EPrints: Repositories for Grassroots Preservation

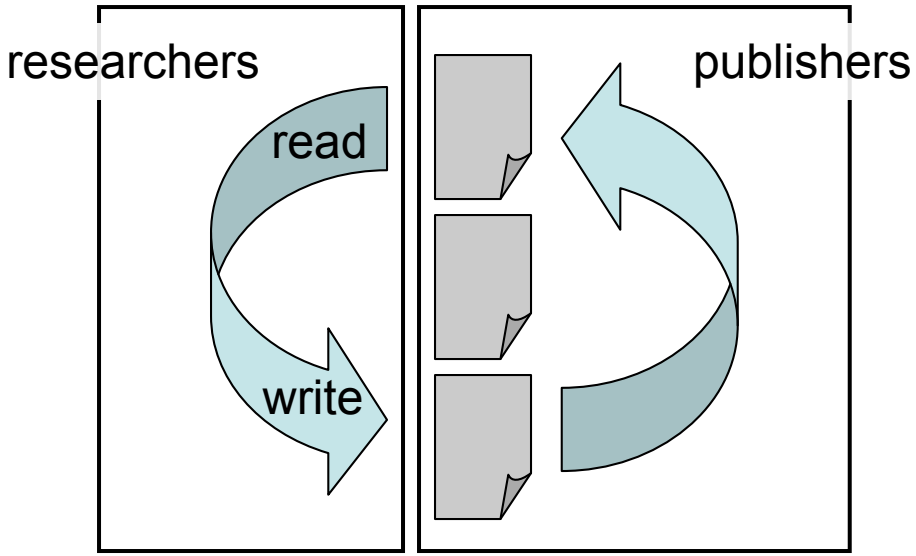
Les Carr, www.eprints.org

Grass roots: preface and précis

- The aim of this presentation is to tell a story. In the context of a meeting which has mainly dealt with the issues of national libraries and enormous digital collections, this is a presentation that addresses a different scale. It is a scale that is both smaller and larger at the same time. It is about collecting individual items from individual researchers - the so-called grass roots - through institutional repositories. Although this seems small and insignificant in comparison to tales of humongous digital collections, the day-by-day aggregated collection of individual items from a community of knowledge producers adds up to the entire scholarly and scientific literature - as well as its supporting data, experimental analyses, discussions and commentaries.
- This story about EPrints focuses on the challenges of acquiring data and documents in order to build up a global collection. The challenges listed are those relating to changing the working practices and use patterns of individuals and their host institutions in order to support long term preservation.
- The story necessarily enalrges on the need to make things easier and more useful for the author/depositor/knowledge producer in order to encourage the first stage of preservation: acquisition.

Problem Space (1)

- Universities and researchers are knowledge producers and knowledge consumers
- Scholarly communications have been outsourced
- Literally nothing to show as evidence of research activities



Problem Space (2)

- Researchers have have hard disks which are just organised enough to support daily activity
  - Disk crashes
  - Stolen laptops
  - Software upgrades that go wrong
  - Backups that never quite get restored
  - Draws and folders full of old stuff that eventually fall off the radar

- “Lost in some research assistant’s computer, the data are often irretrievable or an undecipherable string of digits”

Lost in a Sea of Science Data. S.Carlson, The Chronicle of Higher Education (23/06/2006)

Arts & Humanities Research Council
bbsrc
biotechnology and biological sciences research council
Science & Technology Facilities Council
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
EPSRC
Engineering and Physical Sciences Research Council
Congratulations on your new research project!

This is where your hardware will end up

Make sure your data doesn't!

Research outputs go in research repositories

eprints

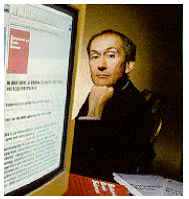
UK Experience

- UK Council of Research Repositories
  - platform agnostic
  - group of repository managers
  - speaks for repository managers

- Most repositories
  - have a part time manager
  - receive little or no technical support

EPrints History

- Open Archiving Initiative - October 1999
  - Originally called UPS
- Among the Participants
  - Paul Ginsparg (Los Alamos, arXiv)
  - Carl Lagoze (Cornell, NCSTRL)
  - Stevan Harnad (Southampton, Cogprints)



