Community Source meets Open Source: An inspired approach to collaborative funding

Allan Bell, University of British Columbia
Alex Garnett, Simon Fraser University
Carla Graebner, Simon Fraser University
Geoff Harder, University of Alberta

IASSIST 2014
• Community Source vs Open Source
• Our communities
• History of collaboration
• Common infrastructure
• Our projects
Community Source + Open Source = Collaborative Open Source

Community Source
- Governance structure
- Formal
- Budgeted
- “Social Entrepreneurship”

Open Source
- Ad hoc
- Agile
- Faster
- Budget? What budget?
Regional → Provincial → Local → Regional

- National
- Provincial
- Local
- Regional
Regional → Provincial → Local → Regional

- National
- Provincial
- Local
- Regional
SFU RaDaR

To undertake a ‘Proof of Concept’ project which will identify the technological infrastructure, policies and procedures, and costs associated with the development of a Research Data Management Repository for research data to ensure its security and integrity over time
Our Collaboration

Unique competencies / contributions / expertise

• Canadian Association of Research Libraries
• Council of Prairie and Pacific University Libraries Digital Preservation Working Group*
• COPPUL Private LOCKSS Network
• Canadian Government Information PLN
• British Columbia Research Libraries Group Government Documents Digitization
• ABACUS
• CARL -> ARC
Contributions ‘in kind’

Identify local expertise

- Identify local expertise
- Develop partnerships
- Choose the vendor/developer
- Communicate objectives, parameters and time frame
Laying the Foundation

Community Source Opportunities

- Institutional priorities
- Existing resources
- Potential partnerships

Open Source Solutions

- Community oriented
- Responsive to funded projects
- Made available to the community under creative commons licensing upon completion
- Collaborative continuum
Open Source Solutions

Archivematica
- Simon Fraser University Library: Research Data Repository [SFU RaDaR]
- Simon Fraser University Archives
- University of British Columbia: Digital preservation program
- University of Alberta: Canadian Polar Data Network
- University of Saskatchewan

Islandora
- Simon Fraser University Library RaDar
- University of British Columbia
- University of Saskatchewan
“I love it when a plan comes together” *

• Serendipitous

• Canada-wide grant which would’ve provided national support for data management failed two years prior, but raised awareness

• Strong leadership from SFU Library Dean and VP Research secured a local grant for our pilot project

• All three institutions had money(!) they wanted to spend in similar areas(!!) at the same time(!!!)

*John “Hannibal” Smith, leader, A-Team
Multi-part open source repository stack

Main Components:

**Islandora**, developed by DiscoveryGarden in PEI

- I think we’re reaching a point where we all know what this is! Credit to USask for plugining it to...

**Archivematica**, developed by Artefactual in BC

- This is the one component we all seem to share across each of our repository implementations (hat’s off to Artefactual for that!) and this is the focus of our collaboration.
Two separate but related customizations:

- **Automated Ingest** of multiple files via a scriptable workflow API.

- A **Storage API** to allow Archivematica to use non-filesystem storage (e.g. NFS mounts or Fedora Commons) for objects.
Automated Ingest

- Archivematica is designed to be *actively used* by depositors, curators, and archivists.

- This doesn’t necessarily scale for us -- we want to use it at the tail (preservation) end of existing repository services for electronic theses and local research data.

- These objects are homogeneous and can receive the same normalization treatment in the background.
Metadata
Cat_videos

Applies to
Cat_videos

Metadata can be added at the SIP/AIP level only

Title
Cat video files

Creator
Memphis Jack

Subject
frollicking

Description

Publisher

Contributor

Date
2012-02-03

Use ISO 8061 (YYYY-MM-DD or YYYY-MM-DD/YYYY-MM-DD)

