **An idea/topic centric information system**

Web2.0 applications

If a simple web interface would show you an article and offer a set of categories to which it could belong, how much time would you spend in this tagging system to give a service to the community?



Vision

Build complete information platforms by disciplines

- In collaboration with all stakeholders, integrating the content of present repositories and databases to host the entire body of metadata and the full-text of all OA publications, past and future;

Enabling new full-text and data-mining applications on all publications

- Detecting **relations** between documents carrying similar information;
- Creating datasets to exercise new hybrid metrics to measure the impact of articles and authors and evaluating the scientific production of research groups;
- Extracting numerical information from figures and tables within published articles

Demonstrate and deploy Web2.0 applications in the domain of sciences

- Involving readers/authors in subject tagging, altering automatically assigned keywords/classification codes;
- Enabling the possibility to review and comment on articles, adding links to additional documents or other digital objects;
- Providing collaborative tools for effective management of co-authorship within distributed collaborations;
- Introducing community-based alternatives to the established peer-review system

Bridging publications with data

The natural evolution of repositories

- Easy-to-add material comes first
- Secondly the repositories will be populated with what is essential to attract users

Academic incentives

- Preparing and publishing data for long-term preservation has to be rewarded to encourage deposit

It is our duty to ensure long term access to data

- Expensive measurements might need to be revisited
- Without access to old climate data we would not know today that the globe is warming ...

Conclusions

- The era of e-Science is still ahead of us
- e-Science requires Open Access to be efficient
- The FP7 programs will open fantastic opportunities for further R&D in the field of building repositories. Investing in a few grand-challenge projects will bring us quicker down the road, aiming for something operational by 2013 that eventually can be adopted and deployed by other players/disciplines
- Consolidation - new services should be built block by block
- Funding on a rolling or long-term basis is essential
- Information specialists have the opportunity to play a key role in the era of e-Science ... provided we listen to our users and offer the tools they need!

I wish you all the best in compiling the report - hopefully it will result in interesting calls followed by a range of exciting bids