Open Access for authors, researchers and their institutions

Presented by Steve Hitchcock, School of Electronics and Computer Science (ECS), Southampton University

These slides prepared for **CRIS2004**, 7th international conference on Current Research Information Systems
http://www.eurocris.org/conferences/cris2004/index.html
on May 13-15, 2004, Antwerp

Abstract

Open access - immediate and permanent free access - will transform the use and impact of published research results. Much of the emphasis has been on the clear benefits of open access for readers and users of this information. First, authors have to be persuaded to adapt and make their papers openly accessible. The benefits for authors and their institutions are significant, but have been given less attention. The talk will highlight examples that reveal motivations and incentives for authors to use open access: increased impact, the ability to link full experimental data with abbreviated published descriptions. Institutions share these interests with their authors, and are setting up institutional eprint archives to provide open access to these materials. If they adopt policies requiring authors to self-archive, institutions can then use these comprehensive archives to produce publication lists and other inputs for research assessment exercises on behalf of research funders. The School of Electronics and Computer Science at Southampton University has done this, and the talk will consider how and why this has worked successfully.

Information systems, CRIS?

Types of information to be considered in this talk:

- **Research data**
- **eprints** (author-self archived versions of published papers, to provide open access)

The **eBank project** (http://www.ukoln.ac.uk/projects/ebank-uk/) is exploring how both types of data can be stored, linked and accessed using services based on **GNU Eprints** (http://software.eprints.org/), free open source software for building institutional eprint archives that are compliant with the Open Archives Initiative (OAI)

- **Citation indexing**

Citebase (http://citebase.eprints.org/) measures the impact of papers in selected archives that use the OAI, e.g. physics arXiv. Data from Citebase are used with archive usage data (Web 'hits') in the **Correlation Generator** (http://citebase.eprints.org/java/correlation/correlation.html) to predict *future* citation impact

Open access and eprints: what researchers want

To maximise research **progress** and their ***rewards***
by maximizing (and accelerating) research ***impact***

Impact has typically been based on citation measures of journals. Now we can measure the impact of individual Web papers and of their authors.

It has been shown that articles freely available online (open access) are more highly cited, i.e. **open access increases impact.**

The easiest and fastest way for authors to make papers freely available, and thereby maximise their impact, is by self-archiving them in **institutional eprint archives**.

Free online availability increases impact

Lawrence, S. (2001) *Nature*: "average of 336% more citations to online articles compared to offline articles published in the same venue" http://www.neci.nec.com/~lawrence/papers/online-nature01/

Kurtz, M. J. (2004) Restrictive access policies cut readership of electronic research journal articles by a factor of two http://opcit.eprints.org/feb19oa/kurtz.pdf

Greg Schwarz (forthcoming): ApJ papers that were also on astro-ph (part of arXiv) have a citation rate that is *twice* that of papers not on the preprint server http://listserv.nd.edu/cgi-bin/wa?A2=ind0311&L=pamnet&D=1&O=D&P=1632

Brody, T., *et al*. (2004) The Effect of Open Access on Citation Impact http://opcit.eprints.org/feb19oa/brody-impact.pdf

National and international policies supporting open access

- Budapest Open Access Initiative (BOAI), 2002
- US Sabo Bill ("Public Access to Science"), 2003
- Berlin Declaration, 2003
- OECD Declaration on Access to Research Data from Public Funding, 2003
- The Wellcome Trust Statement, 2003

See **National Policies on Open Access (OA) Provision for University Research Output: an International meeting**

http://opcit.eprints.org/feb19prog.html

BOAI dual open-access strategy

Gold: Publish your articles in an open-access journal whenever a suitable one exists today (currently <1000, <5%)

and

Green: Publish the rest of your articles in the toll-access journal of your choice (currently 23,000, >95%) **and** self-archive them in your institutional open-access eprint archives.

There is NO immediate alternative to a dual strategy. The Gold strategy, if pursued alone, will not result in universal open access any time soon

Notes. Colours refer to the rights classification of journals adopted by the Romeo project; updated data on publisher copyright policies
http://www.ecs.soton.ac.uk/~harnad/Temp/Romeo/romeo.html

See *OSI EPrints Handbook*: 2. A Guide to Self-Archiving and Open Access
http://software.eprints.org/handbook/

Which archive software? Eprints

There are various working packages, see *OSI Guide to Institutional Repository Software* (2nd edition) http://www.soros.org/openaccess/software/OSI_Guide_to_Institutional_Repository_Software_v2.htm

"The Eprints software has the largest -- and most broadly distributed -- installed base of any of the repository software systems described here"

The primary target of GNU EPrints software are the estimated 2.5M papers published annually in the 24k peer reviewed journals and now it is being adapted for scientific data reports as part of the eBank project

Structure of the talk

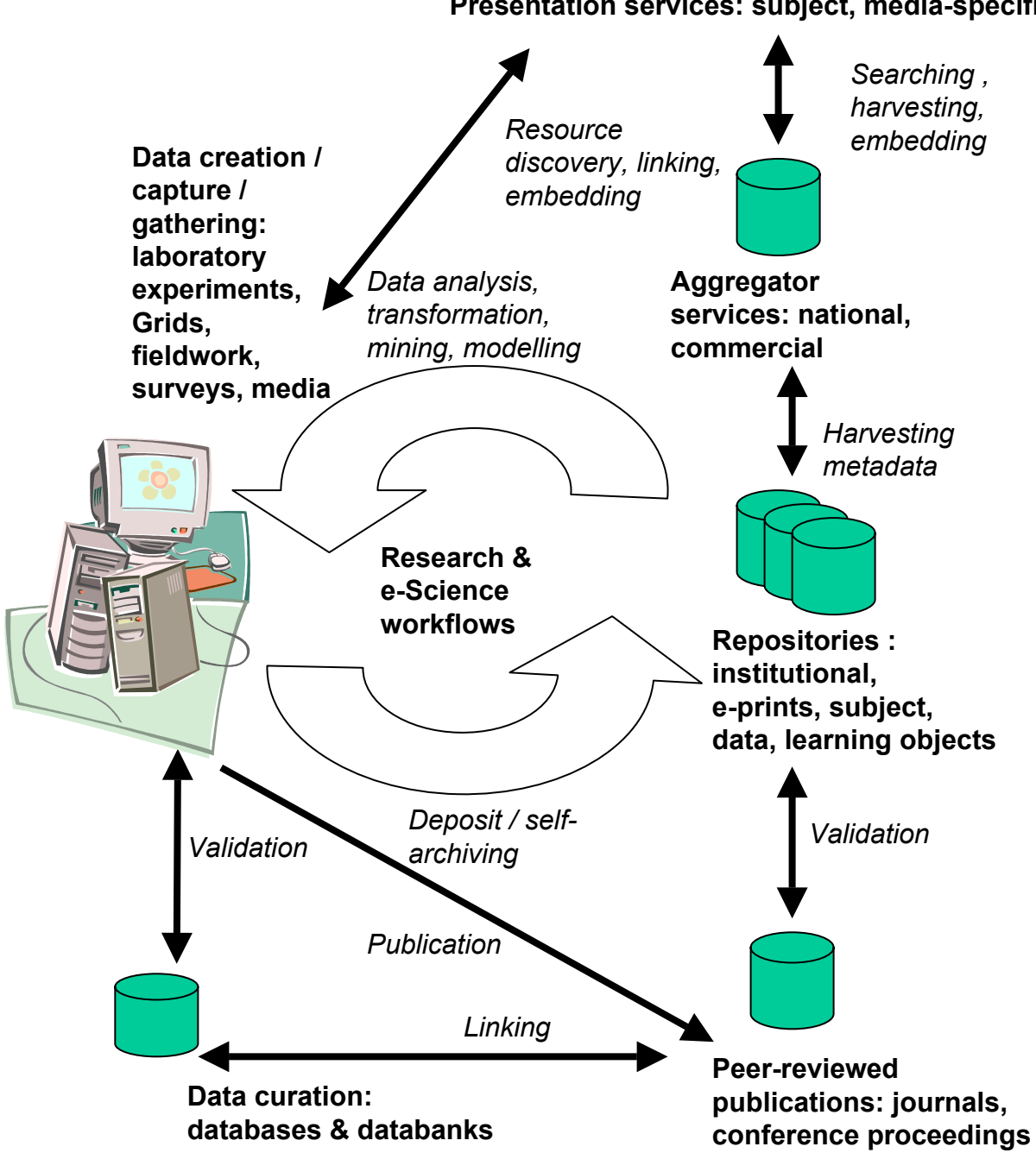
- **eBank project**: capturing research data in Eprints software, metadata schemas, harvesting
- **ECS Eprints (Southampton)**: filling an institutional (school) Eprints archive with self-archived papers, an institutional policy, a research assessment exercise (dry run)
- **Citebase and the Correlation Generator**: measuring research/citation impact on the Web, and predicting future impact

