eResearch: librarians pushing technology to perform

Martie van Deventer
Portfolio Manager
CSIR Information Services

IUG Directors’ meeting
Port Elizabeth 12 November 2008
Changes in Traditional Research

Global Collaboration

Visualisation

Personalisation

e-Research

Opportunities

New technologies

Previously impossible research

Cyber infrastructure

Big science

VRE

South Africa

International

Definitions

Multidisciplinary

HPC
Research is changing

Today much of the work of scientists is internet-based. From their desks scientists access special hardware, software, data and applications as parts of distributed systems. Typically access to them has to be made explicit by keying in specific data. Cooperation with other scientists (e.g. the exchange of interim or final results of research for annotation and further use) typically takes place via the internet. (Osswald, 2007: 517)

Even if publications are still the currency of tenure, data is the currency of science, (Gold, 2007) … !
Ready for eResearch?

• My research involves **working with researchers** from other organisations in South Africa/ Africa and/or overseas …

• Data and/or **equipment** that would be useful to me are located **elsewhere** …

• I have data and/or equipment that other researchers would like to access …

• I need to send/receive **large quantities of data** (e.g. audio, video, datasets) …

• I am aware of international research initiatives in my area and would like to participate in them …

Based upon: [http://www.mcs.vuw.ac.nz/EResearch/KAREN](http://www.mcs.vuw.ac.nz/EResearch/KAREN)
Ready for eResearch?

- I do not **collaborate with other researchers** in other organisations but I would like to in future
- I organise / attend meetings regularly with other researchers
- I have data and/or equipment that other researchers would like to access
- I use software programs to analyse data - it is very complex / takes a long time but produces a **graphic** that I can interpret
- I often email/phone colleagues elsewhere in South Africa/ Africa and overseas to discuss research projects and common research interests

Based upon: [http://www.mcs.vuw.ac.nz/EResearch/KAREN](http://www.mcs.vuw.ac.nz/EResearch/KAREN)
eResearch: challenges for the librarian

• Most librarians are much less familiar with the data-generating research phases of the scientific research cycle than with post-research.

• Data science and data management are an awkward fit with the text-oriented constructs and systems that still dominate library relationships with science communication and publishing.

• Becoming literate in cyberinfrastructure means understanding cyberinfrastructure, eResearch, collaboratories, collaboration science, computational and grid science, data curation, the Semantic Web, open data, data archiving, digital preservation, and data management, and how they relate to each other.

• Phases of reporting, communication and publication are less distinct.

• When used in discussions of eResearch, a vocabulary familiar to librarians (archival, curation, stewardship, provenance) takes on new or specialized meaning … major cause of confusion.
• **What is currently the same**
  
  • Having to adapt to change
  
  • Researchers currently still regard libraries as trusted service providers … but for how long?

• **What’s Different**
  
  • Collaboration … science has to become big and multi-disciplinary to solve the really big challenges
  
  • Research often relies on visualisation
  
  • Researchers need partners in research rather than just suppliers and support staff
  
  • Expanded skill sets – need to learn to do different tasks
What is eResearch?

- eResearch is research done faster, to
- a better quality, or
- by different methods, using
- advanced digital tools and
- services that enable diverse research expertise to be
- assembled in global teams focused on
- very specific research problems

## Components of eResearch

<table>
<thead>
<tr>
<th>eScience</th>
<th>Digital Curation &amp; Preservation</th>
<th>Access to eInformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data transfer &amp; Computation</td>
<td>Tools &amp; applications</td>
<td>Commercial Publishing</td>
</tr>
<tr>
<td>Primary Data Sharing</td>
<td>Digital Curation &amp; Preservation</td>
<td>Open Access Publishing</td>
</tr>
</tbody>
</table>

**by definition this is**

- **Science employing transfer and sharing of large volumes of data**
- **Software that allows manipulation, modeling and analysis of data**
- **Making research data available to other researchers**
- **Active management of databases including promotion of effective and widespread use of the datasets for their scientific & scholarly useful life**
- **Contribution to & use of published resources requiring payment by readers**
- **Contribution to & use of published resources where content is regarded as 'free'**

