Opening aDORe

Ryan Chute [1], Luydimilla Balakireva [2], Herbert Van de Sompel [3]

Digital Library Research & Prototyping Team

Research Library

Los Alamos National Laboratory

[1] rchute@lanl.gov
[2] ludab@lanl.gov
[3] herbertv@lanl.gov

The aDORe Project: Background

- Initial motivation:
  - Severe deficiencies in the information discovery environment developed for the LANL Research Library:
    - Metadata-centric: descriptive metadata records first class citizens; actual digital assets auxiliary data.
    - Tens of millions of digital assets stored as files in file system.
    - Tight integration between application content collection and discovery, preventing other applications from leveraging the rich content base.
  - Obvious solution:
    - Replace metadata-centric approach by compound object approach.
    - Bundle digital assets into storage containers that dramatically reduce the amount of files in a file system.
    - Cleanly separate storage repository from applications that leverage the stored assets by providing necessary machine interfaces.

- Implementation of the obvious solution led to the aDORe R&D project 2003-2007

The aDORe Project: Major Drivers

- Concrete need to design and implement a solution to ingest, store, access the vast and growing collection of the LANL Research Library.
  - ○ Scale, scale, scale!
  - ○ Existing open source solutions (at that time) did not meet our scale requirements
    - - e.g. static binding of disseminators to objects in Fedora.
- Interest in repository interoperability, cf. involvement in OAI-PMH, NISO OpenURL, OAI-ORE
- Interest in digital preservation, cf. National Digital Information Infrastructure and Preservation Program (NDIIP) funding
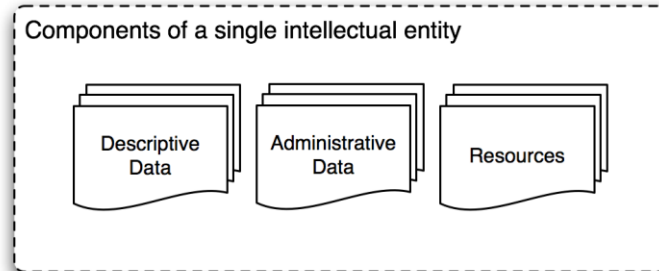
Los Alamos
NATIONAL LABORATORY
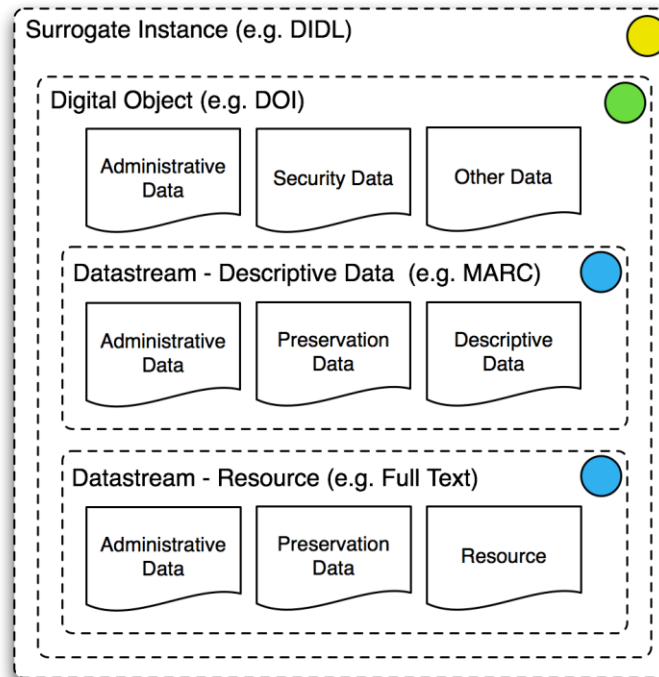
The aDORe Project: Major Design Principles

- Leverage existing standards and technologies to make development and migration more straightforward.
    - Read: Laziness as a strategy
- Use a distributed, component based approach to meet challenges of scale.
- Use Digital Objects, Datastreams, and Surrogate abstractions to characterize content.
- Facilitate a uniform manner for client applications to discover and access content objects available in a group of distributed repositories.
- Provide single repository behavior for a group of distributed repositories.