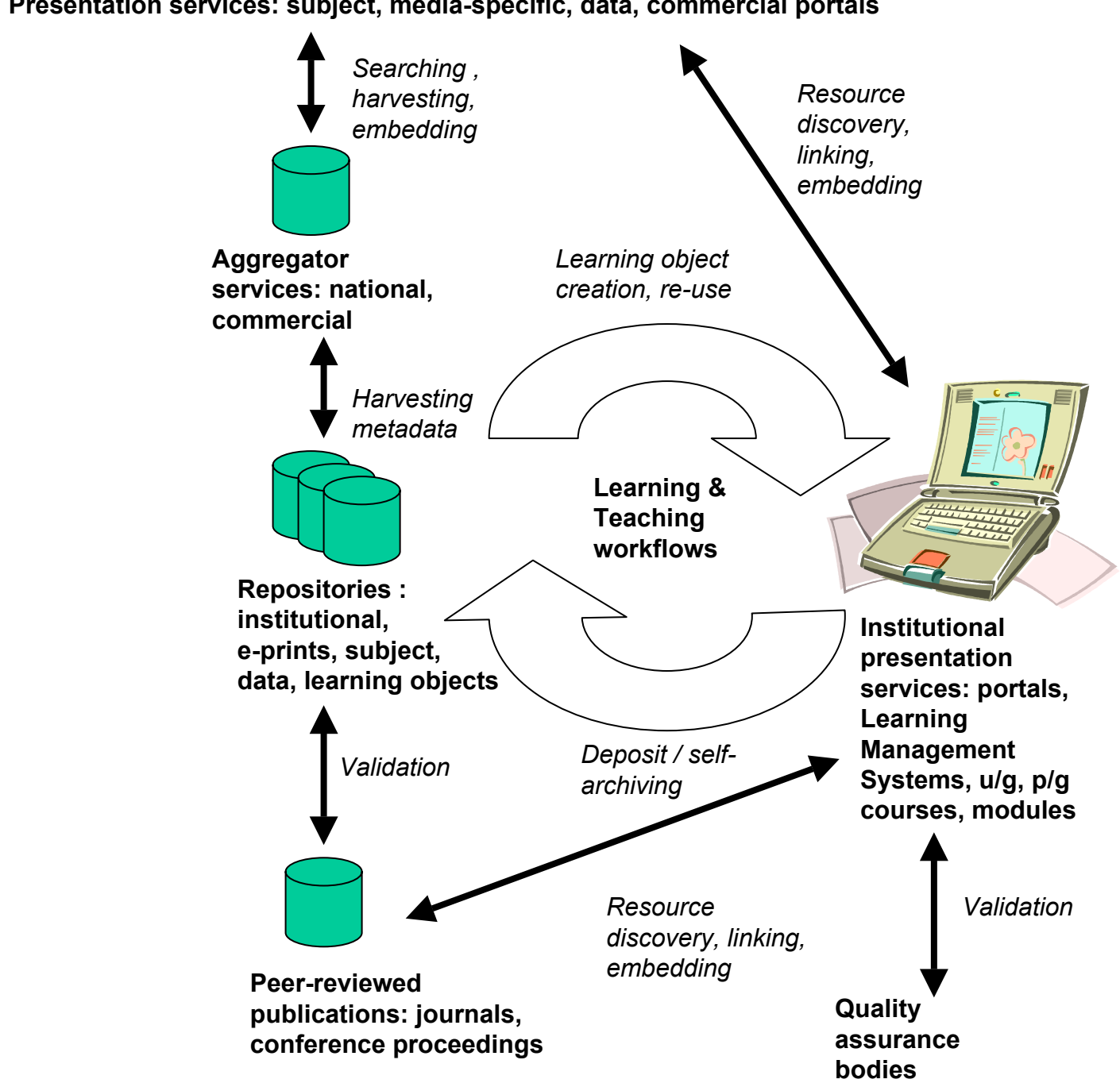
eBank project

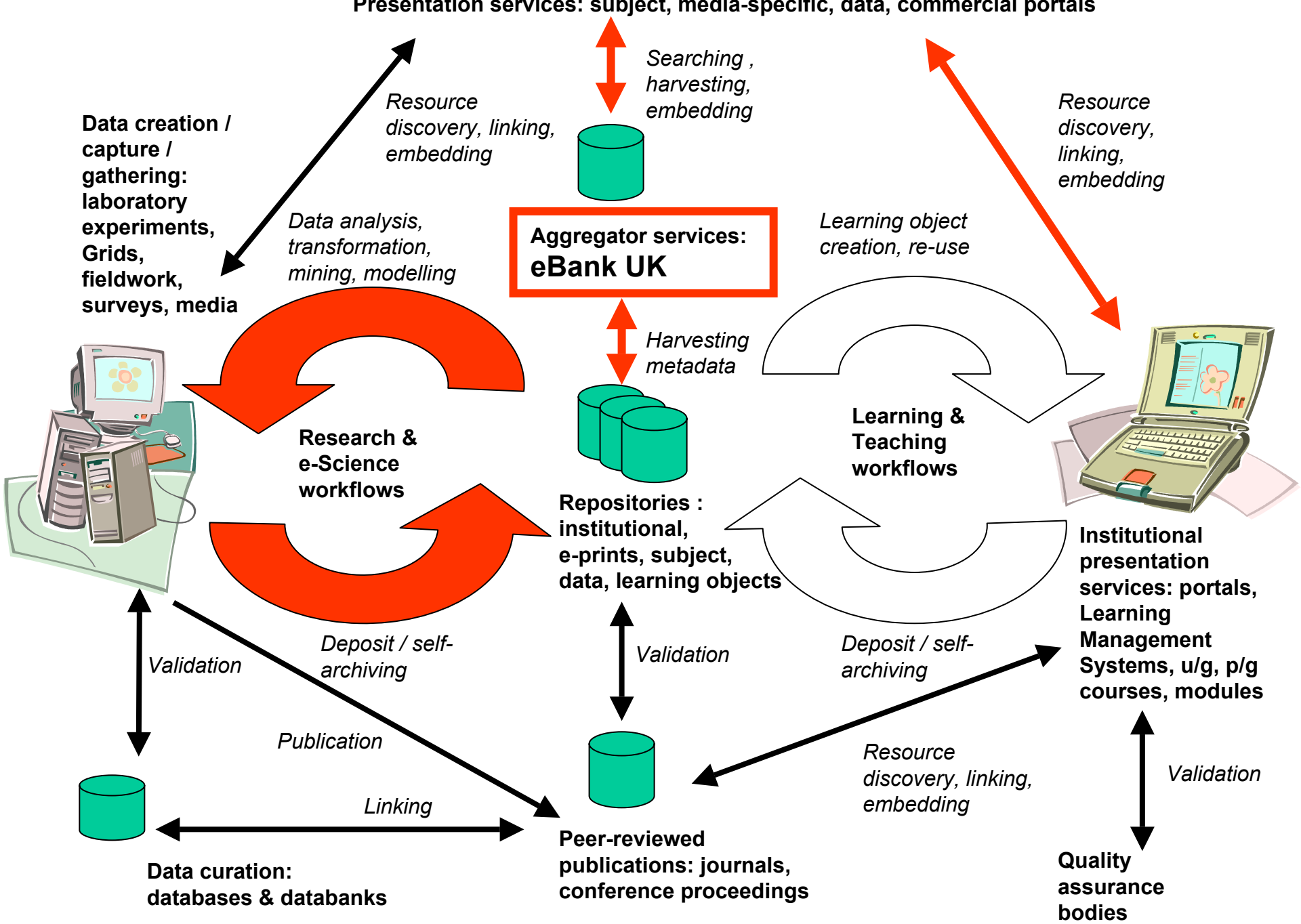
- **JISC-funded** for 1 year from September 2003
- UKOLN, University of Southampton, University of Manchester
- "Building the links between research data, scholarly communication and learning"
- **e-Science testbed Combechem**
  - Grid-enabled combinatorial chemistry
  - Crystallography, laser and surface chemistry
  - Development of an *e*-Lab using pervasive computing technology
  - National Crystallography Service
- Resource Discovery Network PSIgate physical sciences portal

eBank in the scholarly knowledge cycle

The following three slides with schematics are from
Liz Lyon, Realising the scholarly knowledge cycle: The experience of eBank UK**,** *CNI Task Force Meeting Spring 2004*
http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/lyon-cni-spring04-final.ppt


Presentation services: subject, media-specific, data, commercial portals
Searching , harvesting, embedding
Resource discovery, linking, embedding
Data creation / capture / gathering: laboratory experiments, Grids, fieldwork, surveys, media
Data analysis, transformation, mining, modelling
Aggregator services: national, commercial
Harvesting metadata
Research & e-Science workflows
Repositories : institutional, e-prints, subject, data, learning objects
Validation
Deposit / self-archiving
Validation
Publication
Linking
Data curation: databases & databanks
Peer-reviewed publications: journals, conference proceedings


Presentation services: subject, media-specific, data, commercial portals
Searching , harvesting, embedding
Resource discovery, linking, embedding
Aggregator services: national, commercial
Learning object creation, re-use
Harvesting metadata
Learning & Teaching workflows
Repositories : institutional, e-prints, subject, data, learning objects
Institutional presentation services: portals, Learning Management Systems, u/g, p/g courses, modules
Validation
Deposit / self-archiving
Resource discovery, linking, embedding
Validation
Peer-reviewed publications: journals, conference proceedings
Quality assurance bodies


Presentation services: subject, media-specific, data, commercial portals
Searching , harvesting, embedding
Resource discovery, linking, embedding
Resource discovery, linking, embedding
Data creation / capture / gathering: laboratory experiments, Grids, fieldwork, surveys, media
Data analysis, transformation, mining, modelling
Aggregator services: eBank UK
Learning object creation, re-use
Harvesting metadata
Research & e-Science workflows
Learning & Teaching workflows
Repositories : institutional, e-prints, subject, data, learning objects
Institutional presentation services: portals, Learning Management Systems, u/g, p/g courses, modules
Validation
Deposit / self-archiving
Validation
Deposit / self-archiving
Publication
Linking
Resource discovery, linking, embedding
Validation
Data curation: databases & databanks
Peer-reviewed publications: journals, conference proceedings
Quality assurance bodies

Crystallography workflow

- **Initialisation**: mount new sample on diffractometer and set up data collection
- **Collection**: collect data
- **Processing**: process and correct images
- **Solution**: solve structures
- **Refinement**: refine structure
- **CIF**: produce Crystallographic Information File
- **Report**: generate Crystal Structure Report

eBank metadata schema (draft)

Name	Description of the stage	Files associated with this stage			Metadata associated with this s	
		File	Type	Description	Name	Data Typ
Initialisation	Mount new sample on diffractometer Parameterisation to set up data collection	datcol.non i*.kcd .drx	ASCII BINARY ASCII	Parameters for producing i*.kcd files Unit cell determination images Unit cell	Morphology Solvent Sample_image	*STRING *STRING .JPG
Collection	Collect Data	collect.rmat datcol.non .py s*.kcd *scan*.jpg	ASCII ASCII ASCII BINARY JPG	Orientation of the crystal Command file Proprietary configuration file Diffraction images Visual version of .kcd file	Instrument_Type Temperature Crystal_image Software_Name Software_Version	*STRING *INTEGE .JPG STRING INTEGER
Processing	Process and correct images	scale_all.in scale_all.out .hkl .htm	ASCII ASCII ASCII HTML	Result of processing Result of correction on processed data Derived data set Report file	Cell_a Cell_b Cell_c Cell_alpha Cell_beta Cell_gamma Crystal_system Completeness Software_Name Software_Version	*NUMBE *NUMBE *NUMBE *NUMBE *NUMBE *NUMBE *STRING *INTEGE STRING INTEGER
Solution	Solve Structures	.prp xs.lst	ASCII ASCII	Symmetry file, log of process Solution log file	Space_group Figure_of_merit Software_Name Software_Version	*STRING *NUMBE STRING INTEGER
Refinement	Refine Structure	xl.lst .res	ASCII ASCII	Final refinement listing Output coordinates	R1_obs wR2_obs R1_all wR2_all Software_Name Software_Version	*NUMBE *NUMBE *NUMBE *NUMBE STRING INTEGER
CIF	Produce CIF	.cif	ASCII	Final results	Formula_moiety CIF_check	*STRING *STRING
Report	Generate e-Print report	.html	.HTML	Publication format (HTML/XHTML)	EPrint_type Authors Affiliations Formula_empirical Compound_name CCDC_Code Compound_class Keywords Available_data	*CRYSTA STRUCTU *STRING *STRING *STRING *STRING *STRING *STRING *STRING *STRING

eBank schema for harvesting the e-data report

Data Name	Data Description	Data Type	XML wrapped content
EPrint_type	'Crystal Structure'	String	Phrase 'Crystal Structure'
Authors	ePrint creator(s)	String	ePrint authors 'Surname, Christian name, initial'
Affiliations	Institution(s) of creator(s)	String	Various authors addresses
Formula_empirical	Total atom count	String	Atom symbols with their total count (can be real number) subscript
Compound_name	IUPAC Chemical name	String	Chemical name with text & integers
CCDC_Code	Cambridge Structural Database identifier	String	6 character code (may become numeric in future)
Compound_class	Chemical category	String (set)	1 word descriptor of chemical category
Available_data	Actual data available for various ePrint stages (Y/N)	Y/N Toggle	Y or N presence of data associated with RAW & RESULTS stages
Related_publications	Other output containing this compound/structure	String	Literature reference link
Publication_date	Date of releasing ePrint to eBank/world	String	Date of public release of ePrint
Last_revised_date	Date ePrint last revised	String	Date of latest modification to ePrint
Keywords	Categories	String (set?)	Phrase describing chemical relevance
Scheme	2D diagram	String	Two dimensional structural diagram as SMILES string
IChI	International Chemical Identifier	String	Unique compound identifier (contains some structural information)