Type

Format
<table>
<thead>
<tr>
<th>File name</th>
<th>File format</th>
<th>Preservation normalization attempted</th>
<th>Preservation normalization failed</th>
<th>Already in preservation format</th>
<th>Access normalization attempted</th>
<th>Access normalization failed</th>
<th>Already in access format</th>
</tr>
</thead>
<tbody>
<tr>
<td>799px-Euroleague-LE_Roma_vs_Toulouse_IC-27</td>
<td>PUID=&quot;fmt/116&quot;, FormatName=&quot;Windows Bitmap&quot;, FormatVersion=&quot;3.0&quot;, FormatRisk=&quot;&quot;, Extension=&quot;dib.bmp,&quot;</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>799px-Euroleague-LE_Roma_vs_Toulouse_IC-27</td>
<td>image/bmp</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>BBHelmet</td>
<td>PUID=&quot;fmt/19&quot;, FormatName=&quot;Acrobat PDF 1.5 - Portable Document Format&quot;, FormatVersion=&quot;1.5&quot;, FormatRisk=&quot;&quot;, Extension=&quot;pdf,&quot;</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>BBHelmet</td>
<td>application/pdf</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>G31DS</td>
<td>image/tiff</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>G31DS</td>
<td>fmt/9</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>G31DS</td>
<td>fmt/10</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>G31DS</td>
<td>fmt/8</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>G31DS</td>
<td>fmt/7</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>lion</td>
<td>PUID=&quot;fmt/91&quot;, FormatName=&quot;Scalable Vector Graphics&quot;, FormatVersion=&quot;1.0&quot;, FormatRisk=&quot;&quot;, Extension=&quot;svg,&quot;</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Showing 1-10 of 21
• We’re configuring Archivematica to use a local LOCKSS network for storage.
• Archivematica is very good at file-level and metadata preservation, but it doesn’t do much for long-term bit-level preservation (beyond checksums).
• Our use of LOCKSS (which stores multiple copies of each file and verifies them against one another) with the new Storage API makes up for this.
...and a bit more about Islandora if you’re curious

- Primary user-facing research data repository platform (i.e., where end users deposit and browse data)

- We’re a Drupal shop (thinking about a migration path for our Drupal 6 sites, including our IR), and it uses Drupal

- Chosen for its capabilities for working with audio/visual data formats (yup, that’s data); we developed our own Tabular Data Solution Pack to make up the difference in quantitative data functionality: https://github.com/axfelix/islandora_solution_pack_data
Recent deposits

View | Manage

Grid view | List view

Cartoon
History
Test
test
Next steps

- Beta test
- Training / educational component
- Liaising with other stakeholders (REB)
Funding redux

- Pilot funding will eventually (soon!) draw down
- We’ve automated repository design wherever possible
- Researchers can deposit on their own; “curatorial” role is before undertaking research and after deposit
- We’ve hooked into ongoing library work, e.g. thesis management
- Money’s spent (productive customization of open-source software which now costs us nothing and will receive improvements); still need to pay rent (repository staffing!)
The University of British Columbia

- Digital Preservation Program
  - cIRcle, DSpace-based repository
  - Digitized collections in CONTENTdm
  - New and legacy born digital archival material
  - Websites (Archive-IT)
- COPPUL Archivematica as a Service and EduCloud
- Centre for Hip Health and Mobility Research Data Management Pilot
- Research Data Management Training
- Creation of Research Computing Services
CiRcle (DSpace)

- Archivematica receives submissions from DSpace
- Also have Archivematica to DSpace workflow
Digital Preservation Program

CONTENTdm

- Master files uploaded to Archivematica
- Archivematica produces access versions and pushes to CONTENTdm
Digital Preservation Program

RBSC/UA born-digital acquisition workflow

External media transfers

SIP

archivematica

AIP

DIP

ICA AtoM

Legacy descriptions and finding aids

EAD metadata updates
Archivematica-as-a-service

Council of Prairie and Pacific University Libraries (COPPUL) - Digital Preservation Working Group

Univ of Victoria (gold)
Univ of Saskatchewan (gold)
Vancouver Island Univ (silver)
Univ of Lethbridge (bronze)
Mt. Royal Univ (bronze)

Hosting partner’s storage cloud (e.g., UBC’s EduCloud)

End user (Gold)

Gold: Silver + web access to Dissemination Info Packages (DIPs) via AtOM platform (bright archive, 2 TB storage, 15 support tickets)

Silver: Bronze + create Submission Info Packages (SIPs), add descriptive metadata, generate PREMIS, create Bagit Archival Info Packages (AIPs) (dark archive, 1 TB storage, 10 support tickets)

Bronze: Transfer into Archivematica, assign identifiers, generate METS, generate technical metadata, add rights metadata (dark archive, 400 GB storage, 5 support tickets)

Local staff

Archivematica virtual machines, one per institution, running on hosting partner’s cloud
Welcome to Kristof Kessler!