The aDORe Federation: Content Objects

Components of a single intellectual entity
- Descriptive Data
- Administrative Data
- Resources

Surrogate Instance (e.g. DIDL) — Surrogate Identifier

Digital Object (e.g. DOI) — Content Identifier
- Administrative Data
- Security Data
- Other Data

Datastream - Descriptive Data (e.g. MARC) — Resource Identifier
- Administrative Data
- Preservation Data
- Descriptive Data

Datastream - Resource (e.g. Full Text) — Resource Identifier
- Administrative Data
- Preservation Data
- Resource

Legend:
- Surrogate Identifier (yellow)
- Content Identifier (green)
- Resource Identifier (blue)

Los Alamos
NATIONAL LABORATORY

The aDORe Federation: Content Objects

Content is characterized into three types of Content Objects:

- **Digital Objects** - an identified aggregation of one or more Datastreams and properties pertaining to Datastreams and to the aggregation itself.

- **Datastreams** - a retrievable bitstream, of any media type, made available by a repository to the federation.

- **Surrogates** - the serialization of a Digital Object into a machine-readable representation that is made accessible by a repository.
  - Supports Multiple Complex Digital Object Serialization Formats (e.g. MPEG-21 DIDL, METS, ORE Atom/XML, ORE RDF/XML)
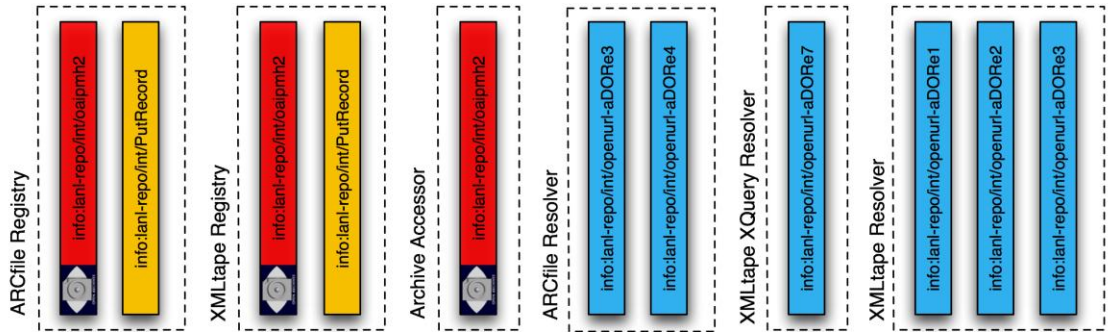
Los Alamos
NATIONAL LABORATORY

The aDORe Federation: Architecture

A 3-Tier architecture for the federation of distributed repositories:
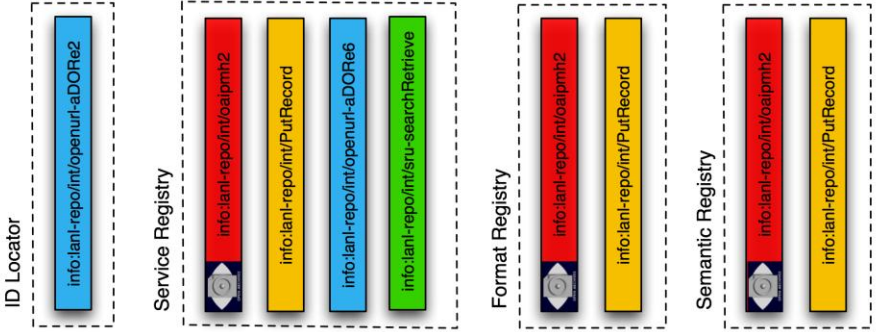
- Tier 1: the aDORe repositories
  - Networked systems that <u>host digital object content</u> and <u>that make that content accessible</u> by exposing core service interfaces.
  - Currently XMLtapes and ARCfiles (aDORe Archive)
  - Other Content Management Systems can be turned into an aDORe repository by implementing the core service interfaces.