EBank Southampton Home || About || Browse || Search || Register || User Area || Help

Bis(mu2-4,6-bis((diphenylphosphino)oxy)-5-methyl-1,3-phenylene-C,C',P,P')-tetrakis(mu2-trifluoroacetato-O,O')-tetra-palladium chloroform solvate

cif

- 02SRC841.cif (20336)

rfne

- 02src841.res (8460)
- 02src841_xl.lst (58006)

soln

- 02src841.PRP (6019)
- 02src841_xs.lst (73362)

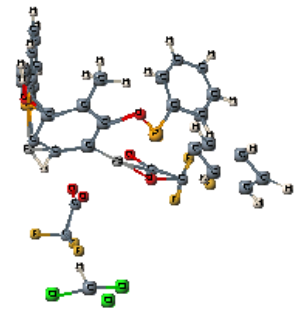
proc

- 02SRC841.HTM (6535)
- 02src841.HKL (1180322)

?

- 02src841_0KL.JPG

Simon J Coles, Robin B Bedford, M E Blake, Michael B Hursthouse and P N Scully.



Creation Date: 18 March 2004

Deposited By: Christopher Gutteridge

Deposited On: 18 March 2004

_CHEMICAL_FORMULA_SUM:	C288 H200 Cl24 F48 O48 P16 Pd16

Example Eprints/eBank crystal structure report from http://eprints.ebank.ecs.soton.ac.uk

The next slide is a placeholder for a poster that displays all the linked resources (data files). A full version can be found at

http://eprints.soton.ac.uk/archive/00001633/

Crystal Structure EPrints:
Publication @ Source Through the Open Archive Initiative

S.J. Coles[a*], J.G. Frey[a], M.B. Hursthouse[a], L. Carr[b] & C.J. Gutteridge[b].

[a]School of Chemistry, University of Southampton, UK.; [b]School of Electronics & Computer Science, University of Southampton, UK.

The Publication Problem

Recent advances in crystallographic instrumentation and computational resources have caused an explosion of crystallographic data, as shown by the exponential growth of the Crystallographic Structural Database over the last few years. The traditional peer review methods of dissemination of chemical data are unable to keep up with this new pace of data generation, causing a publication bottleneck. This problem will become even more severe with developments in high throughput chemistry (Combichem) and the impact of eScience (Combechem). As a result of this situation, the user community is deprived of valuable information, and the funding bodies are getting a poor return for their investments!

The Open Archive Initiative (OAI) approach of EPrints offers a solution to this problem through publically accessable archives They are currently a method for disseminating scholarly and research output that cannot enter the public domain through conventional routes.

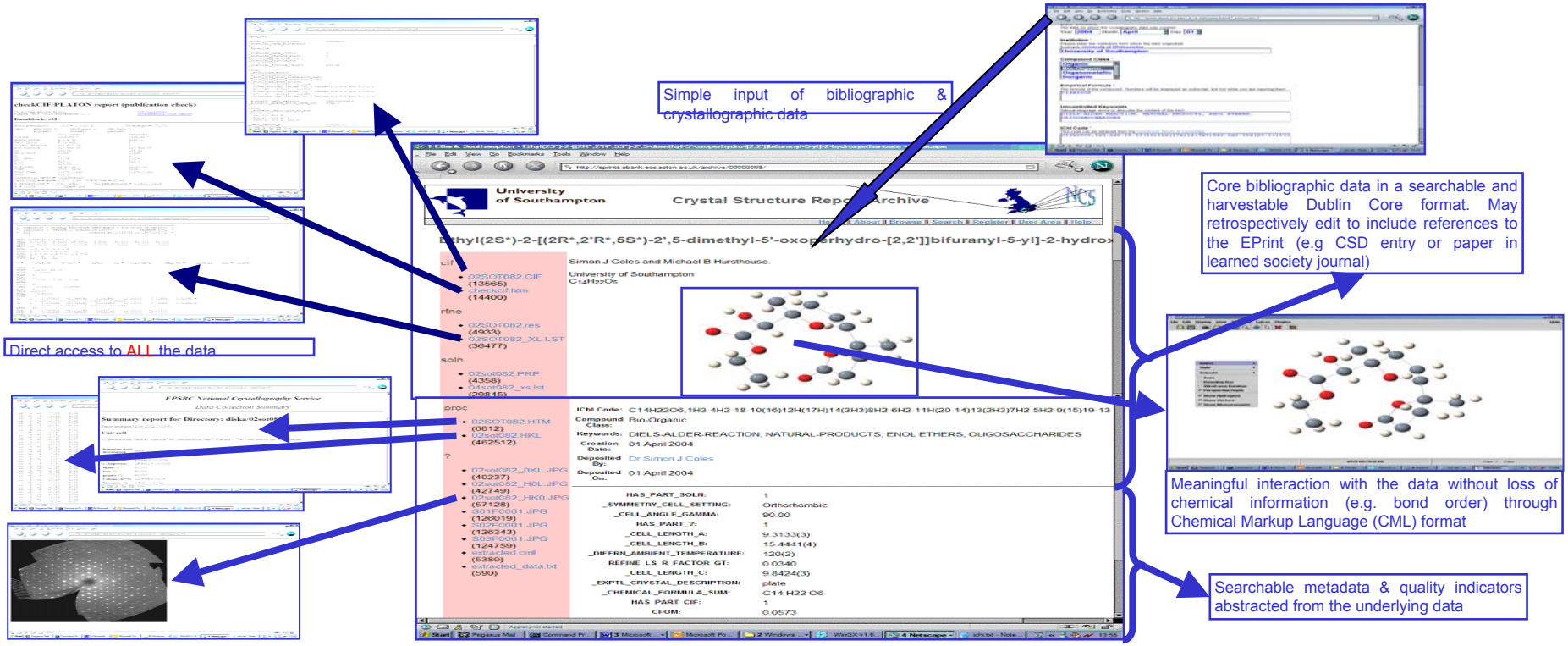
Data Publication @ Source

Crystallographic EPrints use the OAI concept to make available ALL the data generated during the course of a structure determination experiment.

That is: the publishable output is constructed from all the raw, results and derived data that is generated during the course of the experiment.

This presents the data in a searchable and hierarchical system. At the top searchable level this metadata includes bibliographic and chemical identifier items which allow access to a secondary level of searchable crystallographic items which are directly linked to the associated archived data.

Hence the results of a crystal structure determination may be disseminated in a manner that anyone wishing to utilise the information may access the entire archive of data related to it and assess its validity and worth. This way the world becomes the peer reviewers!



The Bigger Picture

All the 'core bibliographic data' is made available in a harvestable format (OAI-PMH).

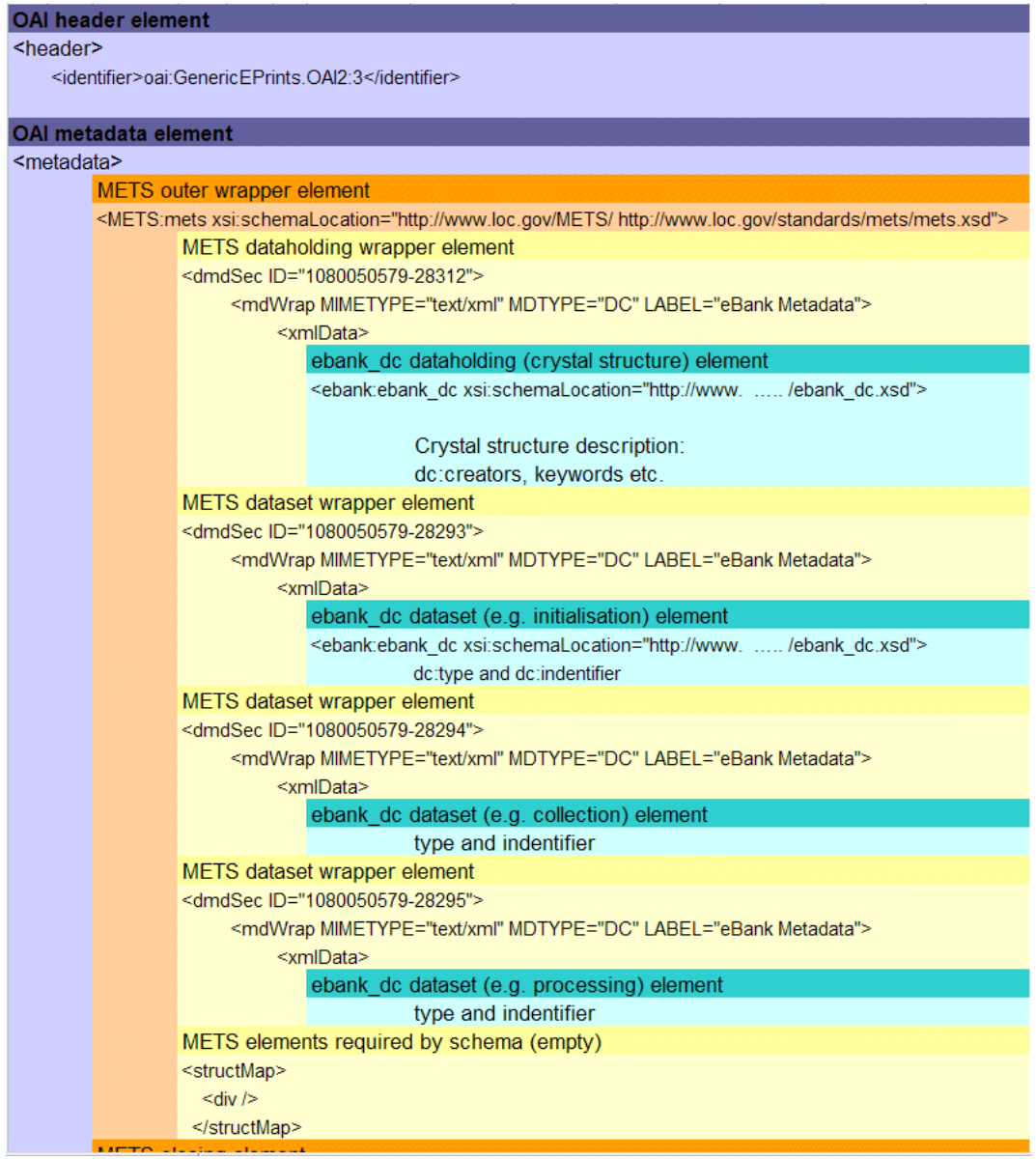
This enables our project partners at UKOLN (Bath University) to automatically extract this metadata from our archive. They can then 'aggregate' this data with similar data and even 'add value' to it. This information is then made available globally by data portals such as PSIgate (also project partners) who are members of the Resource Discovery Network (RDN).

