Date : 01/07/2008

Infrastructure models used by California Digital Library' Preservation Projects

Margaret Low
California Digital Library
Oakland, USA

Meeting: **84. Preservation and Conservation, (PAC), Information Technology, IFLA-CDNL Alliance for Bibliographic Standards (ICABS) and Law Libraries**

Simultaneous Interpretation: Not available

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

The California Digital Library (CDL) manages and operates its own repositories, sharing the development and use of the services with the University of California academic community as well as a wider digital library community. Currently, the main repository models are a digital repository and a web archiving repository. The Digital Preservation Repository (DPR), a self-service repository, is the basic repository model. It has been in production since 2005. The Web Archiving Service (WAS) is the more complex model and also a self-service repository; it is in the final stages of moving into production. A third model, whose sole purpose is to move objects into a repository, is currently being built.

CDL Common Framework (CF)

The repository models are built using the CDL Common Framework (CF), providing a core set of services and supporting systems implemented in Java and J2EE. The architecture follows the Reference Model for an Open Archival Information System (OAIS).

Separation of Functions

Each service within the CF is a well-defined functional component implemented using CDL developed in-house wrappers. The services are independent of each other; and new components unique to a repository can be written and plugged into place. Third party components can be integrated into the CF using the wrappers.

Third Party Standards/Technologies in Common Framework	
ARK	Archival Resource Key
JHOVE	JSTORE/Harvard Object Validation Environment
METS	Encoding and Transmission Standard
MySQL	Relational database
NOID	Nice opaque identifier generator
REST	Representational State Transfer protocol
SOAP	Simple Object Access Protocol
XML	Extensible Markup Language
XSLT	XSL Transformations
XTF	eXtensible Text Framework

This architecture allows for flexible designs specific to the functional needs of repositories.

Separation from Archival Storage
The CF architecture presents CDL with options for its archival storage because the functional services are separated from the storage. The CF uses wrappers to interface between the services and the storage system. All our repositories are currently managed by the Storage Resource Broker (SRB), for their archival storage. SRB is developed by the San Diego Supercomputer Center (SDSC) at the University of California, San Diego. The CF has a wrapper working with SRB's JARGON (a pure java API for programs with a data grid interface) which allows access to the SRB servers. With the changing pace of archival storage technology, the flexibility of the CF architecture allows CDL to move toward different/newer storage equipment.

Repositories Share Code
At the lowest levels, the repositories share the same code built within the common framework. For example, because all of CDL's repositories use SRB for its archival storage, the java storage code is the same. In addition, the repositories share the same code for security, administrative ownership and rights management of the digital objects. Obviously, this reduces the effort of rewriting the code and any enhancements or upgrades can be inherited by all the repositories.