**which requires**

- **Access to remotely held large datasets & high performance computing via affordable high bandwidth**
- **Access to models, source code and open standards**
- **Accessible repositories & quick reference**
- **Preservation & curation repositories & access mechanisms, archival skills & infrastructure**
- **Affordable licenses for researcher access & discovery mechanisms**
- **Serviceable infrastructure for publication and access**

**Researcher Requires:** Perpetual access, Curation, Training, Marketing

**Supplier must ensure:** Security - Access, Authorization, Authentication
Examples of eResearch: Serious Disease Genes Revealed

- Wellcome Trust Case Control Consortium
- 50 research groups
- 200 scientists
- DNA from 17,000 patients
- 15,000 polymorphic markers
- Learned more in 12 months than last 15 years

http://www.mcs.vuw.ac.nz/eresearch
Functional MRI (fMRI) Data Centre

- Online repository of neuroimaging data
- Typical study comprises:
  - 3 groups
  - 20 subjects/group
  - 5 runs per subject
  - 300 volumes per run
  - 90,000 volumes,
  - 60GB raw data
  - 1.2 million files processed
- 100s of such studies
Collaboration, co-ordination & communication

Desktop video-conferencing

Skype

Video-conferencing

Twitter: 21st century telegraph
Visualisation

Tools: proprietary eg Mathlab but open source is rapidly gaining market share

http://www.mcs.vuw.ac.nz/EResearch/Symposium
Working in multidisciplinary teams

Expect: jargon confusion, language barriers, cultural barriers, etc.
The first phase of the computational procurement consists of High-Speed e1350 Linux computer cluster with 160 nodes:

- Each node is equipped with two dual-core AMD Opteron 2.6GHz processors and 16GB of random access memory
- Shared file system from the SAN accessed by the nodes over the infiniband 10 GB cluster interconnecting Clearspeed Floating Point Accelerator
- An aggregate of 640 processing power at approximately 2.5 Terraflops per second
- In addition to local hard disks, all nodes have access to a shared storage system with a capacity of 94TB using General Parallel File System (GPFS)
- The cluster platform, is aptly named “iQudu” (isiXhosa name for Kudu. The name symbolises the agility, speed and size of the cluster). Cluster: iQudu
Examples of what will CHPC be doing

- Develop new regionally specific climate change scenarios for 2025; resource planning in the Western Cape based on models developed from quantifying rainfall vulnerability coupled with impact of increased population growth in the region.

**Speed up the development of vaccines** by studying fragments of 'foreign' proteins that have been shown to trigger a response by the immune system; search for common patterns in the sequences or structures of those proteins and use that information to create a theoretical model of the features that cause the fragments to trigger the immune response. These methods are specifically applicable to the AIDS and malaria vaccine projects.
What are eResearchers worried about?

- How do we describe & catalogue the artefacts, resources, experiments & knowledge we have so that others can find, use, and understand these things?
  - What are effective carriers of meaning (between researchers) in our field?
  - Can these carriers be represented / emulated in systems? How?
  - How much do we need to rely on top-down imposition of meaning? And how much on bottom-up?
  - Can we reason across geological settings?

- How do we fit knowledge computing into current work practices?
  - How do we engineer **knowledge capture** so that it is as **unobtrusive** as possible?
  - How do we maximize utility of captured knowledge in future work?
  - Governance: **who gets to make the rules**?
Further challenges facing eResearchers

- Preserving digital content and even understanding what "preservation" means
- The absence of practices for refreshing and migrating both data and relevant retrieval software
- Capture (archiving) and selection, as well as providing sufficient metadata or other descriptive or administrative information to ensure adequate (as well as legal) access and retrieval over the long term to electronic content created
- Insufficient knowledge of the software applied and an inability to find relevant information
- Missing or ignoring standards as well as obstacles related to licensing and accountability
- **Bottom line:** without ‘librarian’-skills (adapted to the eResearch environment) researchers will not be able to benefit fully from their own past – they do not wish to be librarians
FUNDING
- No detail planning
- Variety of sources
- Generic plans
- Awareness of need to make data available

