“Enabling Success: IT Infrastructure & Repositories”

Andrew Bennett, University of Qld Library

APSR: The Successful Repository
University of Queensland
Brisbane, June 29th 2006
Overview

- A quick retrospective look at UQ experience
- Where does “infrastructure” sit?
- Evolution of Data Storage Requirements
- “Middleware”, Systems and Platform wars
- Staffing Costs & Measurement (statistics)
- Summary & Questions
In the beginning . . .

Early repositories didn’t need a lot of infrastructure: some mud, a sharp stick and a nice dry place to store your “content” was pretty much all you needed – infrastructure was cheap too, when you ran out, you went down to the stream for more!
We started small..

We started with isolated sites that each hosted their own particular type of content .. typically using custom built software and hosted on single-purpose servers

- ePrints
- Theses
- Low resolution images
- Digitised examination papers
And Grew Slowly ...

Our initial philosophy was to keep the different types of content separate. Each new “repository” seemed to need a new operating system, a new development platform and quickly the number of hosts started to multiply.

As the complexity of managing multiple servers and operating systems became more of a cost issue, we started looking at ways of making it more cost-effective.
Technology moved on . . .

By the middle ages, infrastructure had become harder to find: content was chained to desks, in a cloister, available only to a select few.
New Players Emerged …

With the opportunities presented through involvement in projects such as APSR, we were able to look at other products (such as Fedora, DSpace, Greenstone) to become our new repository “platform”

We also started to think more broadly about target content for our “repository” – looking beyond publications
We Entered the Digital Era..

Of course, discovering where your content was had become a lot harder . . . and then there was this thing called Google-Scholar.
Finding our feet …

It became critical to not only understand what content we had, but how it was being used, by whom and where. We started to look at rationalising the number and type of repositories we were running and built some reporting tools to get statistics on their use.

We discovered that we could use the statistics to help recruit more content and to start to demonstrate that there was real value in the work that was being done.
Repositories Everywhere!

So now we have lots of different types of content and some kind of grasp on the number and nature of our institutional repositories .. but there is still all of this “infrastructure stuff” underneath them

No matter what flavour or platform you choose ... many of the IT issues you will encounter are the same
Evolution of Data Storage

- Initial efforts involved storing content on direct attached storage and local file systems on the web servers themselves
- Moving on we started to separate the hosting application and network storage for data
- A radical change was the idea that we might host content in a database, such as MySQL or Oracle but still on a single (or clustered) server
Evolution of Data Storage

- A Storage Area Network (SAN) enabled us to host very large amounts of content online without having to keep everything on one server.
- Backup is done with robotics and multiple instances of the data on remote servers.
- We are looking at the concept of server virtualisation as the next step in reducing hardware infrastructure costs (expected hardware savings might be as much as 15-20%).
# Adding up the pieces

<table>
<thead>
<tr>
<th>Content Type</th>
<th>Current Storage (Tb)</th>
<th>Projected Requirements by end 2006 (Tb)</th>
<th>Projected Requirements by end 2007 (Tb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Collections</td>
<td>0.4</td>
<td>0.6</td>
<td>1.4</td>
</tr>
<tr>
<td>ePrints / pubs</td>
<td>0.3</td>
<td>0.6</td>
<td>1.2</td>
</tr>
<tr>
<td>Digitised Theses</td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Course Materials</td>
<td>1.1</td>
<td>1.9</td>
<td>3.4</td>
</tr>
<tr>
<td>Digitised Exams</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>2.1</strong></td>
<td><strong>3.7</strong></td>
<td><strong>7.1 !!</strong></td>
</tr>
</tbody>
</table>
# Storage Costs

<table>
<thead>
<tr>
<th>Storage Media</th>
<th>Approximate Annual Cost per Gb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Attached Storage (DAS)</td>
<td>$9.65</td>
</tr>
<tr>
<td>Fibre Channel Disk in SAN (FC)</td>
<td>$18.51</td>
</tr>
<tr>
<td>SATA disk in SAN (SD)</td>
<td>$6.33</td>
</tr>
</tbody>
</table>

*Based on (2006) Monash University: The Direct cost of Storage DAS vs SAN*
## Storage Costs (cont)

<table>
<thead>
<tr>
<th></th>
<th>Cost Per GB</th>
<th>Utilisation Rate</th>
<th>True Business Cost Per GB</th>
<th>Example UQL 2006 Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DAS</strong></td>
<td>$9.65</td>
<td>45%</td>
<td>$20.27</td>
<td>$76,798.98</td>
</tr>
<tr>
<td><strong>SAN FC</strong></td>
<td>$18.51</td>
<td>80%</td>
<td>$22.21</td>
<td>$84,149.25</td>
</tr>
<tr>
<td><strong>SAN SATA</strong></td>
<td>$6.33</td>
<td>80%</td>
<td>$7.60</td>
<td>$28,794.88</td>
</tr>
</tbody>
</table>

*Based on (2006) Monash University: The Direct cost of Storage DAS vs SAN*
What Happens When We Run Out of Disk?