- EPrints
  - proposed as a ‘build your own repository’ solution
  - enable institutions and groups to participate in OAI metadata sharing initiative

EPrints History

- First released April 2000
  - to co-incide with OAI-PMH
- Version 3.0 released in Jan 2007
  - at Open Repositories 2007
- Strongly backs *Open Access*
- Used by over 240 registered repositories

The LINNEAN SOCIETY of London

EPrints Management

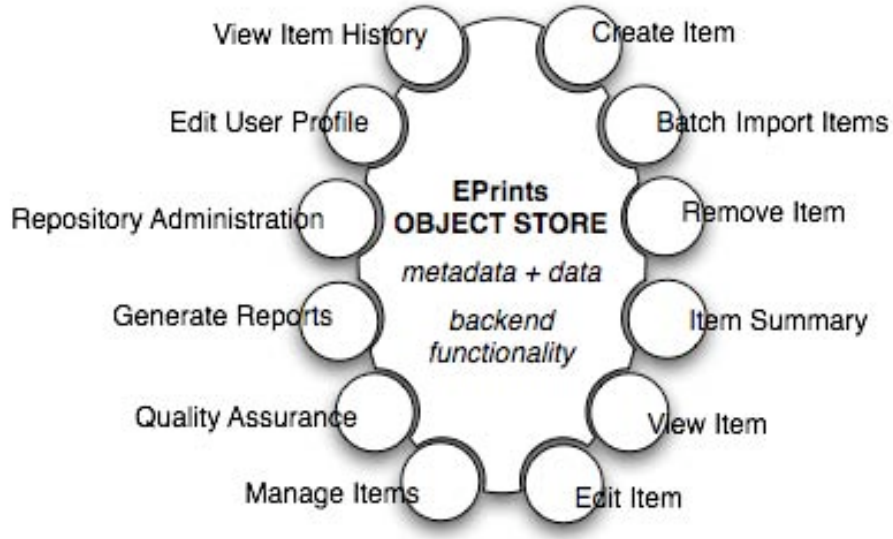
- Open source (GNU license)
- EPrints development model is more centralised than DSpace / Fedora
  - *c.f.* the original problem statement
  - pros and cons *e.g.* faster turnaround on development cycles, more focused, easier quality management
  - All of these platforms are hybrid open source - they were initially bankrolled!
- EPrints Commercial Services
  - repository hosting, bespoke development & training
  - sustain the development team

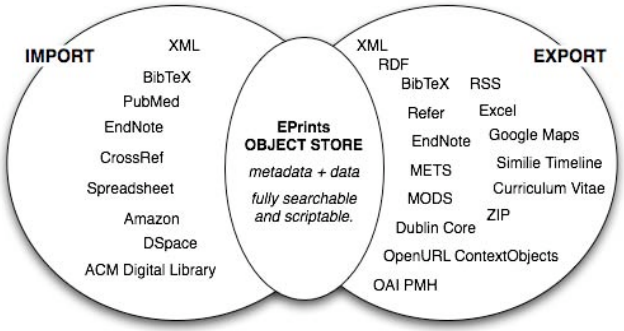
EPrints Core Objectives

- Lower the barrier for depositors while improving metadata quality and ultimate collection value
  - Time saving deposits
  - Import data from other repositories and services
  - Autocomplete-as-you-type for fast data entry
  - Name authorities
- Enter once, reuse often
  - Works with bibliography managers, desktop applications and new Web 2.0 mashups
  - RSS feeds and email alerts keep you up to date
  - Easily integrate reports, bibliographic listings, author CVs and RSS feeds into your corporate web presence
  - Used for corporate reporting and national Research Assessment
- Simple platform for open source contributions
  - Tightly-managed, quality-controlled code framework
  - Flexible plugin architecture for developing extensions