Current Developments

We are now past the 'proof of concept' stage and hence need to apply stylesheets to the publically accessable parts of the archive in order to make an EPrint 'human readable'!

We can search on the core bibliographic data as it is in dublin core, however we need to build the crystallographic part of the search engine.

We need to incorporate some tools to facilitate the deposition of a crystal structure into the EPrints archive.



Schematic view of metadata exchanged in eBank project using OAI-PMH

What next for eBank?
The metadata schema…some issues

- Reduce to its simplest form or reflect the complexity?

  ebank_dc versus **oai_dc**

- Compatibility with other schema

  CLRC Scientific Metadata Model v. 1.0, 2001 (under revision)

- Investigate packaging options, e.g.

  METS, MPEG 21 DIDL

- Integration with library data?

  Functional Requirements of Bibliographic Records is based on a conventional model of publication, dissemination and curation but has nothing to say about pre-publication activities and the distillation of experimental material into data sets which are then described in articles. The concept of versions and revisions is not clearly articulated

Author self-archived papers (eprints): What institutions should do

Heads of schools should lead these initiatives:

• Set up a departmental eprint archive

• Adopt and promote a departmental policy encouraging all authors to self-archive

To accelerate filling of the archive:

• **Use the archive to produce departmental publication lists, manage Research Assessment Exercises (RAEs)**, etc. Authors realise that to be included their records must be complete and up-to-date

When allied to exercises such as these, authors can see a purpose in submitting and it starts to become routine.

See *OSI EPrints Handbook*: 3. Managing an EPrints Service
http://software.eprints.org/handbook/

Example institutional policy: ECS Southampton

Extracts, see full policy http://www.ecs.soton.ac.uk/~lac/archpol.html (still to be officially ratified)

1. It is our policy to maximise the visibility, usage and impact of our research output by maximising online access to it for all would-be users and researchers worldwide.

2. We have accordingly adopted the policy that all research output is to be *self-archived* in the departmental EPrint Archive (eprints.ecs.soton.ac.uk). **This archive forms the official record of the Department's research publications; all publication lists required for administration or promotion will be generated from this source.**

Experience at ECS Southampton: an RAE dry run

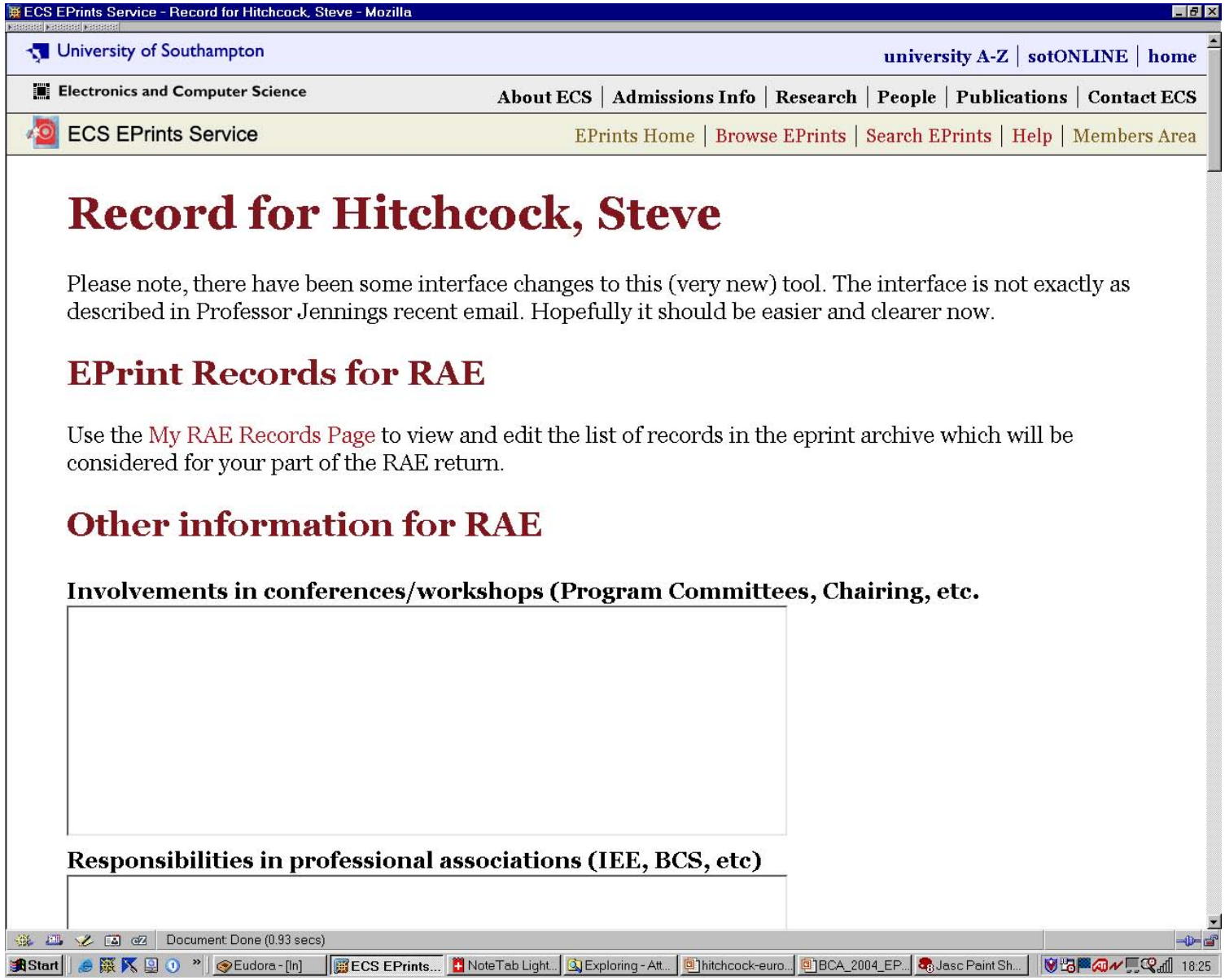
At ECS Southampton we did a Research Assessment Exercise as a dry run and it was almost painless (Hint: the pain came earlier!) **Filling the archive so it is complete is the key**.

The Eprints.org developer created a Web form for author input of honour data and a link to the author's list of publications with 'add', 'remove' buttons to select best publications for the RAE list.

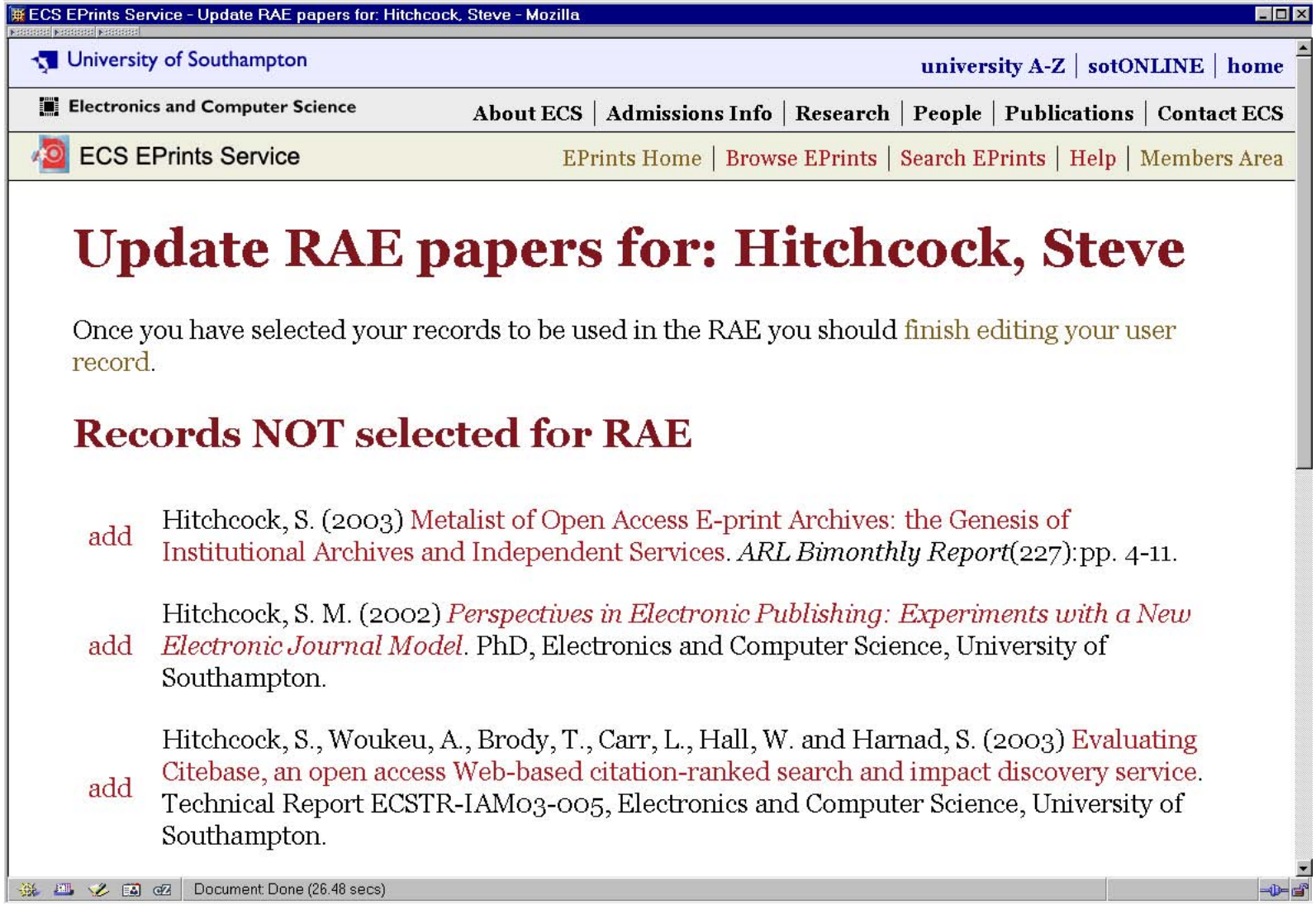
Authors appreciated the ease of completing the exercise, e.g. four clicks to select four RAE publications.