- Tier 2: the aDORe federation management
  - Networked systems that <u>facilitate presenting the aDORe repositories as a single logical repository;</u> these federation components expose core service interfaces to allow access to their content.
  - Federation components are: Identifier Locator, Service Registry, Format Registry, Semantic Registry

- Tier 3: the aDORe front-ends
  - Networked systems that <u>make digital object content hosted in the multitude of physical aDORe repositories accessible</u> by exposing core services interfaces that present those aDORe repositories as a single logical repository
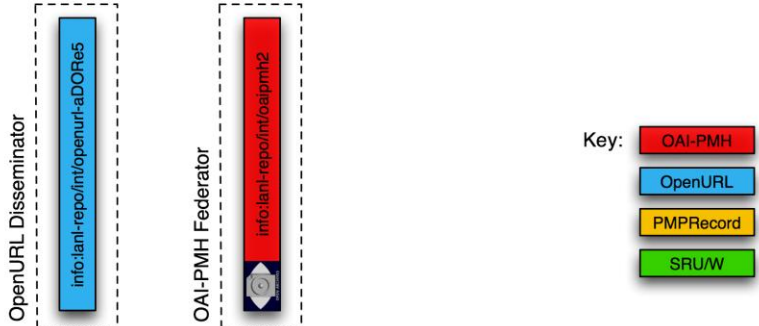 - aDORe front-ends are: OAI-PMH Federator, OpenURL Resolver

aDORe Federation

The aDORe Federation software

- Available today at:

  http://african.lanl.gov/aDORe/projects/adoreFederation

- Released under GNU LGPL Open Source Software License
- This is a major update to the aDORe Archive:
    - Updates the Tier-1 aDORe Archive
    - Implements the 3 Tiers of the architecture instead of only Tier-1