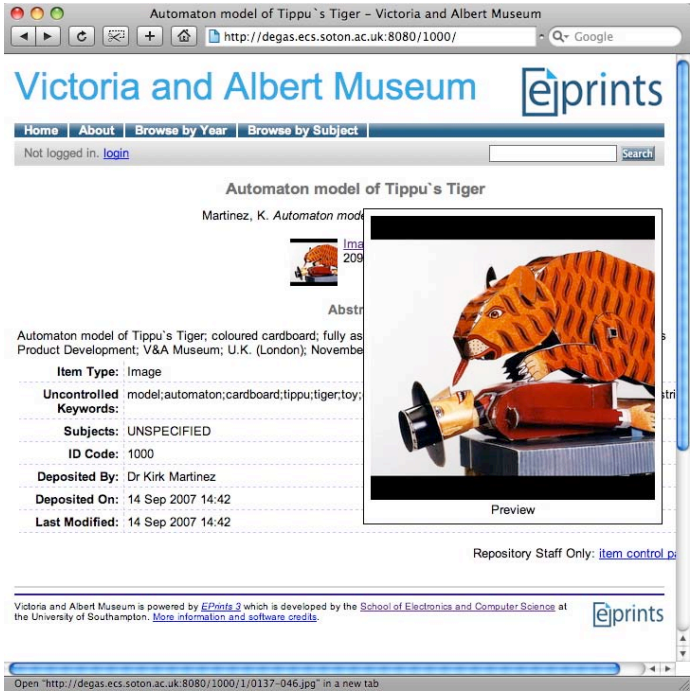
EPrints Flexibility

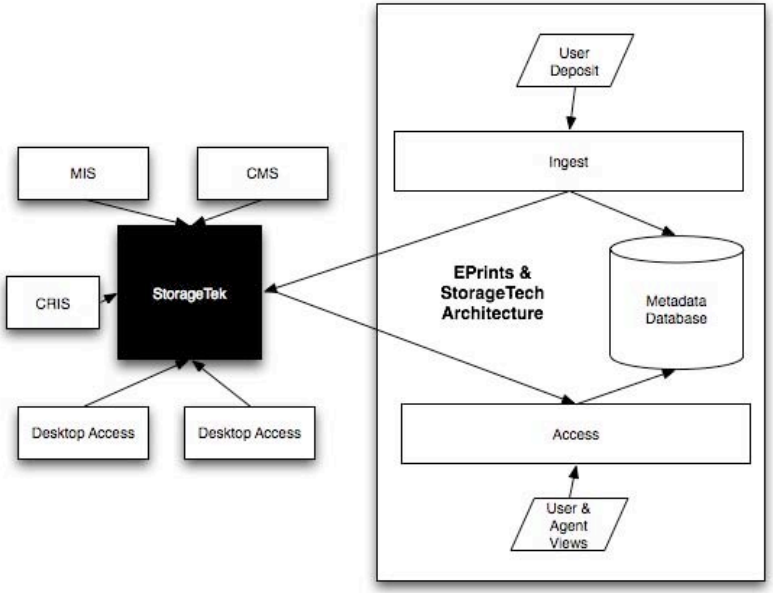
- EPrints backend
  - object store
  - API
- EPrints frontend
  - Screen plugins
  - User interface + methods + REST interface





EPrints + Honeycomb





- **Jam today** - large self-managing storage extends repository bang for library buck
  - New chemistry & artistic objects to be collected
- **Jam tomorrow** - potentially take over part of repository responsibility

