Building an open platform across diverse content and technologies

Stephen Davison, Betsy Coles, R. S. Doiel, Tommy Keswick, and Thomas Morrell

California Institute of Technology Library
Pasadena, CA, USA
An open platform across diverse content and technologies...

**Themes**
- Constraints and Opportunities
- Metadata management and expression
- Data flow and workflow
- Connections: APIs and identifiers

**Outline**
- Introduction
  - Institutional context
  - Repository diversity
- Seed problems
- Tools
- Future directions
- Principles/lessons learned
An open platform across diverse content and technologies...

**Challenges**

Working with limited resources (human, financial)
Matching needs across systems with staff skills
Maintaining multiple workflows and metadata standards

**Responses**

Focused activities
Staff specialization
Repository specialization
  - Institutional repository
  - Digital Library
  - Research data
  - Born digital
Our choices

- Institutional repository: EPrints
- Digital Library: Islandora (Drupal, Fedora Commons)
- Research Data repository: Invenio (TIND RDM)
- Born digital collections: ArchivesSpace, ePADD, etc.
“Building at the edge” : the seed problems ...

**EPrints**
- List of most recently published articles
- “Expensive” query in EPrints
- Decided not to develop plugin but to leverage API and create a data “feed”

**Caltech Archives**
- Migration from ColdFusion & FileMaker Pro to ArchivesSpace
- Needed to replicate existing website functionality using AS data
- Feed of AS data drives website and populates generic “feeds”
Command line tools

- **dataset** - a JSON document manager, on disc or in S3 storage
- **datatools** - utilities for working with JSON, XLS and CSV data
Middleware and systems integration

- **cait** - A set of utilities that augment the ArchivesSpace API
- **ep** - EPrints REST API harvest and client tools
- **caltechdata_feeds** - Use Invenio Read API to harvest metadata
- **OAI-PMH** - Will be used to harvest Islandora
Web applications

- **mkpage** - lightweight tool for generating static web pages
  - [feeds.library.caltech.edu](http://feeds.library.caltech.edu)
  - Content for [archives.caltech.edu](http://archives.caltech.edu)
- **library.caltech.edu** runs on Drupal and pulls data from Feeds
External Data

- Batch update tools
  - Repository management tools and widgets (e.g. type ahead data entry)

- ArchivesSpace
- EPrints
- Invenio (TIND)
- Islandora

- ORCID
- CrossRef
- FundRef

- Feeds
- Library Web Site
- Archives Web Site

- Amazon S3

- IIIF

- ot – Collect info from ORCID API
Backup and Preservation

- Batch update tools
- Repository management tools and widgets (e.g. type ahead data entry)

- ArchivesSpace
- EPrints
- Invenio (TIND)
- Islandora

- Harvest content for preservation
- eprints_bagit
- caltechdata_read
- islandora_bagit

- Dataset
- dsindexer
- dsfind
- dsws

- Amazon S3
- IIIF

- Preservation/Backup Storage

Caltech Library
Welcome to Caltech Library's aggregated feeds

Content is organized around

- **Recent Articles** holds recent articles from CaltechAUTHORS
  - formats available: JSON, HTML include, BibTeX, RSS
- **Recent Publications** holds recent publications from CaltechAUTHORS
  - formats available: JSON, HTML include, BibTeX, RSS
- **Affiliation** holds a list of publications by CaltechAUTHORS group
- **Person** (experimental) listed by ORCID ID, each subdirectory containing publications, articles and recent feeds

About the data in the feeds

Currently we are generating feeds based on the public contents of CaltechAUTHORS. Feeds are provided in the following formats

- **HTML** with the file extension of `.html`
- **HTML Include** (an HTML fragment suitable for including in another website) `.include`
- **BibTeX** with the file extension of `.bib`
- **JSON** with the file extension of `.json`
- **RSS 2** with file extension of `.rss`

If you find a web page with the content you're interested changing the file extension in the URL from `.html` to the format you want is usually all it takes.
Exposing Services to Users

- Batch update tools
- Repository management tools and widgets (e.g. type ahead data entry)
- ArchivesSpace
- EPrints
- Invenio (TIND)
- Islandora

Provide content to various audiences
Library website as starting point for all
Website can display feeds from many sources

Amazon S3

ORCID
CrossRef
FundRef

Feeds
Library Web Site
Archives Web Site

Caltech Library
Future Work: Data Integration

- Batch update tools
- Repository management tools and widgets (e.g. type ahead data entry)

- ArchivesSpace
- EPrints
- Invenio (TIND)
- Islandora

- Connect publication and data records by “IsSupplement” tags
- Links added in one repository are automatically reflected in the other

- ORCID
- CrossRef
- FundRef
- Feeds
- Library Web Site
- Archives Web Site
- Amazon S3
- IIIF
Future Work: Adding Metadata

- Standardize funding information and add identifiers
- Collect and Display CrossRef Event Data for items in repositories
Future Work: Author Identifiers

- Identify and centralize individuals in all repositories
- Determine Caltech affiliation
- Update other repositories with author identifiers
Prefer API over direct DB access
Develop small; iterate frequently

Prefer API over direct DB access
Batch update tools
Repository management tools and widgets (e.g. type ahead data entry)

Keep structures simple

Prefer API over direct DB access

Develop small; iterate frequently
Data flows in one direction

Preservation/Backup

Storage

ArchivesSpace

Invenio (TIND)

Islandora

EPrints

Batch update tools

Repository management tools and widgets (e.g. type ahead data entry)

External data sources

ORCID

CrossRef

FundRef

Feeds

Library Web Site

Archives Web Site

Amazon S3

IIIF

Caltech Library

Caltechdata, read

caltechdata, bagit

epirins, bagit

caltechdata, bagit

OAI-PMH

mkbag

CAI/T
Develop at the edges

Keep structures simple

Prefer API over direct DB access

Develop small; iterate frequently

ArchivesSpace
EPrints
Invenio (TIND)
Islandora
Develop at the edges

Ongoing harvesting, rather than one-off migrations

Keep structures simple

Prefer API over direct DB access

Develop small; iterate frequently
In the abstract...

- Archives management
- Publications repository
- Data repository
- Digital object repository
- External data source
- External data source
- External data source
- Lightweight tools
- Storage
- Feeds
- Web publication
Users and use cases

Archives management
Publications repository
Data repository
Digital object repository
External data source
External data source
External data source

Lightweight tools

Researchers/groups
Collaborators
Harvesters/aggregators
Analysts
Identity management
Compliance assessment
Value assessment/statistics
IR/Archives integration

Storage
Why not Hydra/Samvera [or ...]?  

There will always be many specialized systems 
Systems will come and go 
Migration will be inevitable 
Continuity at the core; change only when necessary 
Continuous change “at the edges”

Concentrate on building user-oriented tools, services, feeds, web sites
Thank you

Stephen Davison
sdavison@caltech.edu

Betsy Coles
bcoles@caltech.edu

R. S. Doiel
rsdoiel@caltech.edu

Tommy Keswick
tkeswick@caltech.edu

Thomas Morrell
tmorrell@caltech.edu

http://feeds.library.caltech.edu
https://github.com/caltechlibrary/dataset
https://github.com/caltechlibrary/dataset-demo