This highlights the *additional* benefits of a managed departmental archive: one-time data input for multiple purposes (avoids multiple keying for different databases for different applications).

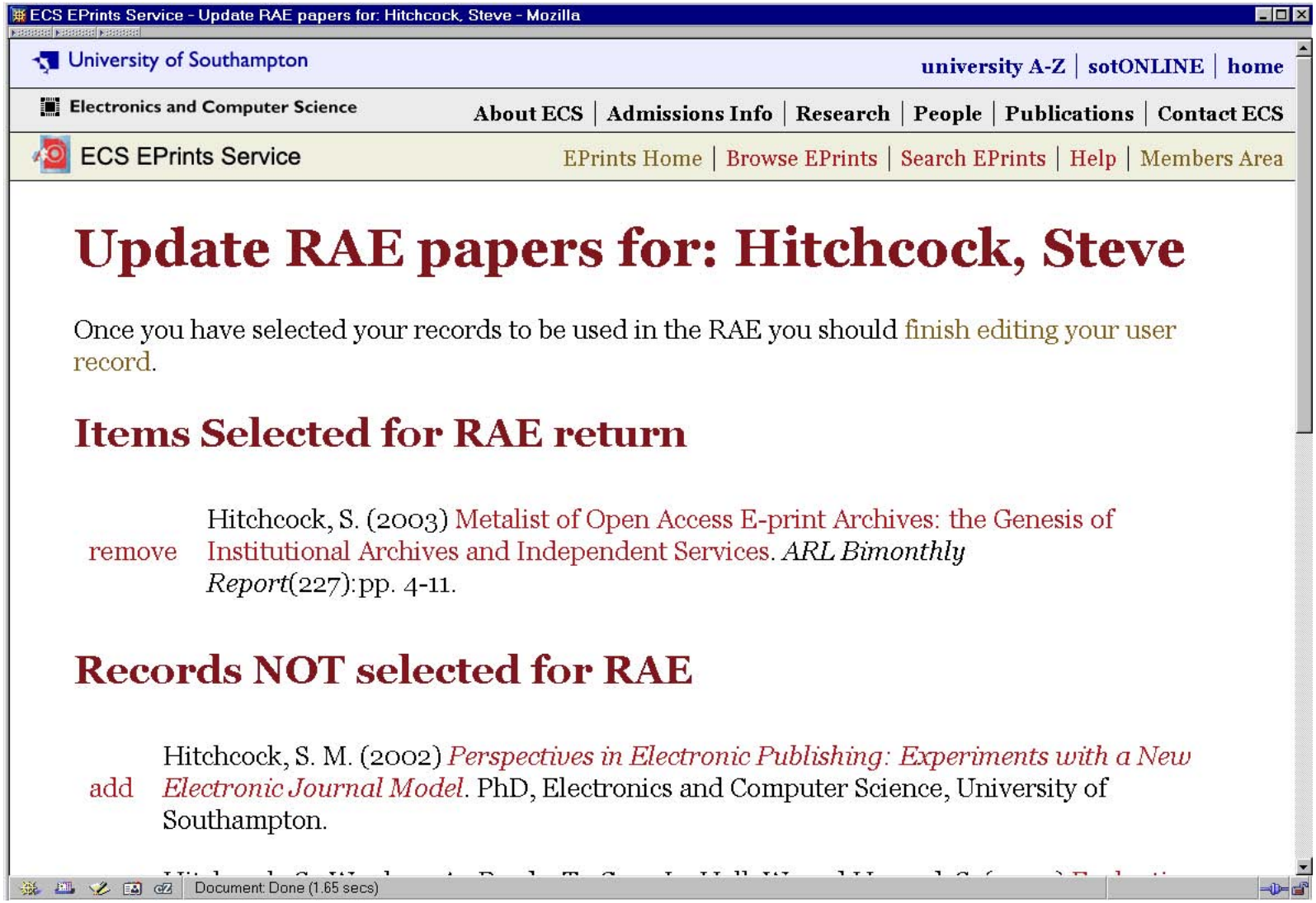
RAE dry run – author input Web form



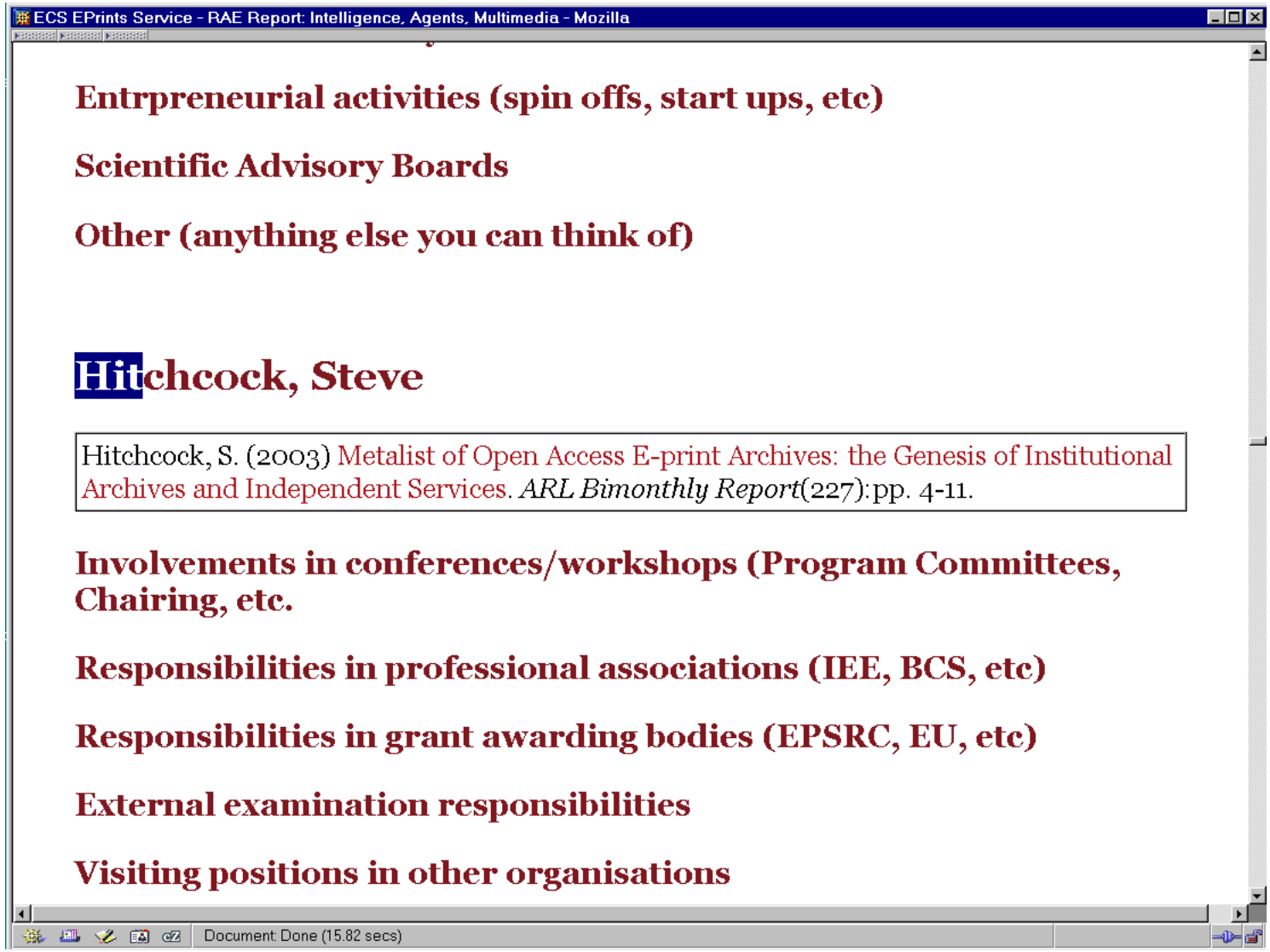
“My RAE records”



