Blending open source and the cloud to build an discipline-specific repository

Al Cornish
Washington State University

XTF – open source digital content platform

- Search and display technology created and maintained by the California Digital Library
- Indexes digital content: XML (including finding aids) PDFs; metadata associated with digital objects, such as JPEGs….

XTF
eXtensible Text Framework

XTF – open source digital content platform

- Java-based framework; uses a Java servlet container, such as Apache Tomcat, to serve an XTF site

- Configuration and customization is accomplished through XSLT programming

XTF
eXtensible Text Framework

more about XTF

- Search – based upon Apache Lucene text search engine library
- Display – based upon Saxon XSLT library
- Active developers/users groups through Google Groups

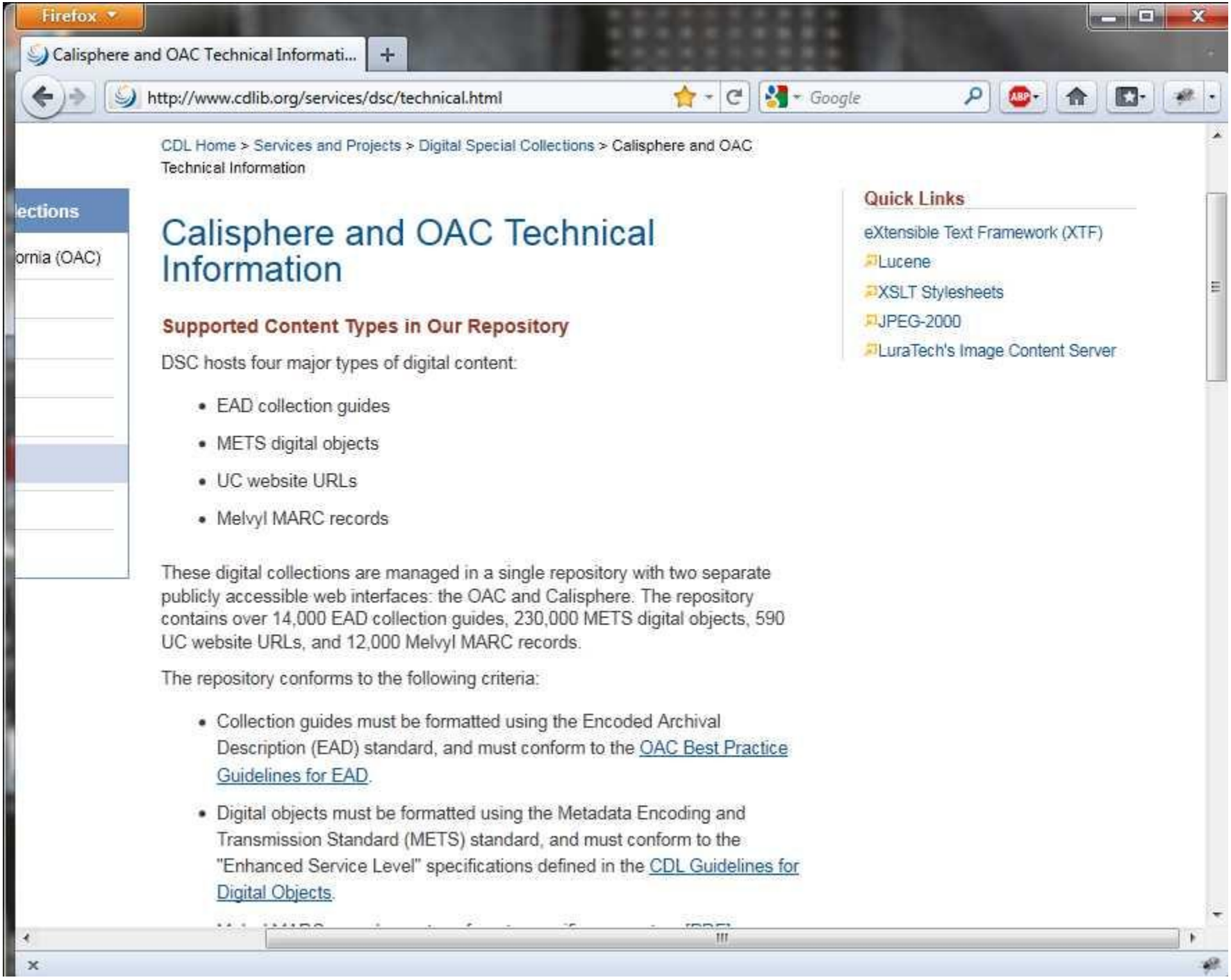
XTF
eXtensible Text Framework

Amazon Web Services EC2

- Enables you to run Linux and Windows servers in one of Amazon's data centers
- A set of persistent resources that is available to support online services
  - Storage volumes, IP addresses

could this work at the institutional level??

- XTF is being used to support large-scale, online digital collections
  - California Digital Library's Calisphere, Mark Twain Project…