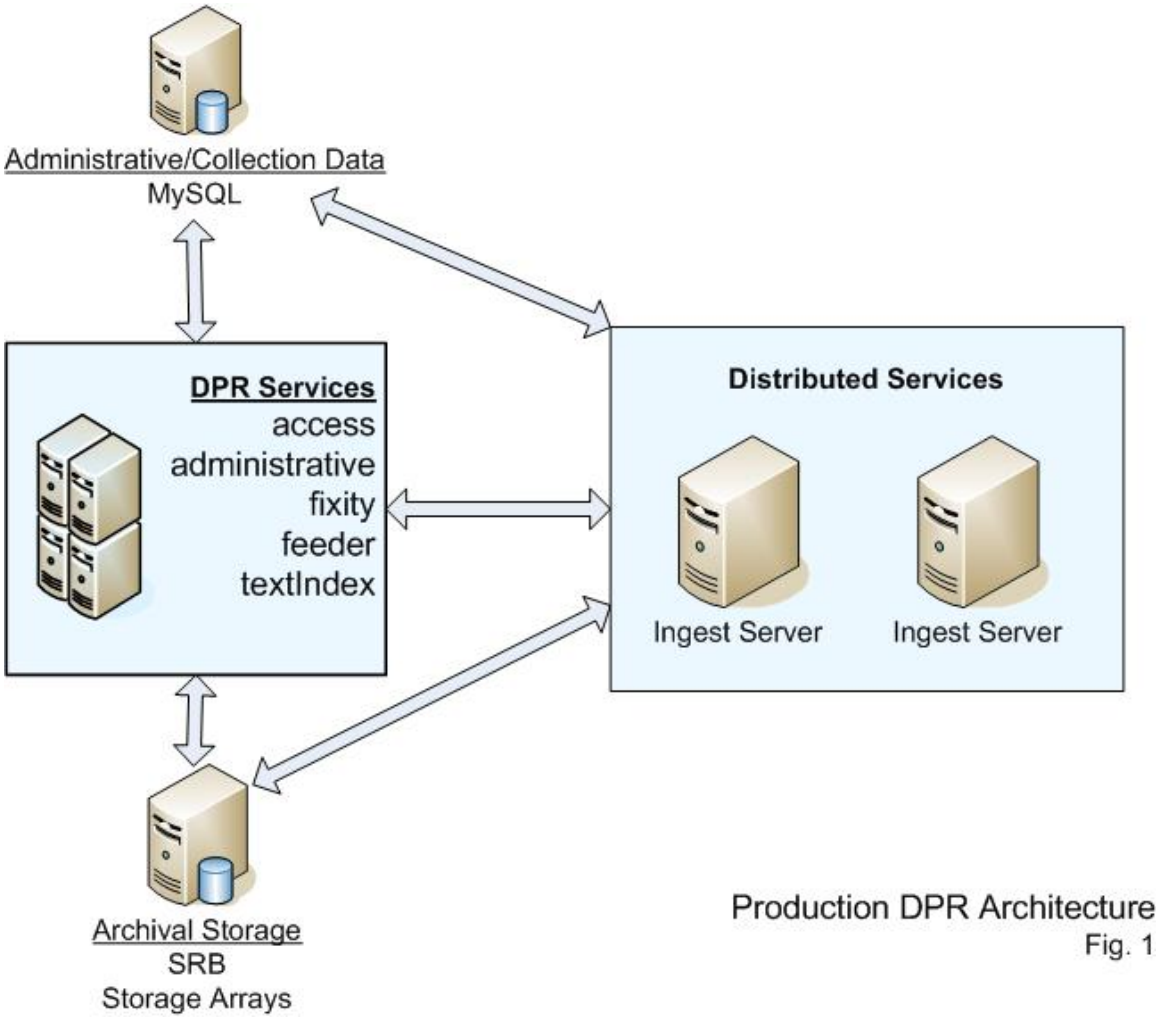
Distributed Architecture for Scalability
Because of the separation of the services, each repository has the option of being built with a distributed architecture. The services can reside in the same server on a machine; or reside in multiple servers on many machines. Resource intensive services such as ingest, can have many instances across multiple servers, distributing the work onto the machines. Each instance of a repository has the ability to add or remove servers to enhance the throughput as needed.

Basic Model - Digital Preservation Repository (DPR)

The DPR is a self-service repository available to the digital library communities in California. This community includes all ten of the University of California Campuses; and public libraries, historical societies and state colleges under the Library Services and Technology Act (LSTA) federal grant.

The DPR was the first instance of the CF. The core services include authentication and administrative functions with access and ingest of the digital objects. There is limited web access to the objects.

There are three test versions of the DPR; two used for development and one used for staging. Since the DPR is a self-service model, the stage version is available for the users to test their digital objects for validation and ingest. These instances are built on different machines going to separate archival storage. Each instance has its core component servers on one machine with the archival storage on a different machine.



Production DPR Architecture
Fig. 1

The production DPR also has its core component servers residing on one machine with the archival storage server on a different machine (see Fig. 1). But for better ingest throughput, there are multiple ingest servers spread across several machines. If there is a need to increase loading, additional ingest servers are built into the production instance.

Complex Model - Web Arching Service (WAS)

Working with the International Internet Preservation Consortium (IIPC), the WAS web archiving repository is in its final phases of testing involving curators from across the nation. Production is scheduled to go live later this year.