We would like to warmly welcome Kristof Kessler who joins CHHM as lead for development of a research data management & curation strategy for the Centre. This position is jointly funded by CHHM (through the CFI Innovation Fund), UBC’s Office of Research Services and the Central Library.

The University has identified support for research data management and curation as a key priority and is working with other universities in Western Canada and across the country to develop strategies to address this issue. The Centre was selected as a key partner in this project at UBC due to the breadth and scope of research data that is acquired/generated, analysed, stored and disseminated by CHHM researchers.

Over the next few months Kristof will carry out a full needs assessment and develop a plan for data acquisition, management and curation at the Centre. Kristof will be working closely with team members from UBC Research IT and the Library to ensure that information obtained from the Centre’s needs assessment provides the framework for implementing a data management and curation strategy for the larger university research community.
Scope

- Interested to understand data needs across many CHHM projects
- 25 researchers and 11 projects were consulted for this pilot
- Other stakeholders - UBC IT, Library, Ethics and Privacy
- Interested to leverage what we have learned for other UBC research groups
Why data needs assessment?

• Different projects and researchers have different data needs
• Structured assessment facilitates revealing and documenting needs more efficiently and effectively
• Need conforming to the requirements (Ethics, NSF, NIH in the US and soon Tri-Council in Canada) - **TC3+**
• **TC3+** -- The Social Sciences and Humanities Research Council (SSHRC), the Natural Sciences and Engineering Research Council (NSERC), the Canadian Institutes of Health Research (CIHR), as well as the Canada Foundation for Innovation (CFI) and Genome Canada
• Capitalizing on Big Data: Agencies launch consultation on digital scholarship, November 2013 - [http://goo.gl/dDuIwB](http://goo.gl/dDuIwB)
• This will be a game changer, like the NSF requirements in 2011!
Existing tools for data assessment

- Data Curation Profiles - Purdue University - http://datacurationprofiles.org/
- Data Asset Framework (DAF) - UK DCC - http://goo.gl/jeS7lh
- CARDIO - http://www.dcc.ac.uk/projects/cardio
- None of these worked for us. Purdue’s questionnaire takes 10-15 hrs per project! However, they provide a comprehensive set of questions to consider in data assessment.
• Pilots were conducted for data curation profiling.
• Identified a need for visual and interactive approach.
• Need to represent three types of objects:
  • data collection instruments
  • processing steps
  • outcomes
• These are the three visual components with which to capture data flows as well as possibilities for improvements in data flows.
Template of Visual and Interactive Approach

Data Collection
- Instrument 1
  - Outcome 1.1
  - Outcome 1.2
- Instrument n
  - Outcome n.1
  - Outcome n.2
  - Outcome n.3

Data Processing and Analysis
- Step 1.1 → Outcome 1.1.2 → Step 1.3
- Step 1.2 → Outcome 1.2.1 → Step 1.4
- Step n.1 → Outcome n.1
- Step n.2 → Outcome n.2
- Step n.3 → Outcome n.3

Finalized Data
- Outcome 1.1
- Outcome 1.x.y

Export and Analysis
- Analysis 1
- Analysis 1
- Analysis 2
- Analysis 3

Sharing and Preservation
- Use data for publications
- Share data with collaborators
- Preserve data for later use
- Other Sharing/Preservation Step

Project Details to be collected:
- Number of assessments/Ts
- # of Participants/Samples, their Identification and Characteristics
- Owner of Data/Intellectual Property
- Data Management Requirements by Funding Agency/Agencies
- Data Services requested/required

Details to be collected for each Instrument:
- Name/ID/Room/Location
- Internal/external data source?
- Activity performed
- Settings
- Operator(s)
- Security
- Use Frequency/Duration

Details to be collected for each Outcome:
- Types of Files/Documents
- Number and Size of Files
- Storage Location
- Description/Organization
- Security and Backup
- Data/File Sharing
- Need for Confidentiality, Integrity and Availability

Details to be collected for each Activity/Processing Step:
- Workstation/Laptop/ID
- Software
- Activity performed
- Security and Backup
- Team Members involved
- Use Frequency/Duration

Lead PI(s)
Key Personnel
Project Duration, Phase and Status
Project Details:
Longitudinal study. Number of participants unknown. Participants generally suffer from some form of more or less benevolent mutation. Data is collected in various locations around the world, but mainly focused on the Eastern United States.