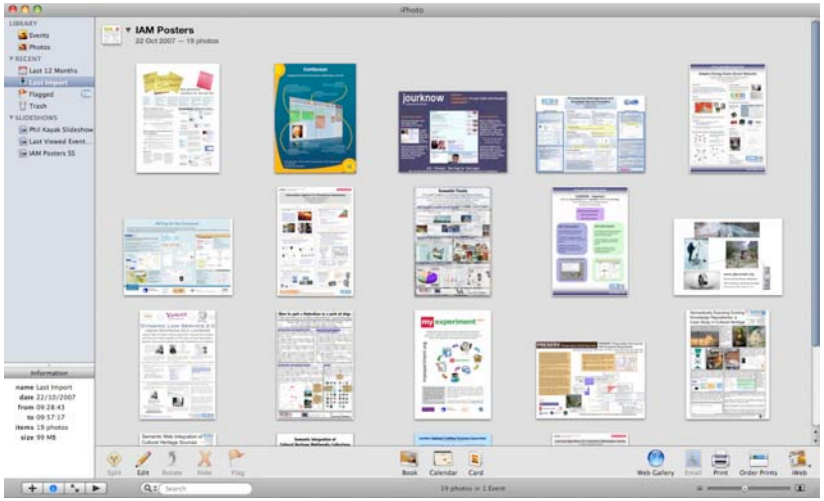
EPrints Challenges

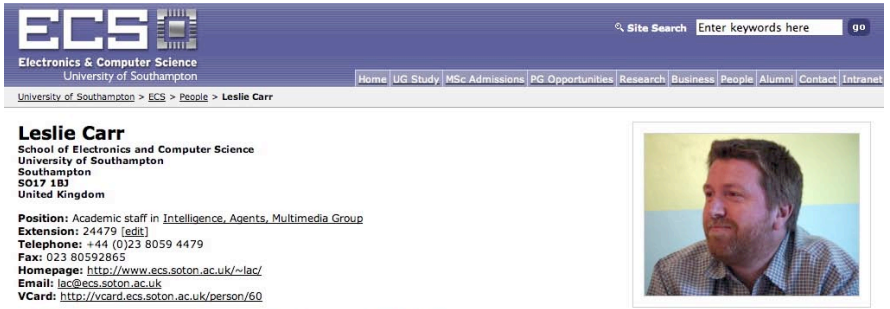
- Small science > big science
  - Data from Big Science is easier to handle, understand and archive. Small Science is horribly heterogeneous and far more vast. In time Small Science will generate 2-3 times more data than Big Science.

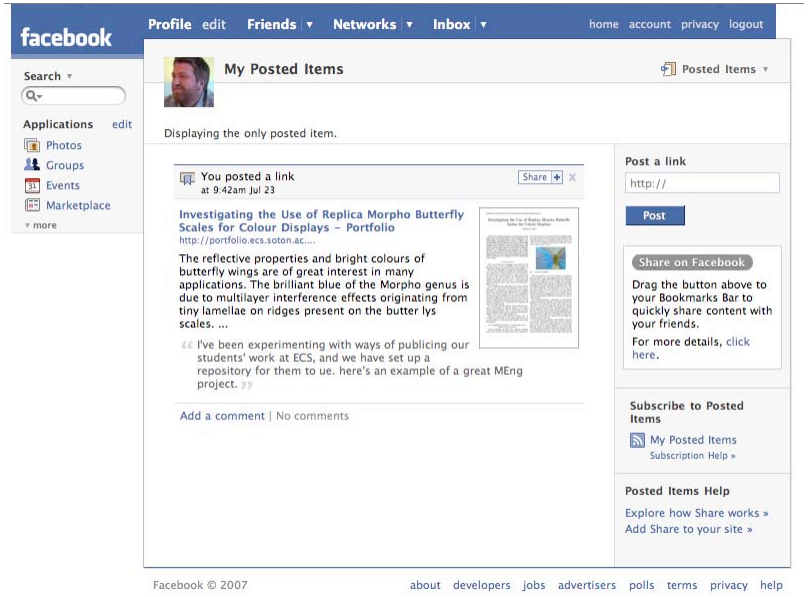
- Lots of inexperienced users
  - Give individuals the tools to become responsible curators of their own intellectual output
  - Give institutions the tools to manage, assist and leverage
  - Give users the tools to access the global literature data - to use and reuse for many, many purposes in many, many contexts by many stakeholders

(Tim O'Reilly)