DATA COLLECTION
- Varied origins, formats, sizes, usefulness to others
- Data collection from printed

PROCESSING
- Stored on departmental servers or desktops
- Some horror stories
- Poor annotation
- Sharing by email or portable media
- Storage and sharing problems with big datasets

PUBLICATION
- Few deposits in national archives
- Publication on the web
- Usefulness of linking data and publications

SUPPORT
- Little and mainly from IT officers

Martinez-Uribe, 2008
When asked, Oxford scientists requested:

- **Advice on practical issues** related to managing data across their life cycle. This help would range from assistance in producing a data management/sharing plan; advice on best formats for data creation and options for storing and sharing data securely; to guidance on publishing and preserving these research data.

- **A secure and user-friendly solution** that allows storage of large volume of data and sharing of these in a controlled fashion way allowing fine grain access control mechanisms.

- **A sustainable infrastructure** that allows publication and long-term preservation of research data for those disciplines not currently served by domain specific services such as the UK Data Archive, NERC Data Centres, European Bioinformatics Institute and others.

- **Funding** that could help address some of the departmental challenges to manage the research data that are being produced.
Single experiment view of data

- Parameter Name
- Date
- Sample ID
- Location

Source: McCord, R. 2005
Different storage areas envisaged in an scientist’s ideal environment, and the relationships between them (Hunter)
Malaria VRE demonstrator

Open space

Personal space

Commercial Resources via Library & IP authorisation

Google tools for document management

Malaria Initiative will facilitate an integrated programme of malaria research and capacity development in South Africa and in the rest of Africa to improve malaria prevention and control. Modern research tools will be applied to malaria research. Outputs will include the identification and validation of drug and insecticide targets, development of insecticidal candidates, improved diagnostics, and new tools for gathering epidemiological information.

Shared space

WebCoLab

ADVT: Slide in Banner Exchange for Webmasters

MSM/Google Groups

Blogger

Media Wiki

DSpace
What can librarians do? … Traditional …

- Select, acquire, and license data and data sets;
- Create metadata (or metadata standards) for discovery and description of data sets … but at eResearch scale
- Creating and/or organize documentation related to data
- Offering preservation services for digital data
- Advise on the appraisal and selection of what data to keep for the long term
What can librarians do? … Traditional 2 …

- Assist users with finding data relevant to their research, using third-party high level directories and data discovery sources
- Develop data publication standards and systems
- Publishing workflows, global identifier schemes, linking schemes, standards for data clean-up and normalization, and also standards for providing credit and recognition to data authors
- Develop practices for rights management ( – how do you site data sets?)
- Offer long-term repositories of scholarly output
What can librarians do? … New

• Position themselves as partners in research – embedding themselves
• Create more dynamic repositories that support pre-publication workflows
• Learn to practice eResearch themselves!
eResearch Support Framework

Senior Management

Development & innovation

- Trial/experimental Resources/Tools and Activities
- Commercial Resources
- Free Resources
- Tools
- New IT solutions (eg HPC)

New initiatives
- Blogs
- Database of Digitization of maps
- Creation of integration standards
- Identification and evaluation of research tools
- VREs
- Repositories

Function

Activities

Service delivery

Unique Resources/Tools
- Commercial (eg SA Journal of Mining)
- Internal (eg Tree database)
- Tools

Shared Tools, Resources & Services
- Commercial (eg SciencDirect)
- Open Access
- Internal (eg ToDB)
- Tools (eg RefWorks)

IT Infrastructure
- Necessary bandwidth
- Link to SANReN/ Cluster/ Cloud
- Maintenance (including security and back-up)

Responsibility

Lead User/ Researcher’s Forum
Learn to visualise (eg find out more about our users)
The architectural wealth of the University of Pretoria represents all the schools of thought of the most important building styles of the past hundred years. It unites a transient and highly accessible collection of imposing buildings that includes the Neo-Classical to the latest innovations of the post-modern period.