  - A number of non-CDL XTF implementations, including the Encyclopedia of Chicago and institutional finding aid databases
  - The online tutorials for XTF suggest how it could be used in an institutional context



CDL Home > Services and Projects > Digital Special Collections > Calisphere and OAC Technical Information

ections

ornia (OAC)

Calisphere and OAC Technical Information

Supported Content Types in Our Repository

DSC hosts four major types of digital content:

- EAD collection guides
- METS digital objects
- UC website URLs
- Melvyl MARC records

These digital collections are managed in a single repository with two separate publicly accessible web interfaces: the OAC and Calisphere. The repository contains over 14,000 EAD collection guides, 230,000 METS digital objects, 590 UC website URLs, and 12,000 Melvyl MARC records.

The repository conforms to the following criteria:

- Collection guides must be formatted using the Encoded Archival Description (EAD) standard, and must conform to the OAC Best Practice Guidelines for EAD.
- Digital objects must be formatted using the Metadata Encoding and Transmission Standard (METS) standard, and must conform to the "Enhanced Service Level" specifications defined in the CDL Guidelines for Digital Objects.

Quick Links

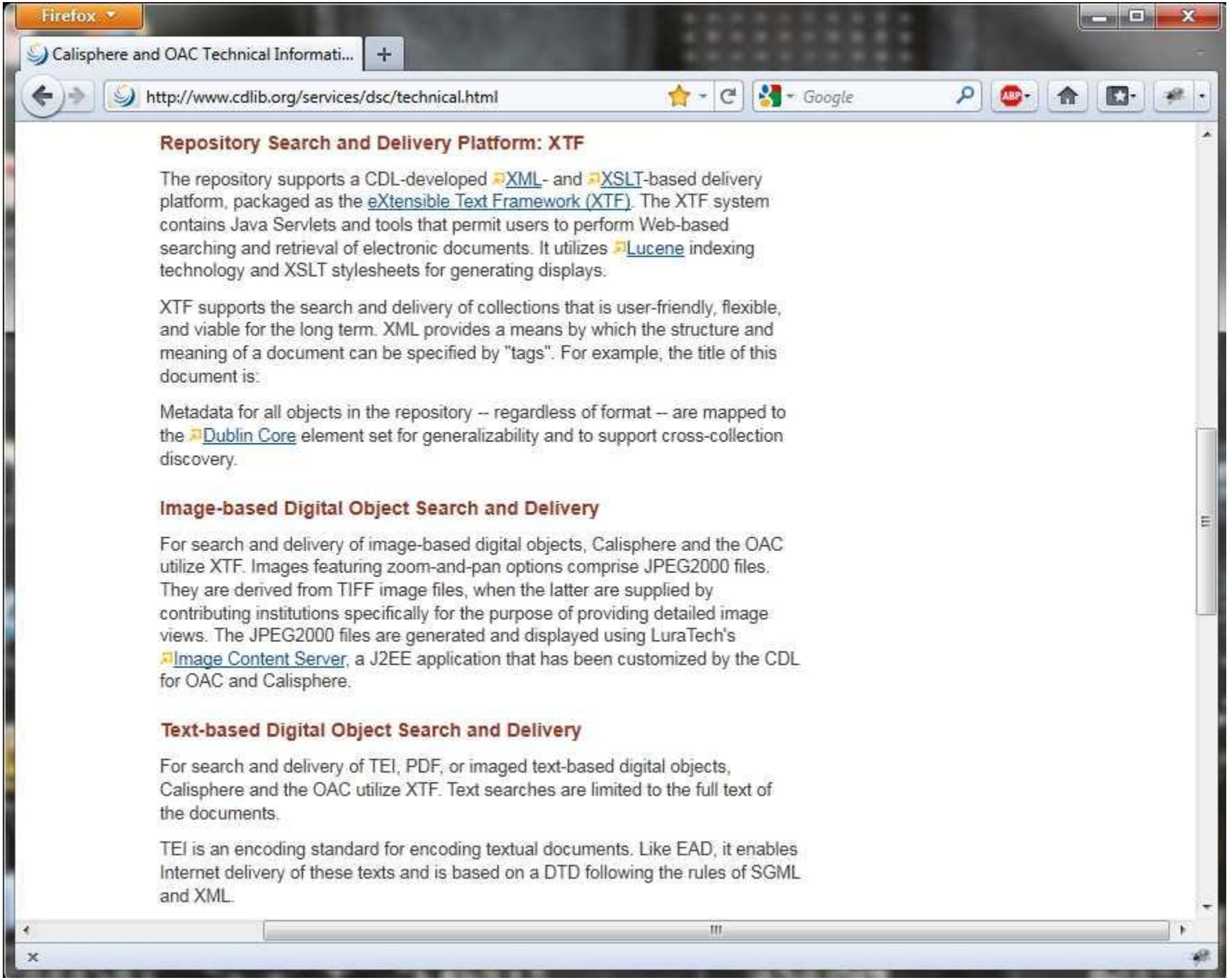
eXtensible Text Framework (XTF)