It's the Data, Stupid

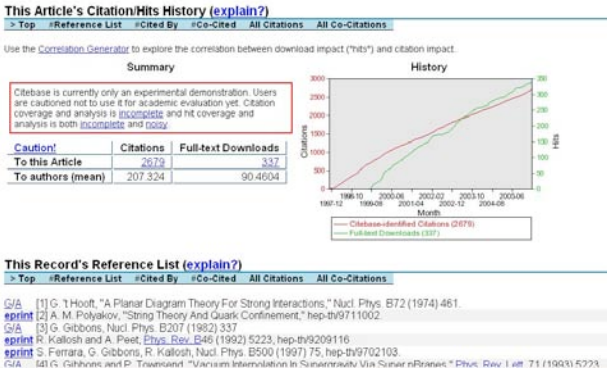


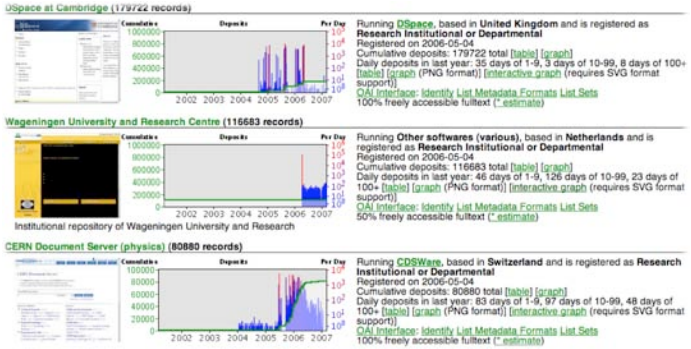




EPrints - beyond the repository

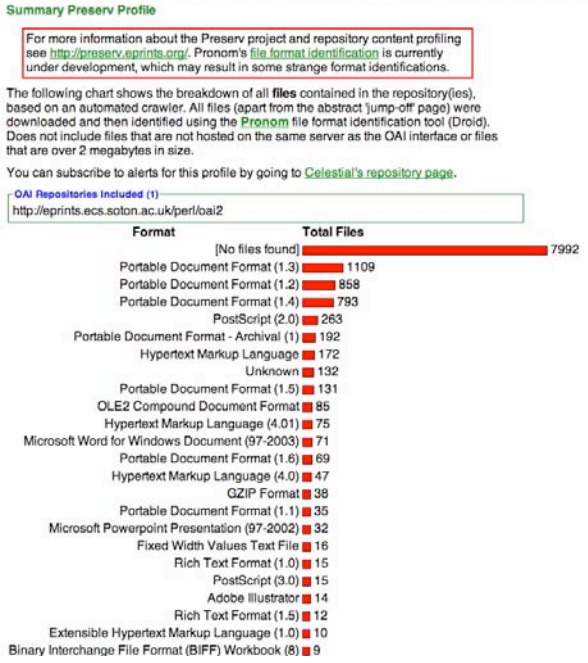
- OAI PMH services
  - Citebase - citation analysis for the Open Access literature.
    - Unfunded PhD work (outshoot of OpCit)
    - 4 million sessions per month
    - Destroy 1 RAID disk every 6 months
  - Celestial - OAI-PMH harvesting proxy
    - Supports Citebase and other services
  - ROAR - registry of Open Access repositories
    - Tracks size and daily deposit profiles over time





EPrints - Preservation Services

- Format profiling using PRONOM-DROID
  - JISC PRESERV project
  - Initially to be applied to two pilot repositories
  - Ultimately applied to over 200 repositories
    - DSpace & EPrints
    - Applied via OAI
    - Delivered through ROAR


Summary Preserv Profile

For more information about the Preserv project and repository content profiling see http://preserv.eprints.org/. Pronom's file format identification is currently under development, which may result in some strange format identifications.

The following chart shows the breakdown of all **files** contained in the repository(ies), based on an automated crawler. All files (apart from the abstract 'jump-off' page) were downloaded and then identified using the Pronom file format identification tool (Droid). Does not include files that are not hosted on the same server as the OAI interface or files that are over 2 megabytes in size.

You can subscribe to alerts for this profile by going to Celestial's repository page.

OAI Repositories Included (1)

http://eprints.ecs.soton.ac.uk/perl/oai2

Format	Total Files
[No files found]	7992
Portable Document Format (1.3)	1109
Portable Document Format (1.2)	858
Portable Document Format (1.4)	793
PostScript (2.0)	263
Portable Document Format - Archival (1)	192
Hypertext Markup Language	172
Unknown	132
Portable Document Format (1.5)	131
OLE2 Compound Document Format	85
Hypertext Markup Language (4.01)	75
Microsoft Word for Windows Document (97-2003)	71
Portable Document Format (1.6)	69
Hypertext Markup Language (4.0)	47
GZIP Format	38
Portable Document Format (1.1)	35
Microsoft Powerpoint Presentation (97-2002)	32
Fixed Width Values Text File	16
Rich Text Format (1.0)	15
PostScript (3.0)	15
Adobe Illustrator	14
Rich Text Format (1.5)	12
Extensible Hypertext Markup Language (1.0)	10
Binary Interchange File Format (BIFF) Workbook (8)	9

- Add Honeycomb to the mix
  - We can 'preserve' repository contents too
  - JISC PRESERV II project

The challenges of human scale institutional repositories versus the challenges of industrial-scale processing of humongous collections.
Lawnmowers *vs* Combine Harvesters?
How do you manage an entire nation's grass clippings?