Your options become more complicated when the amount of content you want to host exceeds your capacity to store it.

- Buy More Storage
- Hierarchical Storage Management (HSM)
- Throw Out some Content ??
- Distributed Storage Options (SRB, FreeLoader etc.)
Shared Storage via SRB

- The purpose of an SRB data grid is to enable the creation of a collection that is shared between academic institutions
  - Register digital entity into the shared collection
  - Assign owners & access controls
  - Assign descriptive, provenance metadata
  - Manage replication information and interactions with storage systems
    - Unix file systems, Windows file systems, tape archives, …
## Storage Resource Broker

<table>
<thead>
<tr>
<th>Federation Management</th>
<th>Database Abstraction</th>
<th>Storage Repository Abstraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency &amp; Metadata Management / Authorization, Authentication, Audit</td>
<td>Databases - DB2, Oracle, Sybase, Postgres, mySQL, Informix</td>
<td>Archives - Tape, Sam-QFS, DMF, ORB</td>
</tr>
<tr>
<td>Logical Name Space</td>
<td>Latency Management</td>
<td>File Systems Unix, NT, Mac OS</td>
</tr>
<tr>
<td></td>
<td>Data Transport</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Metadata Transport</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **C Library, Java**
- **Unix Shell**
- **Linux I/O**
- **C++**
- **NT Browser, Kepler Actors**
- **DLL / Python, Perl, Windows**
- **HTTP, DSpace, OpenDAP, GridFTP**
- **OAI, WSDL, (WSRF)**

- **ORB**
- **Storage Resource Broker**
- **Linux I/O**
- **C++**
- **DLL / Python, Perl, Windows**
- **HTTP, DSpace, OpenDAP, GridFTP**
- **OAI, WSDL, (WSRF)**

- **Databases - DB2, Oracle, Sybase, Postgres, mySQL, Informix**
- **Archives - Tape, Sam-QFS, DMF, ORB**
- **File Systems Unix, NT, Mac OS**
- **Databases - DB2, Oracle, Sybase, Postgres, mySQL, Informix**

- **THE UNIVERSITY OF QUEENSLAND**
- **AUSTRALIA**
What Good is “Middleware”?  

- Addresses problems of shared access and identity management in distributed, secure environments  
- Helps when you need to know **who** someone is, rather than just their role/attributes?  
- Helps get around issues of inflexible, “hardwired” access policies at the system level  
- eg. Shibboleth, MAMS, Athens, eduPerson . . .  

- **BUT**, you need everyone to play the same game
**Systems and Platform Wars**

*Is it better to have many different custom purpose repository systems or one which can be many things?*

We chose Fedora for its flexibility and the fact that it aligned with our development expertise but at the end of the day, as long as we are all speaking the same language it doesn’t matter.

*If you choose a platform which requires your institution to develop, customise or add code for enhancements, be wary of the human resourcing costs*
FEZ Development @ UQ

Currently, Fez development is being carried out by 2 developers at UQ with contributions from other sites worldwide. Open-source is great but the community has to be self-sustaining when federal funding runs out.

If all you wanted to do was run the repository without further development, a single competent IT person with some expertise in the development tools used could fairly easily manage it.
Who Manages IT?

- Is the “system” managed by someone in your library?
- Is your IR managed by an “IT” person or by someone with specialised metadata skills?
- Options of “Do it yourself” ..vs.. SLA with a central IT service ..or.. Outsource to a 3rd Party?
Measurement

Statistics are probably the key thing that will help drive the idea that your institution is getting value from the investment in you IR and as a sector we also have to be able to make meaningful comparisons (JISC: Interoperable Repository Statistics)

BUT .. any metrics you supply have to be useful to the executive and consistent with other institutional data collection efforts
Summary

- *Infrastructure costs are a real barrier to repository success, make sure you understand them when talking to your executive.*
- *Look for cost-efficiencies in storage, staffing and management of overheads such as identity management*
- *Choose a platform that fits not just your budget but your environment and make sure you integrate with other enterprise systems*
- *Make sure you have measures and metrics to prove the value of your investment, especially when it comes time to ask for more*
Thankyou!!

Questions if we have time?