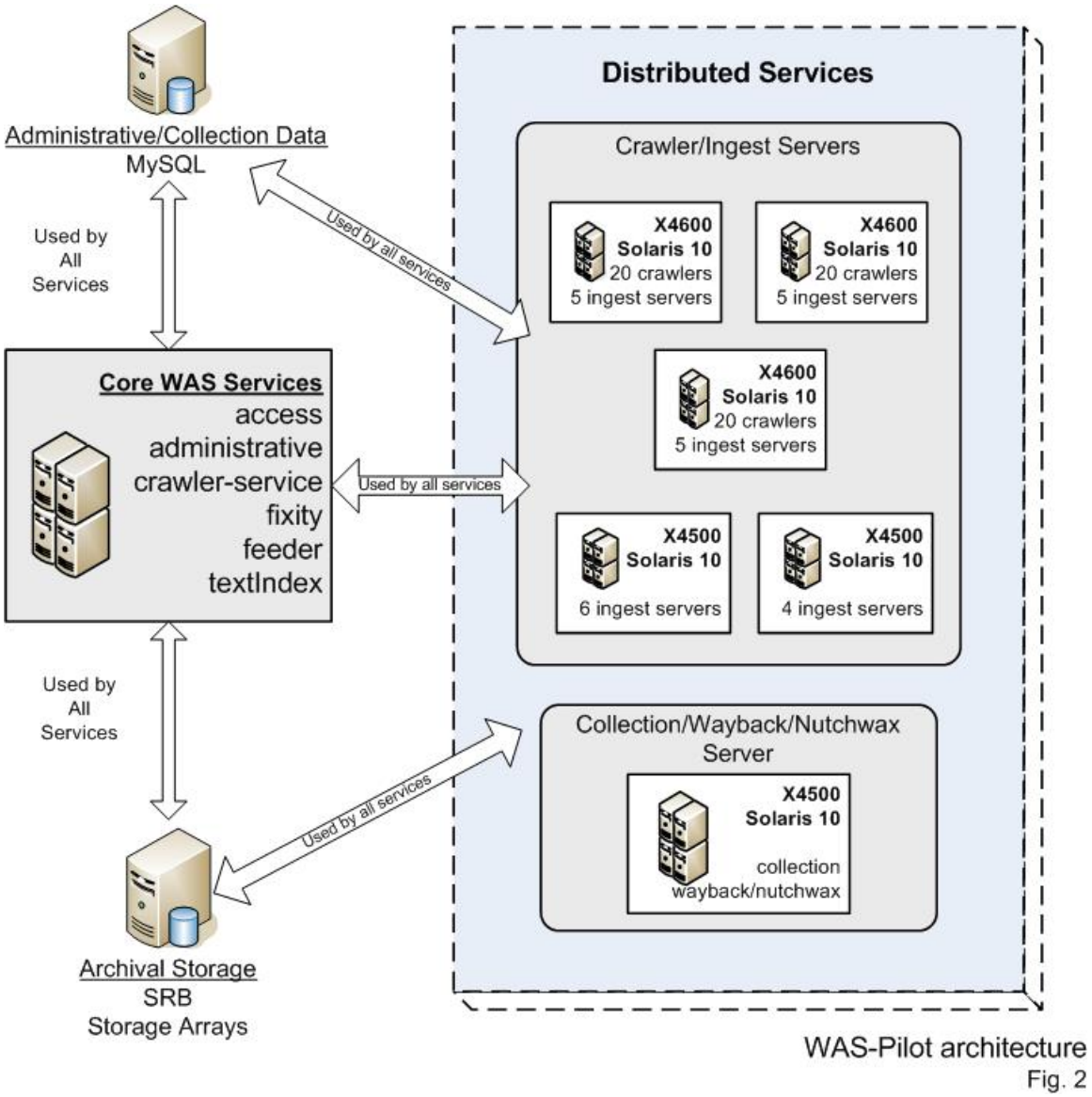
The WAS repository is more complicated CF model, using the core CF services as well as incorporating services developed by the Internet Archive with the Nordic National Libraries. For web crawling, it uses an open-source, extensible, web-scale, archiving quality web crawler, Heritrix. For access and finding aids, it incorporates NutchWAX (Nutch with Web Archive eXtensions) a tool for indexing and searching web archives.

Additional Third Party Standards/Technologies for WAS	
Heritrix	Web crawler
NutchWAX	Searches web archive collections

There are two instances of this model, one development and one pilot (staging) instance. The development model, used for CDL testing, is built on one machine with separated archival storage. It has the minimum number of services to allow for testing of new code and features.

The pilot instance is used by the curators of the project to crawl and archive web sites. It is built on seven different machines, load balancing the resource intensive services for throughput. For this instance there were 31 crawlers and 20 ingest servers (see Fig. 2). Again, as with the DPR, each of these instances has its own separate archival storage.

The production instance is due to go live in mid-2008 and is built on nine separate machines following the pilot model with its own instance of the archival storage.