For each Instrument:
- Name/ID/Room/Location
- Internal/external data source?
- Activity performed
- Settings
- Operator(s)
- Security
- Use Frequency/Duration

For each Outcome:
- Types of Files/Documents
- Number and Size of Files
- Storage Location
- Description/Organization
- Security and Backup
- Data/Files Sharing
- Need for Confidentiality, Integrity and Availability

For each Processing Step:
- Workstation/Laptop/ID
- Software
- Activity performed
- Security and Backup
- Team Members involved
- Use Frequency/Duration
Examples

• Example of data assessment profile - http://goo.gl/Q3QZ8J

• This work has allowed us to collate research data outcomes to groups that we called data streams

• Example of data streams, including all data formats - http://goo.gl/AwQEnQ
Research Data Management Training

• Setting the context: UBC Library’s involvement in data management activities (January 2014)

• Data 101: introduction to research data and data management for Librarians (February 2014)

• Storing and working with data (March 2014)
Development of Research Computing Services

- A centralized/distributed research IT support program is urgently needed to advance and coordinate the complex and specialized technology-intensive needs of the research community, address granting agencies requirements to maximize investments, mitigate risks and costs to the institution, and support an environment of research excellence at UBC.
- Research Engagement and Consultation Hub: REaCH-IT proposal
# RCS: Service and Product Design

## 1. Categories

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Engagement &amp; Enablement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Pre and Post Award Consultation</td>
</tr>
<tr>
<td></td>
<td>• Research Data Management</td>
</tr>
<tr>
<td></td>
<td>• Training &amp; Knowledge Transfer</td>
</tr>
<tr>
<td></td>
<td>• Service Pipeline and Continual Improvements</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Intake &amp; Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Software, Systems &amp; Infrastructure Support</td>
</tr>
<tr>
<td></td>
<td>• Advanced Computing and Scientific Support</td>
</tr>
<tr>
<td></td>
<td>• Support Escalations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operations</th>
<th>RCS Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Executive: Strategic Planning, Funding &amp; Partnerships</td>
</tr>
<tr>
<td></td>
<td>• Business Services: Project, Change &amp; Communications Management</td>
</tr>
<tr>
<td></td>
<td>• Internal Operations: Infrastructure, Client Services &amp; Vendor Management</td>
</tr>
</tbody>
</table>
RCS: Suggested Organizational Structure

Organizational Reporting

Research Computing Services

Executive

Engagement & Enablement
- Enterprise Architecture
- Engagement & Enablement
- Research Data Management

Intake & Operations
- Research Intake
- Research Operations
- Advanced Computing
- Scientific

RCS
IT
Library
Compute Canada
University of Alberta

• Strong support for growing our research data management, curation, and preservation programs and services (support digital scholarship)
• Scalable, stable, sustainable and successful
• Open and transparent toolkit - enabling, not defining services
  • Modular, not monolithic
  • Trust what we can see, not what we have been told
  • Scaleable - anticipate growth
  • Sustainable - swim, not sink
• Funding – planning assumption: no new money
  • *do more with the same, do more with less*
• Data services/components include: DMPonline, Dataverse, Hydra/Fedora, Archivematica and investment in people
Partnership

• Archivematica contract and partnership
  • An opportunity to study the issues related to the design of Archival Information Packages (AIPs) for research data sets
    • An opportunity for the development of consistent, shareable AIPs for curated research datasets
    • An opportunity to better understand data formats - automated and manual normalization workflows as part of Format Policy Registry (FPR) development
    • An opportunity to build further footings for ongoing and new work with regional, national and international partners like UBC, SFU and CARL
  
• Use Cases:
  Canadian Polar Data Network
Welcome

The Canadian Polar Data Network aims to develop a research data management infrastructure for the most recent International Polar Year program. We have significant experience with developing a distributed preservation backbone, curating data across the research lifecycle, producing standards-based research metadata, contributing to data and metadata standards, working collaboratively with multiple institutions and across sectors to preserve data, sharing knowledge and aligning with communities of practice. We encourage and welcome new collaborations with researchers and partners.