Lucene

XSLT Stylesheets

JPEG-2000

LuraTech's Image Content Server



Repository Search and Delivery Platform: XTF

The repository supports a CDL-developed XML- and XSLT-based delivery platform, packaged as the eXtensible Text Framework (XTF). The XTF system contains Java Servlets and tools that permit users to perform Web-based searching and retrieval of electronic documents. It utilizes Lucene indexing technology and XSLT stylesheets for generating displays.

XTF supports the search and delivery of collections that is user-friendly, flexible, and viable for the long term. XML provides a means by which the structure and meaning of a document can be specified by "tags". For example, the title of this document is:

Metadata for all objects in the repository -- regardless of format -- are mapped to the Dublin Core element set for generalizability and to support cross-collection discovery.

Image-based Digital Object Search and Delivery

For search and delivery of image-based digital objects, Calisphere and the OAC utilize XTF. Images featuring zoom-and-pan options comprise JPEG2000 files. They are derived from TIFF image files, when the latter are supplied by contributing institutions specifically for the purpose of providing detailed image views. The JPEG2000 files are generated and displayed using LuraTech's Image Content Server, a J2EE application that has been customized by the CDL for OAC and Calisphere.

Text-based Digital Object Search and Delivery

For search and delivery of TEI, PDF, or imaged text-based digital objects, Calisphere and the OAC utilize XTF. Text searches are limited to the full text of the documents.

TEI is an encoding standard for encoding textual documents. Like EAD, it enables Internet delivery of these texts and is based on a DTD following the rules of SGML and XML.

could this work at the institutional level??

- Downside:
  - XTF does not include ingest and administrative modules
  - Use of Amazon Web Services EC2 can pose challenges, in terms of setting up payment

Server hosting costs

	Payment
EC2 Reserved Instance Linux OS small instance, 1 year	$227.50
January 1 - January 31	$25.49
February 1 - February 28	$23.85
March 1 - March 31	$25.63
April 1 - April 30	$15.98
May 1 – May 31	$28.23
GoDaddy.com domain name, 1 year	$27.00
2011 total cost estimate	$542.58

amazon web services™