It is the sensitive integration of functional needs and contemporary technological solutions that distinguishes the UP campus and makes each facility and building unique to be able to appreciate.
Learn to collaborate and co-create
13, 23, 43 Things: Create. Share. Network

- Blogging
- Photos & Images – Flickr & mashups
- RSS & Newsreaders
- Social Tagging – del.icio.us, Connotea
- Social Networking – Facebook, Linkedin
- Tagging, Folksonomies & Technorati
- Wikis
- Online Applications & Tools – online office/ design
- Podcasts, Video & Downloadable Audio - YouTube
- Next Generation OPACs
Next generation OPACs

- As times have changed, our OPACs have not kept up. End-users want more from their library catalogs. They want the ability to create mash-ups with other services (e.g., LibraryThing), save catalog data in new and different ways (e.g., Zotero), and much, much easier findability (e.g., through facetted browsing.)
• "If you love books, and love people who love books, LibraryThing is for you. Start by using the service to catalog your book collection: Tag your books by topic, share your catalog with others, and then endlessly browse the titles that they have on their shelves. The utterly book obsessed can add the LibraryThing widget to a blog to show visitors what they have been reading lately."  
PCWorld (February 2006)
- Automatic capture of citation information from web pages
- Flexible notetaking with autosave
- Playlist-like library organization, including saved searches (smart collections) and tags
- Runs right in your web browser
- Free and open source
- Saves records and notes in any language; interface available in more than 20 languages
- Storage of PDFs, files, images, links, and whole web pages
- Fast, as-you-type search through your materials
- Platform for new forms of digital research that can be extended with other web tools and services
- Formatted citation export (over 1100 styles and growing)
- Integration with Microsoft Word and OpenOffice
- Integration with WordPress and other blogging software

Our Future Through Science
Librarian 2.0

• Recognises that the Library is human
• Plans for the users: breaks down barriers and allow users access wherever they are: home, work, commuting, school, or at the library.
• Embraces Web 2.0 tools recognizing how services might be enhanced by the Read/Write web and how new services might be born in a climate of collaboration.
• Controls techno-lust - does not buy/use technology for the sake of technology.

See: http://www.oclc.org/nextspace/002/3.htm
Web 2.0 Librarian cont

• Makes good, yet fast decisions: comfortable with perpetual beta & redesigns for ease of use, user involvement and easily added/re-configured pieces

• Trendspotter:
  • seeks out information and news that may impact future services
  • reads outside the profession and
  • watches for the impact of technology on users and new thinking on business

• Knows how to get content - understands that the future of libraries will be guided by how users access, consume and create content

• Listens to staff and users when planning, tells the stories of successes and failures, learns from both, celebrates those successes, allows staff time to play and learn, and never stops dreaming about the best library services
In the Job Jar of Oxford’s Digital Repositories Research Co-ordinator

- Capture and document researcher's requirements for digital repositories services to handle research data
- Participate actively in the development of an interoperability framework for the federated digital repository at Oxford
- Make recommendations to improve and coordinate the provision of digital repository services for research data
In the Job Jar of Oxford’s Digital Repositories Research Co-ordinator 2

- Initiate and develop collaborations with the different repository activities already occurring to ensure that communication takes place in between them
- Raise awareness at Oxford of the importance and advantages of the active management of research data
- Communicate significant national and international developments in repositories to relevant Oxford stakeholders, in order to stimulate the adoption of best practices
Key question: do you manage a …

- Paper based library which also gives access to electronic resources?

- Electronic resources library which also gives access to paper collections?

- Challenge: perhaps the library should become more like a lab and less like a warehouse! (… with apology to Jim Jacobs – retired data-librarian University of California, San Diego)
“The world we have created today has problems which cannot be solved by thinking the way we thought when we created them.”

-Albert Einstein

… also applicable to librarians
Scientific progress depends on speedy and open access to the full spectra of scientific data and derived products

... remember ...

(Hunter, 2006)
Additional references