If you are interested in depositing data, see the Deposit page.
If you are interested in becoming a Network partner, see the Membership page.

Latest Updates

May 2013:

The CPDN is pleased to announce its completion of The Canadian High Arctic Research Station (CHARS): Data and Information Management for Science and Technology report, which outlines requirements, expectations, and goals for CHARS with respect to Arctic data and information management. This substantial document will aid the CHARS program as they move forward with planning.

January 2013:

Final versions of the Data Deposit Agreement and Governance Charter have been uploaded to the Resources page. The Governance Charter has also been summarized and split between the Purpose, History, Principles, and Governance pages of this website for your convenience.
Building Up with Archivematica

- Format policy registry - automated and manual normalization
  - Research data doesn’t always play nice

- Should an AIP be kept small, pairing a single data file with its metadata?
  - If a set of AIPs are part of a larger aggregation, for example, a "dataset", "study" or "project," should a separate AIP exist to serve as the parent of a group of related AIPs? Or should the parent information be bundled within each offspring AIP?

- Should a single AIP consist of all files in a “dataset”?
  - Different definitions of what equates to a dataset. A related set of files serving some defined purpose; act as a unit.

- Should a single AIP exist for a study or a project?
  - Is this the right scope?

- How important is it to be consistent in the design of AIPs?
  - Is it acceptable to produce AIPs on the basis of all three of the above structures?

- How much should today's physical archival storage limitations play in determining the design of AIPs?
  - For example, if the maximum storage block is three terabytes, how much does this contribute to the decision around smaller versus larger AIPs? The maximum block size certainly limits the size of a single AIP, but should it also contribute to an overall principle around the design of AIPs.
Stick framing

• OAIS Reference Model
• Make use of the concept of Archival Information Collection (AIC) in combination with the Archival Information package (AIP)
• Model facilitates encapsulation of data files, metadata and any other necessary documentation in one AIC.
• An AIC can have many AIPs but logically is a single entity for transmission or for archival storage.
• If parts of an AIC are separated due to any incident, the whole structure can be re-constructed using this model

Image source: www.snakeroot.net/farm/BuildingTechniques.shtml
Archival Information Collection (AIC)

- An AIC is a type of AIP that contains metadata pertaining to a class of AIPs. All AIPs that are members of a particular class will inherit the metadata of that AIC (kept in a documentation AIP).
- The structure of an AIC is identical to that of an AIP.
- The use of AICs offers the following advantages:
  - the type and amount of information contained in an AIC is discretionary;
  - large volumes of metadata can be accommodated easily;
  - the information contained in an AIC is preserved in the same archive as its AIPs;
  - certain types of metadata for an entire class of AIPs need only be stored once;
  - a new AIP can be added without significant changes to the overall structure.
- Any orphaned AIP can easily be linked backed to its parent
AIC Creation workflow

1. **Start dataset workflow**
   - User begins AIC* transfer
   - User clicks on metadata icon
   - User inputs AIC metadata

2. **User finishes processing and places AIC in archival storage**
   - User begins AIP** transfer
   - User clicks on metadata icon
   - User inputs metadata linking AIP to AIC

3. **User finishes processing and places AIP in archival storage**
   - User repeats process for all AIPs
   - In archival storage tab, user searches for AIC & AIPs
   - User clicks “Generate AIC METS file”

4. **AIC METS file is generated**
   - End dataset workflow

*An AIC consists of program/project-level metadata and documentation
**An AIP consists of one or more data files and related metadata
Next steps

- Connect Archivematica up with existing CPDN workflows
- Connect Archivematica up with existing and new infrastructure and workflows – Fedora, OpenStack, Dataverse, etc.
  - Leverage Archivematica and ICA-AtoM for University Archives; handshake with Alfresco ERDMS
- Sustainability
  - Partner investment in common technologies will help ensure our tools remain healthy and strong
  - Growing expertise that can be shared between institutions to allow our brightest minds to do less of the same and more of what is really needed
- Succession planning for data; builds a bridge for easier sharing and collaboration with other entities
QUESTIONS?