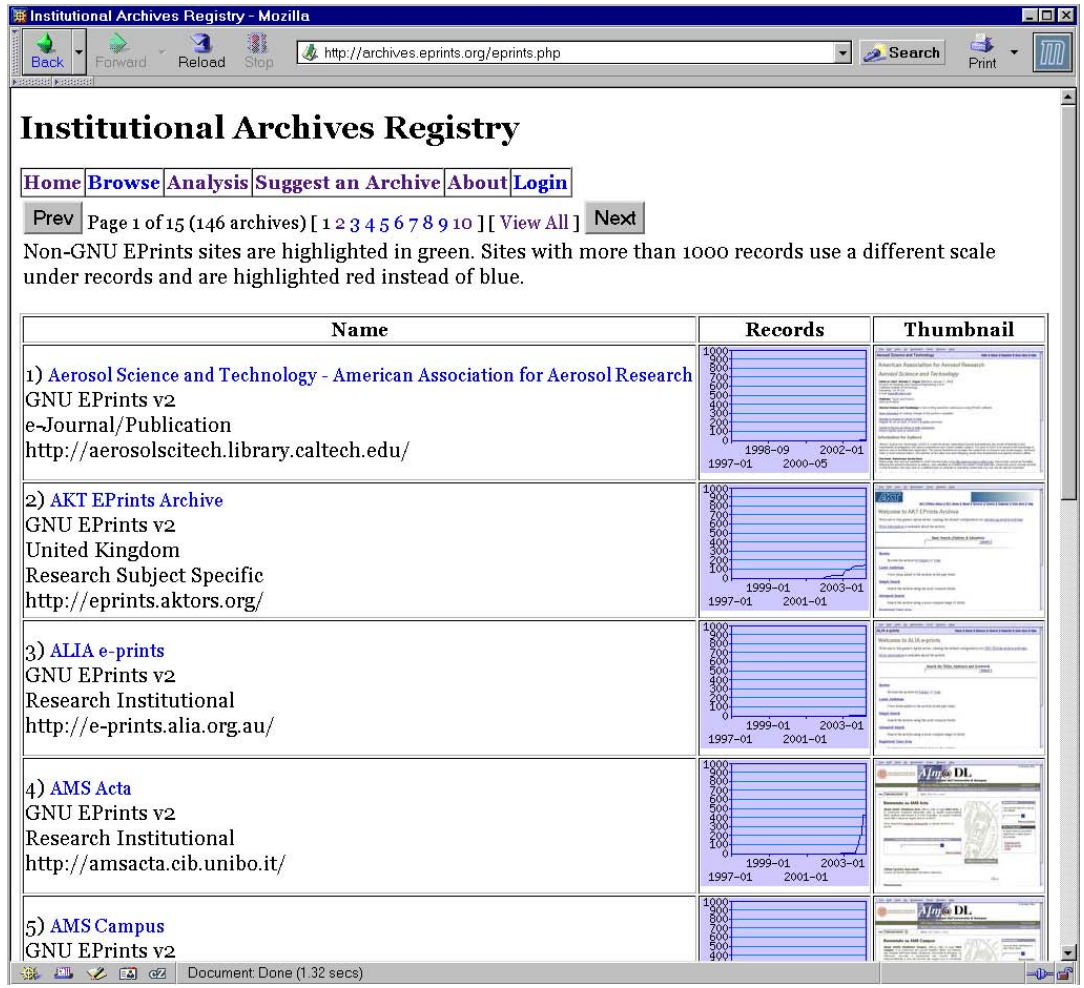
Items selected for RAE return



RAE report


ECS EPrints Service - RAE Report: Intelligence, Agents, Multimedia - Mozilla
Entrpreneurial activities (spin offs, start ups, etc)
Scientific Advisory Boards
Other (anything else you can think of)
Hitchcock, Steve
Hitchcock, S. (2003) Metalist of Open Access E-print Archives: the Genesis of Institutional Archives and Independent Services. ARL Bimonthly Report(227):pp. 4-11.
Involvements in conferences/workshops (Program Committees, Chairing, etc.
Responsibilities in professional associations (IEE, BCS, etc)
Responsibilities in grant awarding bodies (EPSRC, EU, etc)
External examination responsibilities
Visiting positions in other organisations
Document: Done (15.82 secs)

Monitor growth of institutional archives and content



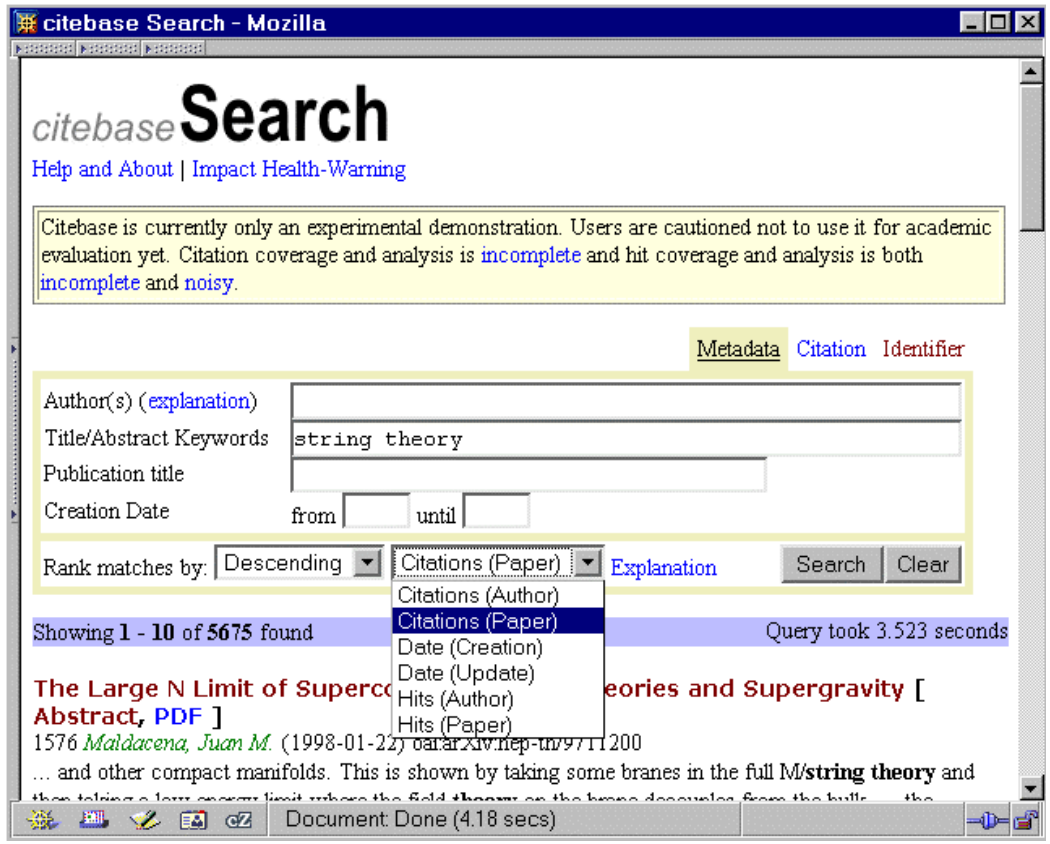
Institutional Archives Registry http://archives.eprints.org/eprints.php

Research impact

1. Measures the size of a research contribution to further research ("publish or perish"), e.g. citation-counts, co-citations, now we also have **usage-measures** ("hits", webmetrics), **time-course analyses, early predictors**, etc.
2. Generates further research funding
3. Contributes to the research productivity and financial support of the researcher's institution
4. Advances the researcher's career
5. Promotes research progress

Note the direct connection between open access, impact, research assessment and funding

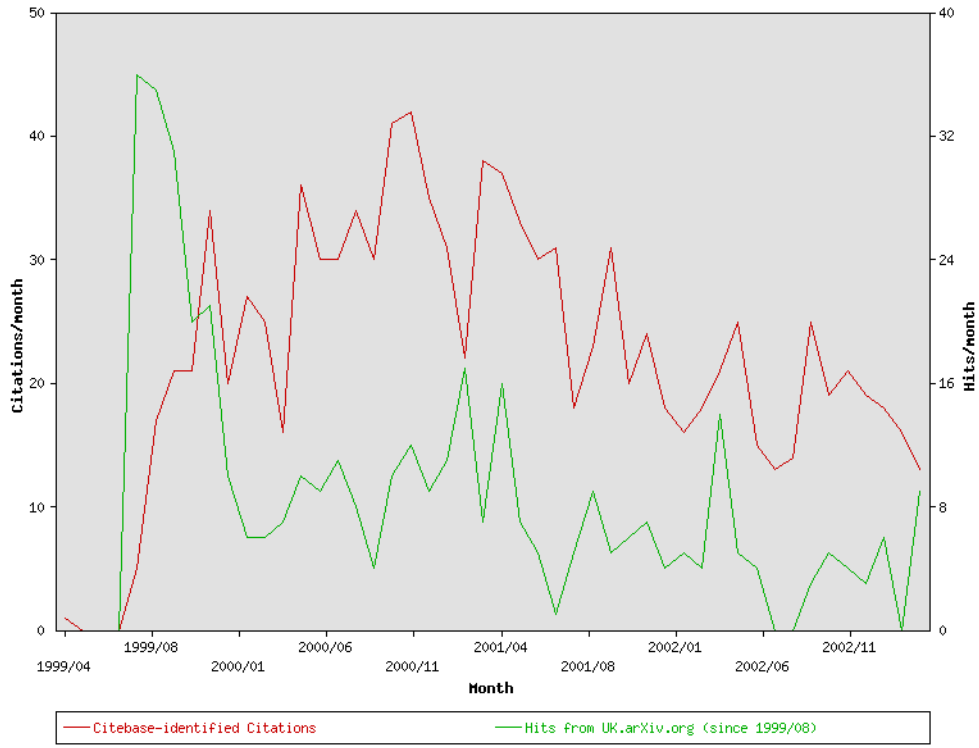
Citebase, a new interface to the scholarly literature



Citebase (http://citebase.eprints.org/) was originally produced as part of the Open Citation Project (http://opcit.eprints.org/). It is now a featured service of arXiv.

Time-course of citations (red) and usage (hits, green)

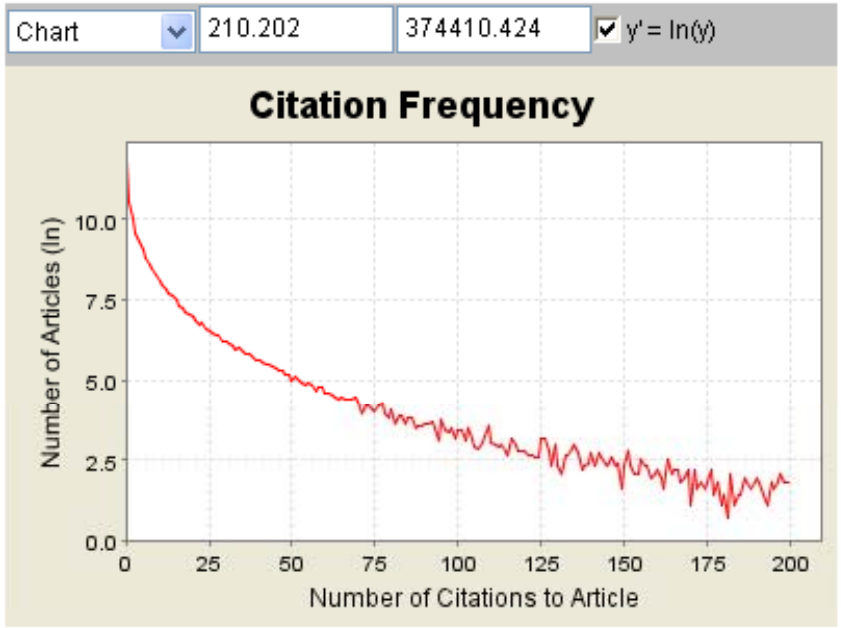
Witten, Edward (1998) String Theory and Noncommutative Geometry *Adv. Theor. Math. Phys.* 2 : 253

