From Creation to Dissemination
A Case Study in the Library of Congress’s use of Open Source Software

DLF Spring Forum
Corey Keith
ckeith@loc.gov
Introduction

- What is OSS
- Why OSS
- OSS used at LC

- XML is a favorite tool
Purpose

- Share information on available tools
  - Especially those outside the library community
- Promote using OSS
- Promote developing OSS
Open Source Initiative

- Open source promotes software reliability and quality by supporting independent peer review and rapid evolution of source code.
What is OSS?

- Software with source code attached
- Access to underlying code
- Derivative works
  - Add functionality
  - Enhance functionality
  - Remove functionality
- Redistribution (Gotcha)
Why OSS?

- Low experimentation cost
  - Technologies
    - Rapid evolution
    - Move to commercial solution if needed
- Not profit driven community
What types of OSS?

- **Low Level**
  - Programming Libraries
  - Databases
  - Languages
  - Operating Systems
  - Servers
  - Internet Infrastructure

- **High Level**
Selecting OSS for use

- Size
  - Active developers
- Alive/Dead projects
- Reliable Support
- Sponsored Projects
  - Apache Foundation
- Licenses
Benefits of OSS

- Peer reviewed
- Less defects
  - “Many eyes makes all bugs shallow”
- Increased security
Standards

- Promote standards through software development
- Collaboration by development
Benefits for Software Producers

- Development speed
  - Interesting projects attract developers
  - Less likely to switch once started
- Lower overhead
- Smaller institutions can handle larger projects
- If you release OSS, hire from OSS community
Benefits for Software Consumers

- No vendor lock in
- License maintenance
- Own vs rent
Costs for Producers

- Supporting the project
  - Goal: Self sustaining
- Preparing for open source distribution
Costs for Consumers

- Finding technical support
- Services
  - Free software
  - $$$ Services
- Possible increased outsourcing costs
LC Case Study

- I Hear America Singing
- Digital Production
  - METS Making
- Dissemination
  - METS Transformation
- Project Management
Philosophies

- Get data to XML as early as possible
- Flexible manipulation of XML with XSLT
- Aggregation of XML from variety of sources
- Pipelined XSLT
Digital Production

- Aggregate from different sources
  - MySQL DB Descriptive Data
  - Filesystem for Structure
- Apache Cocoon
Apache Cocoon

- XML Glue
- XML Pipeline
- XML Publishing Framework
- Separation of content and style
Cocoon Pipeline

- Generate XML
  - From files (or URLs)
  - From databases
  - From code (XSP, JSP, PHP)
  - From MARC21 records
- Transform XML
  - XSLT
- Serialize
  - XML, HTML, PDF, SVG, JPEG
- Caching, Logging
- Pass parameters to XSLT’s
- Maps URL to Generate/Transform/Serialize
- Command line interface
Dissemination

- Apache – Web Server
- Tomcat – Application Server
- Cocoon – XML Transformation
- Lucene – Searching/Indexing
Apache Lucene

- Text Indexing API
- Flexible
- Fast
- Hierarchical
- Cross-collection
- Depends how you build indexes
Project Management

- CVS
- Bugzilla
- Ant
CVS

- Version Control
- Manage multiple document versions
  - Schemas
  - Documentation
  - Code
  - Libraries
- Multiple developer projects
Bugzilla

- Web-based defect tracking
- Avoids flurry of emails
- Products – Components
- Accountability
- Reporting
Ant

- Automated build tool
- Packaging
- Deployment
What you can do to support OSS

- Demand (well documented) source code when outsourcing
  - Release this code
- Release code your internal projects
Why this is hard

- Clean-up code
  - Embarrassing
- Documentation
- Preparing for release
- Fears of supporting the code
- Hosting
- Marketing
- Generalizing for others use
Summary

- OSS Software
  - Benefits
  - Costs
- Uses at LC
  - I Hear America Singing
- Questions