- Use Cases
    - Large collections of relatively stable objects
    - A plug-in storage component for Institutional Repository and Archive solutions
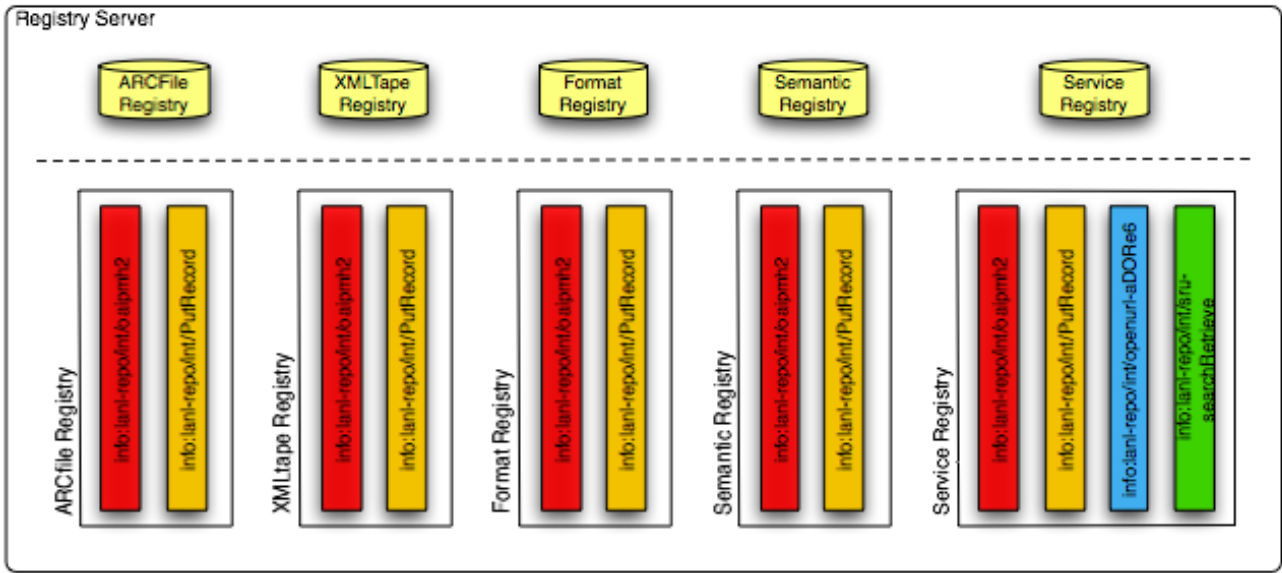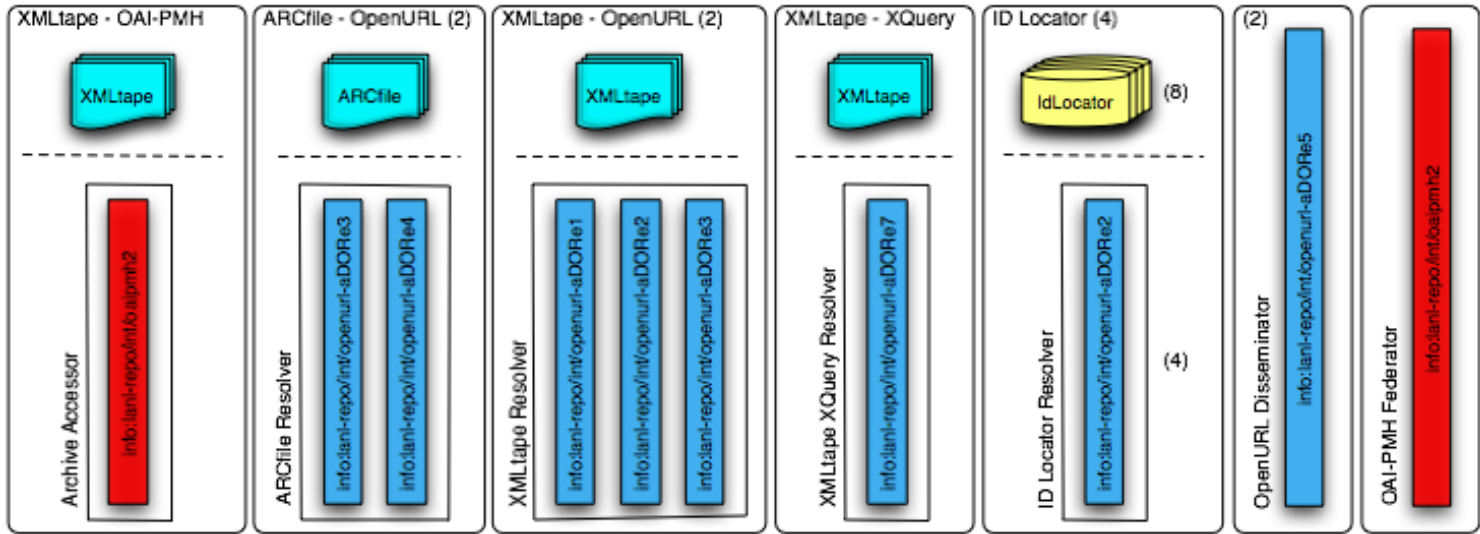
The aDORe Archive @ LANL, August 23, 2008

- In production at LANL Research Library for over 1 year

- 90,000,000 Compound Digital Objects
- 216,000,000 Stored bitstreams
- ~ 10,000 autonomous repositories:
    - ~ 4,500 XMLtapes: XML serializations of Digital Objects
    - ~ 5,500 ARCfiles: bitstreams
- > 617,000,000 identifiers

The aDORe Project: Current / Future Work

- djatoka - An Open Source JPEG 2000 Image Server
  - Compression of JPEG 2000 files using the Kakadu JPEG 2000 Library
  - Dynamic extraction of resolutions and Regions from JPEG 2000 files;
  - Support for a rich set of input/output formats (e.g. BMP, GIF, JPG, PNG, PNM, TIF, JPEG 2000).
  - Extensible interfaces to request image services and manipulations (e.g. watermarking);
  - A rich service framework, based on the OCLC OpenURL Resolver, to facilitate the transfer of service parameters via an OpenURL compliant HTTP GET request.
  - Pluggable as an aDORe Disseminator Service
  - Release: Mid-Sept. 2008
- adore-searcher - A SOLR-based search engine for aDORe repositories
  - Currently used in production aDORe instance (> 90M bib records)
  - Release: ?