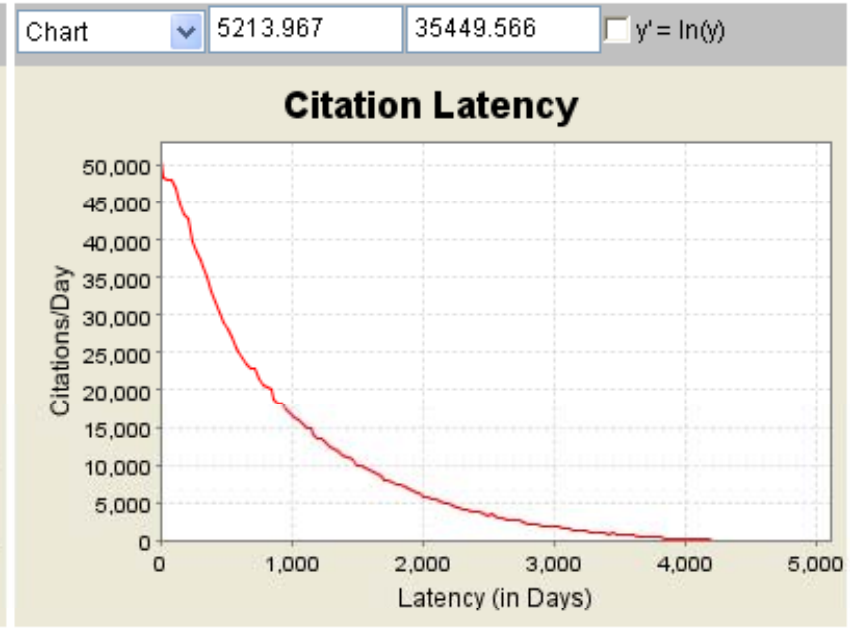


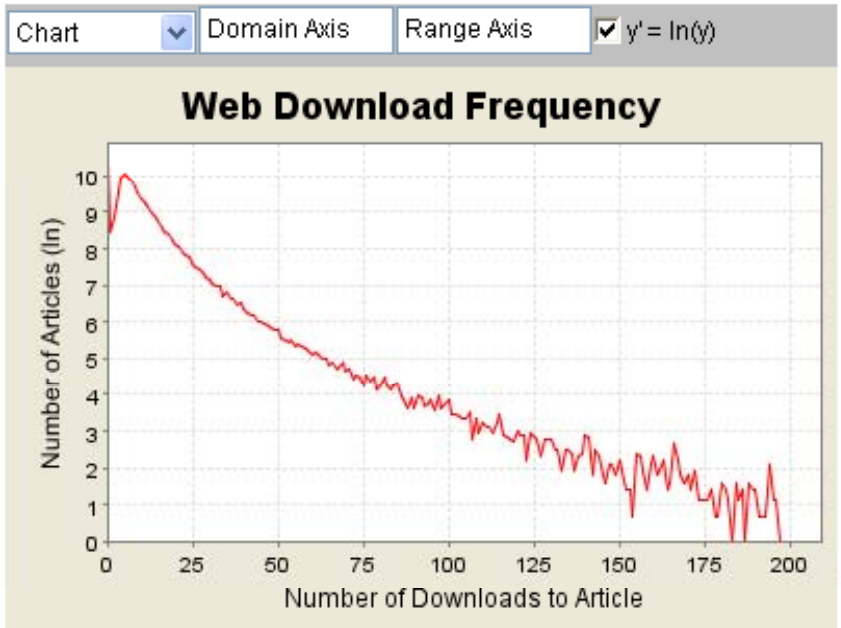
1. Preprint or Postprint appears. 2. It is downloaded (and sometimes read).
3. Eventually citations may follow (for more important papers).
4. This generates more downloads,
etc.

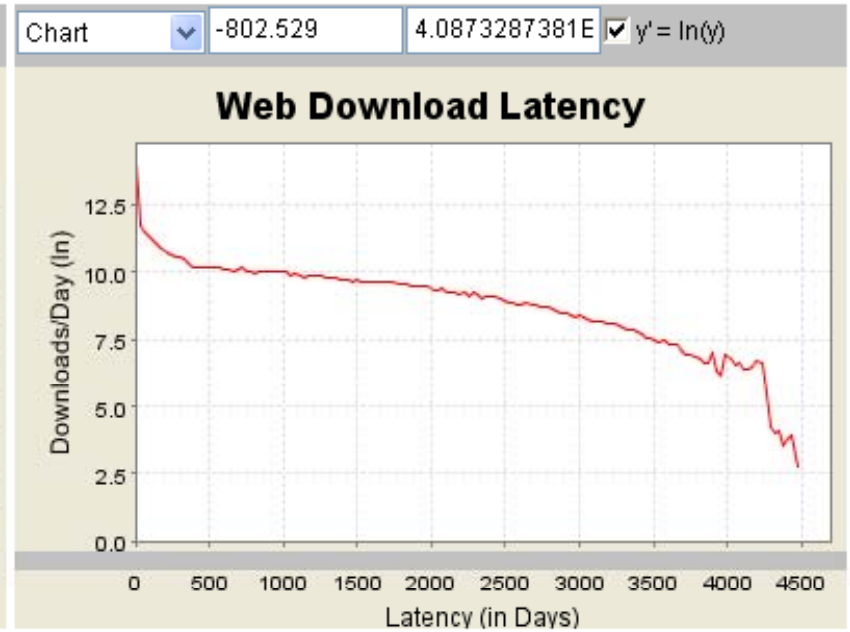
Ref. Hitchcock *et al*., "Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service". Technical Report ECSTR-IAM03-005, School of Electronics and Computer Science, University of Southampton http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report-tr.html

Correlation Generator: citations vs hits

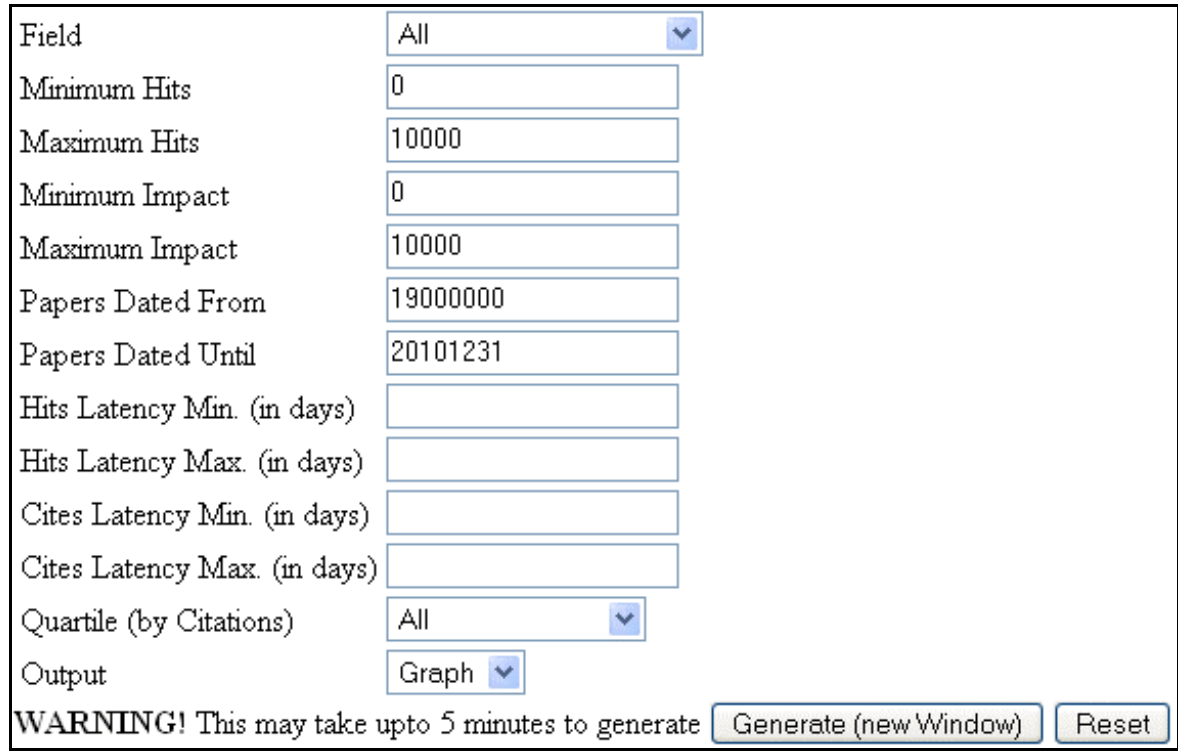






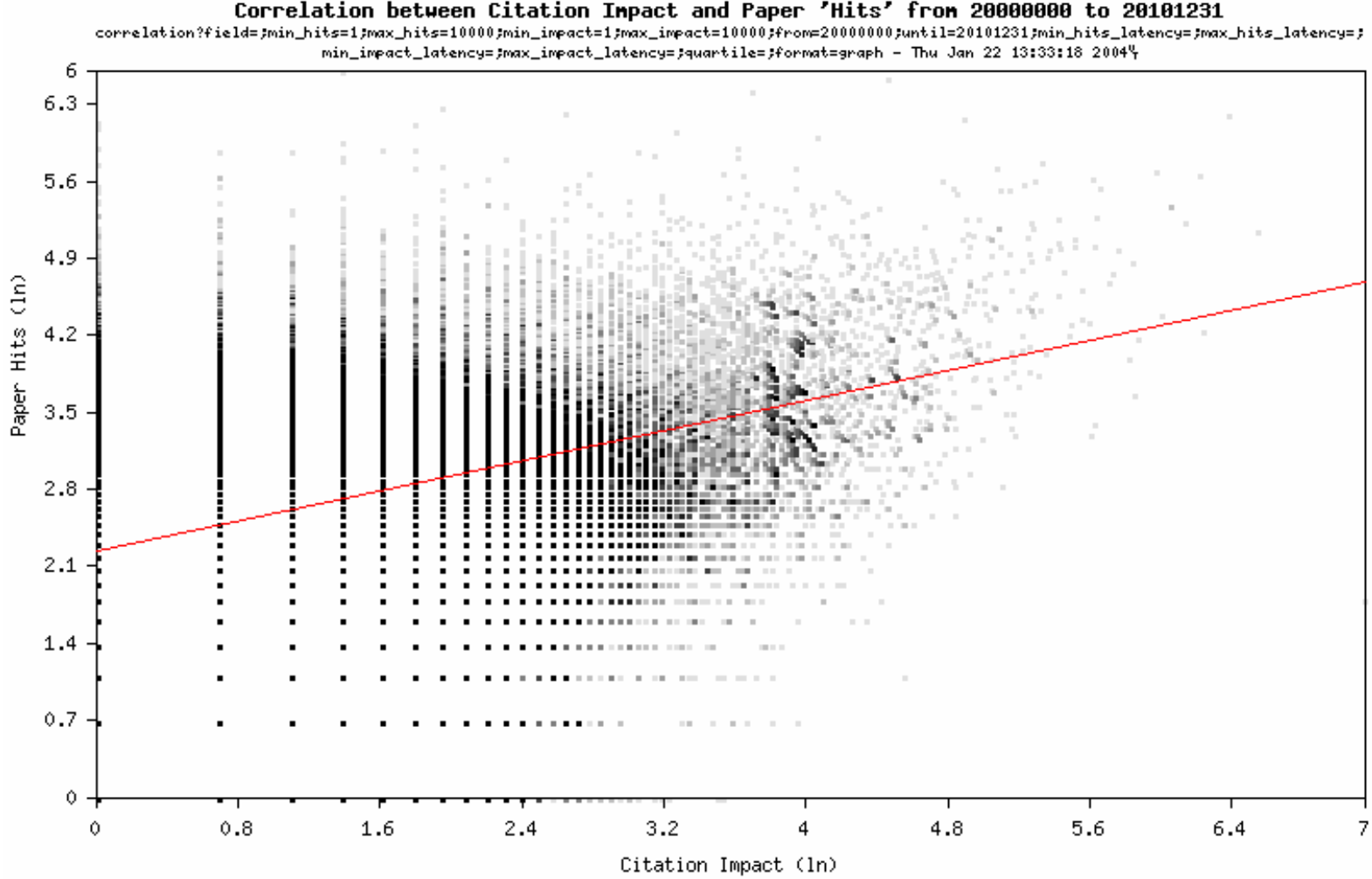


Correlation Generator: users set the parameters



Correlation Generator http://citebase.eprints.org/java/correlation/correlation.html