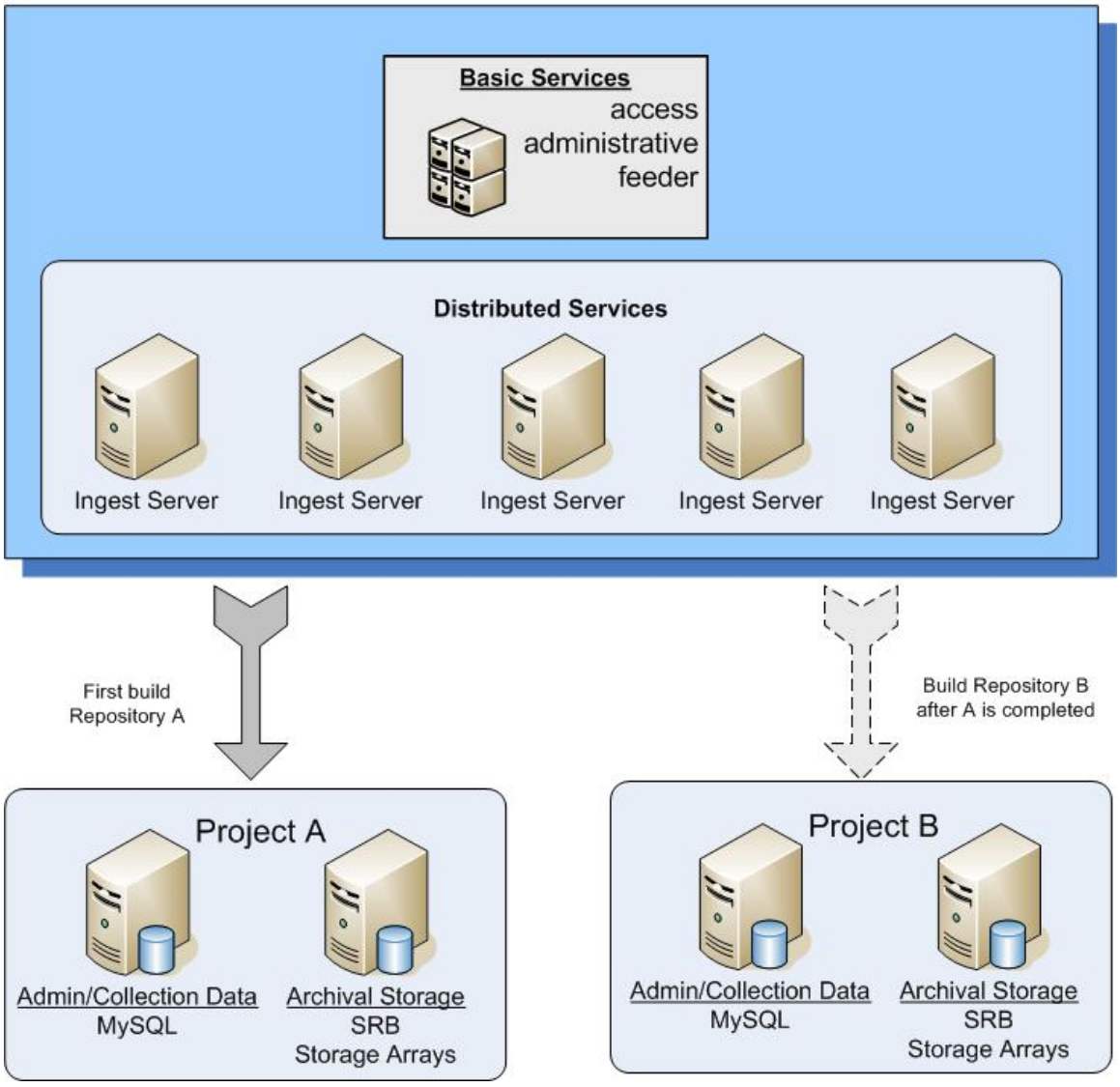


WAS-Pilot architecture
Fig. 2

Flexible Scalable Model

CDL has end-in-time (or end-in-sight) digital preservation projects where there is definite group of objects to ingest. Once the objects are ingested the project is completed. The model (see Fig. 3) for these projects is a flexible scalable model where available resources are used for that purpose. In addition, this model can be re-architected for other end-in-time projects. It has the core services; administrative ownership, access and ingest. The number of ingest servers would be related to the availability of the equipment and the deadline required to complete the task. When the project is completed, the instance could be re-architected for other projects.

Flexible Scalable Model
Fig. 3



Conclusion

As these digital repositories mature, they represent many challenges for CDL. There is the challenge to work with the local, national and worldwide digital library community, keeping abreast and involved with the goals of digital preservation; the challenge to keep up with the latest technology, choosing a course that moves ahead without falling into the 'latest and newest' trap; the challenge of managing the archival storage keeping it reliable and secure and finally, the challenge is to choose technical and organizational models that will ensure that the current California Digital Library repositories, as well as future ones, will be available for the next generation. The flexibility of the CF is helping CDL to meet the challenges of the ever changing digital library community.