Warning, data-intensive Java process, can be slow to download

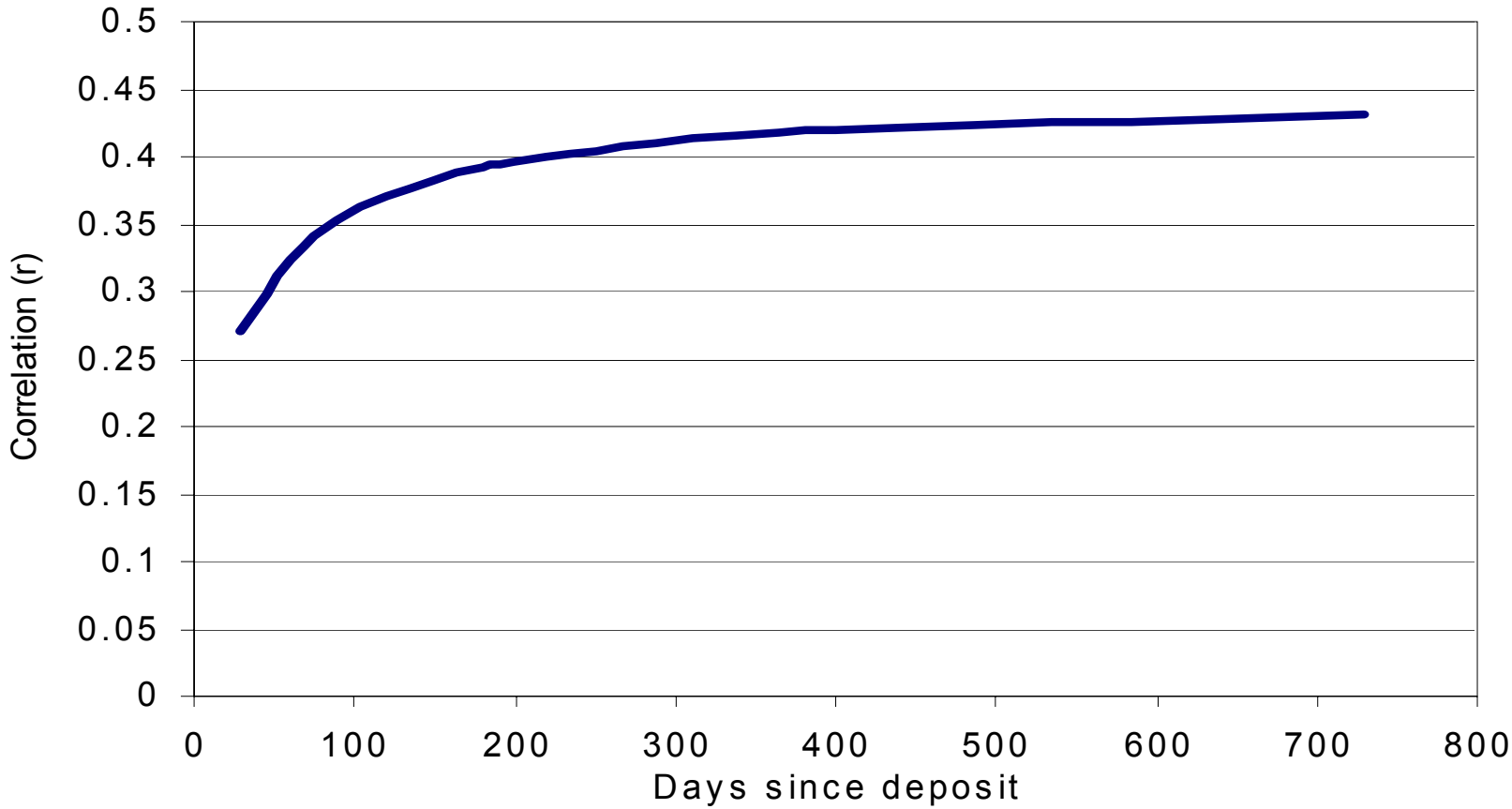


	Min Filter	Max Filter	Sum (0.43)	Mean	Standard Dev.
Impact	1	10000	83400.23	1.15	1.07
Hits	1	10000	190160.63	2.63	0.89

r (squared)	0.3432 (0.1177)
n (t)	72279 (98.2384)
Non-Dir. Sig.	<0.0001

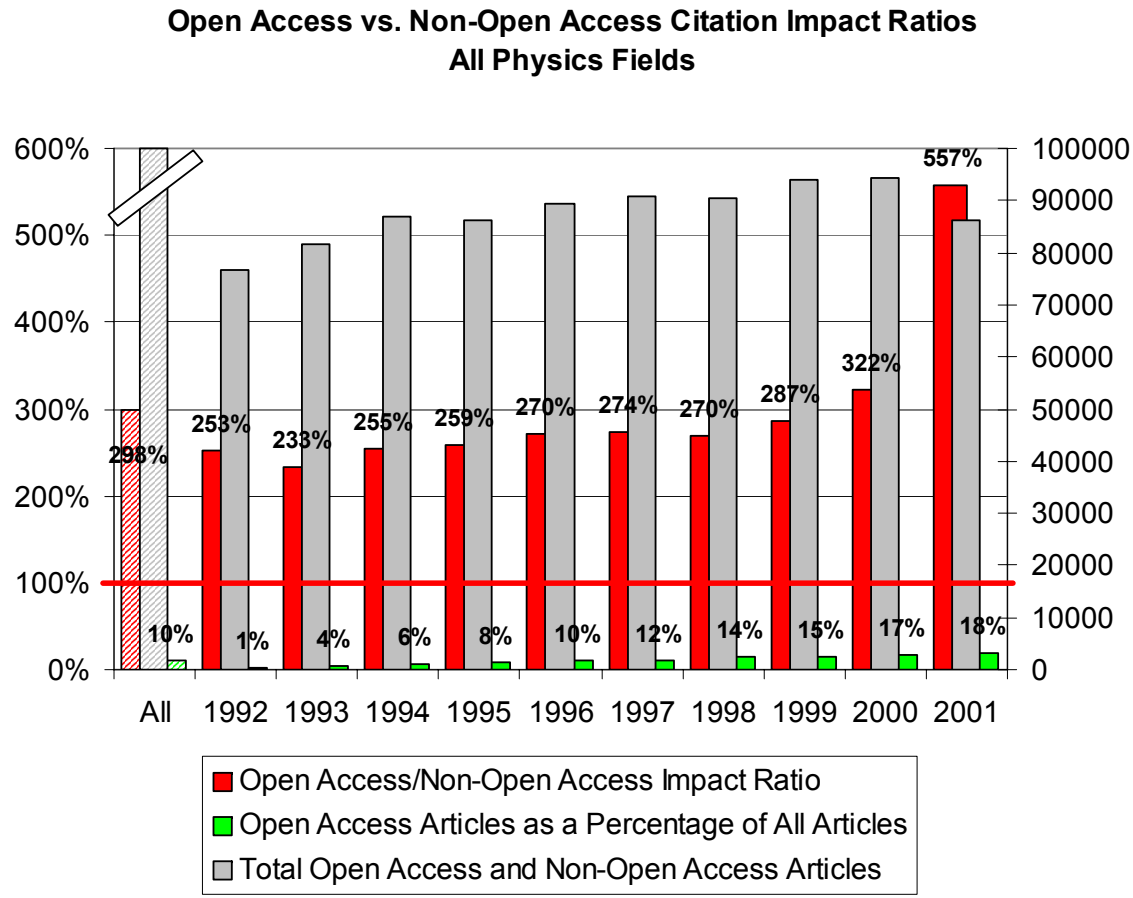
Correlation scatter-graph generated for all papers deposited between 2000-current. The correlation for these 72,279 papers is r=0.3432 (the probability that a downloaded paper will be cited). From the distribution in the scatter graph it can be seen that the distribution is noisy, but that few articles with high citation impact receive low hits impact

Correlation generator: predicting citation impact



How soon can hits impact be used to predict citation impact? This shows the correlation increases with time, approximating the final correlation after 6-7 months. (This and previous three slides from Brody *et al*., paper in preparation)

Citation impact ratios



From: Brody, T., *et al*. (2004) The Effect of Open Access on Citation Impact
http://opcit.eprints.org/feb19oa/brody-impact.pdf

Conclusion

We are seeing the emergence of a unified, but *very large*, research information system covering

- Raw research data (e.g. eBank)
- Reports, publication (e.g. ECS Eprints)
- Search, analysis and metrics for assessment (e.g. Citebase)

It is entirely digital, all made possible by open access, and is mediated via the Web.

In our case two other intrinsic components are

- **Eprints software**: for storage management and user/author interfaces
- **OAI**: for discovery

Credits

eBank @ UKOLN

• Michael Day, Monica Duke, Rachel Heery, Liz Lyon, Andy Powell

eBank @ Southampton

• Les Carr, Simon Coles, Jeremy Frey, Chris Gutteridge, Mike Hursthouse

eBank @ Manchester

• John Blunden-Ellis

Eprints.org @ Southampton

• Stevan Harnad, Les Carr, Christopher Gutteridge

• Citebase and the Correlation Generator are produced by Tim Brody

For more about Eprints.org see http://www.eprints.org/

These slides can be found at http://opcit.eprints.org/opcitpapers.shtml

Contact **Steve Hitchcock**: sh94r@ecs.soton.ac.uk